

Alexander Mehler  
Serge Sharoff  
Marina Santini  
*Editors*

TEXT, SPEECH AND LANGUAGE TECHNOLOGY SERIES 42

# Genres on the Web

*Computational Models and  
Empirical Studies*

# Genres on the Web

# Text, Speech and Language Technology

---

VOLUME 42

---

## *Series Editors*

Nancy Ide, *Vassar College, New York*

Jean Véronis, *Université de Provence and CNRS, France*

## *Editorial Board*

Harald Baayen, *Max Planck Institute for Psycholinguistics, The Netherlands*

Kenneth W. Church, *Microsoft Research Labs, Redmond WA, USA*

Judith Klavans, *Columbia University, New York, USA*

David T. Barnard, *University of Regina, Canada*

Dan Tufis, *Romanian Academy of Sciences, Romania*

Joaquim Llisterri, *Universitat Autònoma de Barcelona, Spain*

Stig Johansson, *University of Oslo, Norway*

Joseph Mariani, *LIMSI-CNRS, France*

For further volumes:

<http://www.springer.com/series/6636>

# Genres on the Web

## Computational Models and Empirical Studies

Edited by

Alexander Mehler

*Goethe-Universität Frankfurt am Main, Germany*

Serge Sharoff

*University of Leeds, United Kingdom*

and

Marina Santini

*KYH, Stockholm, Sweden*

*Editors*

Alexander Mehler  
Computer Science and Mathematics  
Goethe-Universität Frankfurt am Main  
Georg-Voigt-Straße 4,  
D-60325 Frankfurt am Main  
Germany  
Mehler@em.uni-frankfurt.de

Serge Sharoff  
University of Leeds  
LS2 9JT Leeds  
United Kingdom  
s.sharoff@leeds.ac.uk

Marina Santini  
Varvsgatan 25  
SE-117 29 Stockholm  
Sweden  
marinasantini.ms@gmail.com

ISSN 1386-291X

ISBN 978-90-481-9177-2

e-ISBN 978-90-481-9178-9

DOI 10.1007/978-90-481-9178-9

Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010933721

© Springer Science+Business Media B.V. 2010

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Foreword

As a reader, I'm looking for two things from a new book on genre. First, does it offer some new tools for analysing genres; and second, does it explore genres that haven't been much studied before? *Genres on the Web* delivers brilliantly on both accounts, introducing as it does a host of computational perspectives on genre classification and focussing as it does on a range of newly emerging electronic genres. Lacking expertise in the computational modelling thematised throughout the book I can't do much more here than express my fascination with the questions tackled and methods deployed. Having expertise in functional linguistics and its deployment in genre-based literacy programs I can perhaps offer a few observations that might help push this and comparable endeavours along.

First some comments as a functional linguist. Characterising almost all the papers is a two-level approach nicely summarised by Stein et al. in their Table 8.1. On the one hand we have a web genre palette, with many alternative classifications of genres; on the other hand we have document representation, with the many alternative sets of features used to explore web data in relation to genre. The most striking thing about this perspective to me is its relatively flat approach as far as social context and its realisation in language and attendant modalities of communication is concerned.

In systemic functional linguistics for example, it is standard practice to explore variation across texts from the perspectives of field, tenor and mode as well as genre. Field is concerned with institutional practice – domestic activity, sport and recreation, administration and technology, science, social science and humanities and so on. Tenor is concerned with social relations negotiated – in relation to power (equal/unequal) and solidarity (intimate, collegial, professional etc.). Mode is concerned with the affordances of the channel of communication – how does the technology affect interactivity (both type and immediacy), degree of abstraction (e.g. texts accompanying physical behaviour, recounting it, reflecting on it, theorising it) and intermodality (the contribution of language, image, sound, gesture etc. to the text at hand). In my own work genre is then deployed to describe how a culture combines field, tenor and mode variables into recurrent configurations of meaning and phases these into the unfolding stages typifying that social process.

When I referred to a flat model of social context above what I meant was that in this book these four contextual variables tend to be conflated into a single taxonomy of text types, without there being any apparent theoretically informed set of

principles for the flattening. It may well be of course that for one reason or another we do want a simple model of social context and may wish to foreground one field or mode or tenor variable over another. But it might prove more useful to begin with a richer theory of context than we need for any one task, and flatten it in principle, than to try and build a parsimonious model from the start, and complicate it over time.

Turning to document representation, once again from the perspective of systemic functional linguistics, it is standard practice to explore representation in language (and other modalities of communication) from the perspective of various hierarchies and complementarities. The chief hierarchies used are rank (how large are the units considered – e.g. word, phrase, clause, phase, stage, text) and strata (which level of abstraction from materiality is being considered – phonology/graphology, lexicogrammar or discourse semantics). The chief complementarity used is meta-function (are we considering the ideational meanings used to naturalise a picture of reality, the interpersonal meanings used to negotiate social relationships or the textual meanings used to weave these together as waves of information in interpretable discourse).

The meanings dispersed across these ranks, strata and metafunctions are regularly collapsed into a list of descriptive features in this volume, when for different purposes one might want to be selective or value some features over others. Exacerbating this is an apparent need to foreground relatively low-level formal features which are easily computable, since manual analysis is too slow and costly, and in any case so much of the research here is focussed on the automatic retrieval of genres. Beyond this, as Kim and Ross point out, texts are regularly treated as bags of features, as if the timing of their realisation plays no significant part in the recognition of a genre. What saddens me here is the gulf between computational and linguistically informed modelling of genres, for which I know my colleagues in linguistics are responsible – since for the most part they work on form not meaning, and focus on the form of clauses and syllables, not discourse (they still think a language is a set of sentences rather than a communication system instantiated through an indefinitely large lattice of texts).

Next some comments as a functional linguist working in language and education programs over three decades. From the start we of course faced the problem of classifying texts – in our case the genres that students needed to read and write in primary, secondary and tertiary sectors of education, and their relation to workplace discourse and professional development therein. One thing we learned from this work was to be wary of the folk-classifications of genres used by educators. Our primary school teachers for example called everything their students wrote a story, when in fact, from a linguistic perspective, the students engaged in a range of genres. Complicating this was their tendency to evaluate everything the students wrote as a story, in spite of suggesting to students that they choose their own topics or even that they write in any form they choose. As an issue of social justice, we felt we had to replace the folk-categorisation with a linguistically informed one, and take the further step of insisting that this uncommon sense classification be shared between teachers and students. The moral of this experience I feel is that we need to treat

“folksonomies” with great caution when classifying genres, and not expect users to be able to easily bring to consciousness or even demonstrate in practice a genre classification that will best suit the purposes of our own research.

Throughout this literacy focussed action research we have lacked the funding and computational tools to undertake the systematic quantitative analysis thematised in this volume. Instead we had to rely on manual analysis of texts our teacher linguists selected as representative (depending as they did on their own experience, advice from teachers, assessment processes and textbook exemplars). This meant we could build up a picture of genres based on thick descriptions of all the levels of analysis I worried about being flattened above; the great weakness of this approach of course is replicability – were our few texts in fact representative and would quantitative analysis support our findings over time? In practice, the only confirmation we received that we were on the right track lay in the literacy progress of our students, since we were interested in genre because we wanted to redistribute the meaning potential of our culture more evenly than schools have been able to do in the past.

At this point I suspect that most of the authors in this volume would throw up their hands in despair of finding anything useful in our work. So let me just end on a note of caution. What if genres cannot be robustly characterised on the basis of just a few easily computable formal features? What if a flat approach to contextual variables and representational features simplifies research to the point where it is hard to see how the texts considered could have evolved as realisations of the genres members of our culture use to live? Would we be wise to complement flat computationally based quantitative analysis with thick manual qualitative description and see where the two trajectories lead us? And do we need to balance commercially driven research with ideologically committed initiatives (who for example will benefit from the genre informed search engines inspiring so many of the papers herein)?

I’ll stop here, concerned that this preface is turning into a post-script, or even a chapter in a book where prefacing is where I barely belong! My thanks to the editors for opening up this work, which will prove indispensable for readers with many converging concerns. I’ll do what I can to point my students and colleagues in the direction of the transdisciplinary dialogue which I’m sure will be inspired by the genre analysts dialoguing here.

Sydney, Australia  
March 2009

James R. Martin



# Personal Note

*Here let us breathe and haply institute  
A course of learning and ingenious studies.  
Shakespeare, The taming of the shrew, Act I, scene I*

To all of you who have been involved in this book I want to say: Thank you! This book is very much the result of your collective efforts. It would not have come about without your commitment and interest in the concept of genre, this untamed shrew.

My first mention goes to the *authors* who readily accepted to contribute to this volume. Many thanks for your chapters, dear Authors, that show the state of the art of empirical and computational genre research.

I am also most grateful to our *reviewers* whose comments were most valuable. Many thanks for your detailed feedback, dear Reviewers, that has improved the content, presentation and style of our chapters.

Thank you to everybody for sharing your knowledge and dedication to make this volume possible.

Have we started taming the shrew? I am sure we have.

Marina Santini  
Book Coordinator



# Contents

## Part I Introduction

- 1 Riding the Rough Waves of Genre on the Web** ..... 3  
Marina Santini, Alexander Mehler, and Serge Sharoff

## Part II Identifying the Sources of Web Genres

- 2 Conventions and Mutual Expectations** ..... 33  
Jussi Karlgren
- 3 Identification of Web Genres by User Warrant** ..... 47  
Mark A. Rosso and Stephanie W. Haas
- 4 Problems in the Use-Centered Development of a Taxonomy of Web Genres** ..... 69  
Kevin Crowston, Barbara Kwaśnik, and Joseph Rubleske

## Part III Automatic Web Genre Identification

- 5 Cross-Testing a Genre Classification Model for the Web** ..... 87  
Marina Santini
- 6 Formulating Representative Features with Respect to Genre Classification** ..... 129  
Yunhyong Kim and Seamus Ross
- 7 In the Garden and in the Jungle** ..... 149  
Serge Sharoff

**8 Web Genre Analysis: Use Cases, Retrieval Models, and Implementation Issues** ..... 167  
 Benno Stein, Sven Meyer zu Eissen, and Nedim Lipka

**9 Marrying Relevance and Genre Rankings: An Exploratory Study** ... 191  
 Pavel Braslavski

**Part IV Structure-Oriented Models of Web Genres**

**10 Classification of Web Sites at Super-Genre Level** ..... 211  
 Christoph Lindemann and Lars Littig

**11 Mining Graph Patterns in Web-Based Systems: A Conceptual View** . 237  
 Matthias Dehmer and Frank Emmert-Streib

**12 Genre Connectivity and Genre Drift in a Web of Genres** ..... 255  
 Lennart Björneborn

**Part V Case Studies of Web Genres**

**13 Genre Emergence in Amateur Flash** ..... 277  
 John C. Paolillo, Jonathan Warren, and Breanne Kunz

**14 Variation Among Blogs: A Multi-Dimensional Analysis** ..... 303  
 Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova

**15 Evolving Genres in Online Domains: The Hybrid Genre of the Participatory News Article** ..... 323  
 Ian Bruce

**Part VI Prospect**

**16 Any Land in Sight?** ..... 351  
 Marina Santini, Serge Sharoff, and Alexander Mehler

**Index** ..... 355

# Contributors

**Douglas Biber** English Department, Northern Arizona University, Flagstaff, AZ, USA, douglas.biber@nau.edu

**Lennart Björneborn** Royal School of Library and Information Science, Copenhagen, Denmark, lb@iva.dk

**Pavel Braslavski** Institute of Engineering Science RAS, 620219 Ekaterinburg, Russia, pb@imach.uran.ru; pb@yandex-team.ru

**Ian Bruce** University of Waikato, Hamilton, New Zealand, ibruce@waikato.ac.nz

**Kevin Crowston** School of Information Studies, Syracuse University, Syracuse, NY, USA, crowston@syr.edu

**Matthias Dehmer** Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Vienna, Austria; Institute for Bioinformatics and Translational Research, Hall in Tyrol, Austria, matthias.dehmer@univie.ac.at; mdehmer@geometrie.tuwien.ac.at; Matthias.Dehmer@umit.at

**Frank Emmert-Streib** Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, UK, v@bio-complexity.com

**Eric Friginal** Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, GA, USA, efriginal@gsu.edu

**Jack Grieve** QLVL Research Unit, University of Leuven, Leuven, Belgium, Jack.Grieve@arts.kuleuven.be

**Stephanie W. Haas** School of Information & Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA, shaas@email.unc.edu

**Jussi Karlgren** Swedish Institute of Computer Science (SICS), Stockholm, Sweden, jussi@sics.se

**Yunhyong Kim** Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, Glasgow, UK; School of Computing, Robert Gordon University, Aberdeen, UK, ykim1@rgu.ac.uk

**Breanne Kunz** School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA, bkunz@indiana.edu

**Barbara Kwasnik** School of Information Studies, Syracuse University, Syracuse, NY, USA, bkwasnik@syr.edu

**Christoph Lindemann** Department of Computer Science, University of Leipzig, Leipzig, Germany, cl@rvs.informatik.uni-leipzig.de

**Nedim Lipka** Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, nedim.lipka@uni-weimar.de

**Lars Littig** Department of Computer Science, University of Leipzig, Leipzig, Germany, littig@rvs.informatik.uni-leipzig.de

**Alexander Mehler** Computer Science and Mathematics, Goethe-Universität Frankfurt am Main, Georg-Voigt-Straße 4, D-60325 Frankfurt am Main, Germany, Mehler@em.uni-frankfurt.de

**Sven Meyer zu Eissen** Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, sven@meyer-zu-eissen.de; sven.meyer-zu-eissen@medien.uni-weimar.de

**Tatiana Nekrasova** English Department, Northern Arizona University, Flagstaff, AZ, USA, Tatiana.Nekrasova@nau.edu

**John C. Paolillo** School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA, paolillo@indiana.edu

**Seamus Ross** iSchool, University of Toronto, Toronto, CA, seamus.ross@utoronto.ca

**Mark A. Rosso** School of Business, North Carolina Central University, Durham, NC 27707, USA, mrosso@nccu.edu

**Joseph Rubleske** School of Information Studies, Syracuse University, Syracuse, NY, USA, jrublesk@gmail.com

**Marina Santini** KYH, Stockholm, Sweden, marinasantini.ms@gmail.com

**Serge Sharoff** Centre for Translation Studies, University of Leeds, LS2 9JT Leeds, UK, s.sharoff@leeds.ac.uk

**Benno Stein** Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, benno.stein@uni-weimar.de

**Jonathan Warren** School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA, jowarren@indiana.edu



**Part I**  
**Introduction**



# Chapter 1

## Riding the Rough Waves of Genre on the Web

### Concepts and Research Questions

Marina Santini, Alexander Mehler, and Serge Sharoff

#### 1.1 Why Is Genre Important?

*Genre*, in the most generic definition, takes the meaning “kind; sort; style” (OED). A more specialised definition of genre in OED reads: “A particular style or category of works of art; esp. a type of literary work characterised by a particular form, style, or purpose.” Similar definitions are found in other dictionaries, for instance, OALD reads “a particular type or style of literature, art, film or music that you can recognise because of its special features”. Broadly speaking, then, generalising from lexicographic definitions, genre can be seen as a classificatory principle based on a number of characterising attributes.

Traditionally, it was Aristotle, in his attempt to classify existing knowledge, who started genre analysis and defined some attributes for genre classification. Aristotle sorted literary production into different *genre classes* by focussing on the attributes of purpose and conventions.<sup>1</sup>

After him, through the centuries, numberless definitions and attributes of the genre of written documents have been provided in differing fields, including literary criticism, linguistics and library and information science. With the advent of digital media, especially in the last 15 years, the potential of genre for practical applications in language technology and information technology has been vigorously emphasised by scholars, researchers and practitioners.

---

M. Santini (✉)  
KYH, Stockholm, Sweden  
e-mail: marinasantini.ms@gmail.com

<sup>1</sup> More precisely, “in the *Poetics*, Aristotle writes, ‘the medium being the same, and the objects [of imitation] the same, the poet may imitate by narration – in which case he can either take another personality as Homer does, or speak in his own person, unchanged – or he may present all his characters as living and moving before us’ . . . . The *Poetics* sketches out the basic framework of genre; yet this framework remains loose, since Aristotle establishes genre in terms of both convention and historical observation, and defines genre in terms of both convention and purpose”. Glossary available at The Chicago School of Media Theory, retrieved April 2008.

But why is genre important? The short answer is: because it reduces the cognitive load by triggering expectations through a number of conventions. Put in another way, genres can be seen as sets of *conventions* that transcend individual texts, and create frames of recognition governing document production, recognition and use. Conventions are *regularities* that affect information processing in a repeatable manner [29]. Regularities engage *predictions* about the “type of information” contained in the document. Predictions allow humans to identify the *communicative purposes* and the *context* underlying a document. Communicative purposes and context are two important principles of human communication and interactions. In this respect, genre is then an implicit way of providing background information and suggesting the cognitive requirements needed to *understand a text*. For instance, if we read a sequence of short questions and brief answers (*conventions*), we might surmise that we are reading FAQs (*genre*); we then realize that the purpose of the document is to instruct or inform us (*expectations*) about a particular topic or event of interest. When we are able to identify and name a genre thanks to a recurrent set of regular traits, the functions of the document and its communicative context immediately build up in our mind. Essentially, knowing the genre to which a text belongs leads to predictions concerning form, function and context of communication. All these properties together define what Bateman calls the “the most important theoretical property” of genre for empirical study, namely the power of *predictivity* [9, p. 196]. The potential of predictivity is certainly highly attractive when the task is to come to terms with the overwhelming mass of information available on the web.

### ***1.1.1 Zooming In: Information on the Web***

The immense quantity of information on the web is the most tangible benefit (and challenge) that the new medium has endowed us as web users. This wealth of information is available either by typing a URL (suggested by other web external or web internal sources) or by typing a few keywords (the query) in a search box. The web can be seen as the *Eldorado* of information seekers.

However, if we zoom in a little and focus our attention on the most common web documents, i.e. written texts, we realize that finding the “right” information for one’s need is not always straightforward. Indeed, a common complaint is that users are overwhelmed by huge amounts of data and are faced with the challenge of finding the most relevant and reliable information in a timely manner. For some queries we can get thousands of hits. Currently, commercial search engines (like Google and Yahoo!) do not provide any hint about the *type of information* contained in these documents. Web users may intuit that the documents in the result list contain a *topic* that is *relevant* to their query. But what about other dimensions of communication?

As a matter of fact, Information Retrieval (IR) research and products are currently trying to provide other dimensions. For instance, some commercial search engines provide specialised facilities, like Google Scholar or Google News. IR research is

active also in plagiarism detection,<sup>2</sup> in the identification of context of interaction and search,<sup>3</sup> in the identification of the “sentiment” contained in a text,<sup>4</sup> and in other aspects affecting the reliability, trust, reputation<sup>5</sup> and, in a word, the appropriateness of a certain document for a certain information need.

Still, there are a number of other dimensions that have been little explored on the web for retrieval tasks. *Genre* is one of these. The potential of genre to improve information seeking and reduce information overload was highlighted a long time ago by Karlgren and Cutting [47] and Kessler et al. [48]. Rosso [76] usefully lists a pros and cons of investigating web retrieval by genres. He concludes on a positive note, saying that genre “can be a powerful hook into the relevance of a document. And, as far as the ever-growing web is concerned, web searches may soon need all the hooks they can get”. Similarly, Dillon [29] states “genre attributes can add significant value as navigation aids within a document, and if we were able to determine a finer grain of genre attributes than those typically employed, it might be possible to use these as guides for information seekers”.

Yet, the idea that the addition of genre information could improve IR systems is still a hypothesis. The two currently available genre-enabled prototypes – X-SITE [36] and WEGA (see Chapter 8 by Stein et al., this volume) – are too preliminary to support this hypothesis uncontroversially. Without verifying this hypothesis first, it is difficult to test genre effectiveness in neighbouring fields like human-computer interaction, where the aim is to devise the best interface to aid navigation and document understanding (cf. [29]).

IR is not the only field that could thrive on the use of genre and its automatic classification. Traditionally, the importance of genre is fully acknowledged in research and practice in qualitative linguistics (e.g. [96]), academic writing (e.g. [18]) and other well-established and long-standing disciplines.

However, also empirical and computational fields – the focus of this volume – would certainly benefit from the application of the concept of genre. Many researchers in different fields have already chosen the *genre lens*, for instance in corpus-based language studies (e.g. [14, 24, 58]), automatic summarisation [87], information extraction [40], creation of language corpora [82], e-government (e.g. [37]), information science (e.g. [39] or [68]), information systems [70] and many other activities.

The genres used by Karlgren and Cutting [47] were those included in the Brown corpus. Kessler et al. [48] used the same corpus but were not satisfied with its genre taxonomy, and re-labelled it according to their own nomenclature. Finding the appropriate labels to name and refer to genre classes is one of the major obstacles

---

<sup>2</sup> For instance, see “PAN’09: 3rd Int. PAN Workshop – 1st Competition on Plagiarism Detection”.

<sup>3</sup> For instance, see “ECIR 2009 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation”.

<sup>4</sup> For instance, see “CyberEmotions” <http://www.cyberemotions.eu/>

<sup>5</sup> For instance, see “WI/IAT’09 Workshop on Web Personalization, Reputation and Recommender Systems”.

in genre research (see Chapter 3 by Rosso and Haas; Chapter 4 by Crowston et al., this volume). But, after all, the naming difficulty is very much connected with the arduousness of defining genre and characterising genre classes.

## 1.2 Trying to Grasp the Ungraspable?

Although undeniably useful, the concept of genre is fraught with problems and difficulties. Social scientists, corpus linguists, computational linguists and all the computer scientists working on empirical and computational models for genre identification are well aware that one of the major stumbling blocks is the lack of a shared definition of genre, and above all, of a shared set of attributes that uncontroversially characterise genre.

Recently, new attempts have been made to pin down the essence of genre, especially of web genre (i.e. the genre of digital documents on the web, a.k.a. cyber-genre).

A useful summary on the diverse perspectives is provided by Bateman [9]. Bateman first summarises the views of the most influential genre schools – namely *Genre as social action* put forward by North American linguists and *Genre as social semi-otic* supported by systemic-functional linguistics (SFL)<sup>6</sup> – then he points out the main requirements for a definition of genre for empirical studies:

Fine linguistic detail is a prerequisite for fine-grained genre classification since only then do we achieve sufficient details (i) to allow predictions to be made and (ii) to reveal more genres than superficially available by inspection of folk-labelling within a given discourse community. When we turn to the even less well understood area involved in multimodal genre, a fine-grained specification employing a greater degree of linguistic sophistication and systematicity on the *kind of forms that can be used for evidence for or against the recognition of a genre category is even more important* ([9, p. 196] – italics in the original)

Bateman argues that the current effort to characterise the kinds of documents found on the web is seriously handicapped by a relatively simple notion of genre that has only been extended minimally from traditional, non-multimodal conceptions. In particular, he claims that the definition of cybergenre, or web genres, in terms of <content, form, functionality>, taken as an extension of the original tuple <content, form> is misleading (cf. also Karlgren, Chapter 2 in this volume). Also the dual model proposed by Askehave and Nielsen [4], which extends the notion of genre originally developed by Swales [89], is somewhat unsatisfying for Bateman. Askehave and Nielsen [4] propose a two-dimensional genre model in which the generic properties of a web page are characterised both in terms of a traditional text perspective and in terms of the medium (including navigation). They motivate this divide in the discussion of the homepage web genre. The traditional part of their model continues to rely on Swales' view of genre, in which he analyses genres at

---

<sup>6</sup> The contraposition between these two schools from the perspective of teaching is also well described in Bruce [18], Chapter 2.

the level of purpose, moves and rhetorical strategies. The new part extends the traditional one by defining two modes that users take up in their interaction with new media documents: users may adopt either a reading mode or a navigation mode. Askehave and Nielsen argue that hyperlinks and their use constitute an essential extension brought about by the medium. Against this and all the stances underpinning hypertext and hyperlinking facilities as the crucial novelty, Bateman argues that the consideration that a more appropriate definition of genre should not open up a divide between digital and non digital artefacts.

Other authors, outside the multimodal perspective underpinned by Bateman [9], propose other views. Some recent genre conceptions are summarised in the following paragraphs.

Bruce [18] builds upon some of the text types proposed by Biber [11] and Biber [12] to show the effectiveness of his own genre model. Bruce proposes a two-layered model and introduces two benchmark terms: social genres and cognitive genres. Social genres refer to “socially recognised constructs according to which whole texts are classified in terms of their overall social purpose”, for instance personal letters, novels and academic articles. Cognitive genres (a.k.a. text types by some authors) refer to classification terms like narrative, expository, descriptive, argumentative or instructional, and represent rhetorical purposes. Bruce points out that cognitive genres and social genres are characterised by different kinds of features. His dual model, originally devised for teaching academic writing, can be successfully applied to web genre analysis, as shown by Bruce’s chapter in this volume.

The genre model introduced by Heyd [43] has been devised to assess whether email hoaxes (EH) are a case of digital genre. Heyd provides a flexible framework that can accommodate for discourse phenomena of all kinds and shapes. The author suggests that the concept of genre must be seen according to four different parameters. The vertical view (parameter 1) provides levels of descriptions of increasing specificity, that start from the most general level, passing through an intermediate level, down to a sublevel. This view comes from prototype theory and appears to be highly applicable to genre theory (cf. also [53]), with the intermediate level of genre descriptions being the most salient one. The horizontal view (parameter 2) accounts for genre ecologies, where it is the interrelatedness and interdependence of genre that is emphasised. The ontological status (parameter 3) concerns the conceptual framework governing how genre labels should be ascribed, i.e. by a top-down or a bottom-up approach. In the top-down approach, it is assumed that the genre status depends upon the identification of manifest and salient features, be they formal or functional (such a perspective is adopted also in Chapter 7 by Sharoff, this volume); by contrast a bottom up approach assumes that the genre status is given by how discourse communities perceive a discourse phenomenon to be a genre (see Chapter 3 by Rosso and Haas; Chapter 4 by Crowston et al., this volume). The issue of genre evolution (parameter 4) relates to the fast-paced advent and evolution of language on the Internet and to the interrelation with socio-technical factors, that give rise to genre creation, genre change and genre migration. Interestingly, Heyd suggests that the frequently evoked hybridity of Computer Mediated Communication (CMC) genres can be accounted for by the “transmedial stability that

predominates on the functional sublevel while genre evolution occurs on the formal sublevel: this explains the copresence of old and new in many digital genres” [43, p. 201].

Martin and Rose [60] focus on the relations among five major families of genres (stories, histories, reports, explanations and procedures) using a range of descriptive tools and theoretical developments. Genre for Martin and Rose is placed within the systemic functional model (SFL). They analyse the relationship between genres in terms of a multidimensional system of oppositions related to the function of communication, e.g. instructing vs. informing.

This overview on recent work on genre and web genre shows that the debate on genre is still thrilling and heated. It is indeed an intellectually stimulating discussion, but do we need so much theory for a definition of web genre for empirical studies and computational applications?

### ***1.2.1 In Quest of a Definition of Web Genre for Empirical Studies and Computational Applications***

Päivärinta et al. [70] condense in a nutshell the view on genre for information systems:

[...] genres arguably emerge as fluid and contextual socio-organisational analytical units along with the adoption of new communication media. On the other hand, more stabilised genre forms can be considered sufficiently generic to study global challenges related to the uses of communications technology or objective enough to be used as a means for automatic information seeking and retrieval from the web.

Essentially, an interpretation of this statement would encourage the separation of the theoretical side from the practical side of genre studies. After all, on the empirical and computational side, we need very little. Say that, pragmatically, genre represents a *type of writing, which has certain features that all the members of that genre should share*. In practical terms, and more specifically for automatic genre classification, this simply means:

1. take a number of documents belonging to different genres;
2. identify and extract the features that are shared within each type;
3. feed a machine learning classifier to output a mathematical model that can be applied to unclassified documents.

The problem with this approach is that without a theoretical definition and characterisation underpinning the concept of genre, it is not clear how to select the members belonging to a genre class and in which way the genre labels “represent” a selected genre class. A particular genre has conventions, but they are not fixed or static. Genre conventions unfold along a continuum that ranges from weak to strong genre conformism. Additionally, documents often cross genre boundaries and draw on a number of characteristics coming from different genres. Spontaneous questions then arise, including:

(A) Which are the features that we want use to draw the similarities or differences between genre classes? (B) Who decides the features? (C) How many features are really the core features of a genre class? (D) Who decides how many raters must agree on the same core feature set and on the same genre names in order for a document to belong to a specific genre? (E) Are the features that are meaningful for humans equally meaningful for a computational/empirical model? (F) Are genre classes that are meaningful for humans equally meaningful for a computational model? And so on and so forth.

Apparently, theoretical/practical definitions of genres have no consequence whatsoever when deciding about the *actual typification* of the genre classes and genre labels required to build empirical and computational models. This gap between definitions and empirical/classification studies has been pointed out by Andersen, who notes that freezing or isolating genre, statistically or automatically, dismantles action and context (Andersen, personal communication; cf. also Andersen [2, 3]), the driving forces of genre formation and use. In this way, genres become *lifeless* texts, merely characterized by formal structural features.

In summary, we are currently in a situation where there is the need to exploit the *predictability* inherent in the concept of genre for empirical and computational models, while genre researchers are striving to find an adequate definition of genre that can be agreed upon and shared by a large community. Actually, the main difficulty is to work out optimal methods to define, select and populate the constellation of genres that one wishes to analyse or identify without hindering replication and comparison.

### 1.3 Empirical and Computational Approaches to Genre: Open Issues

Before moving on to the actual chapters, the next three sections focus on the most important open issues that characterise current empirical and computational genre research. These open issues concern the nature of web documents (Section 1.3.1), the construction and use of corpora collected from the web (Section 1.3.2) and the design of computational models (Section 1.3.3).

#### 1.3.1 Web Documents

While paper genres tend to be more stable and controlled given the restrictions or guidelines enforced by publishers or editors, on the web centrifugal forces are at work. Optimistically, Yates and Sumner [97] and Rehm [75] state that the process of imitation and the urge for mutual understanding act as centripetal forces. Yet, web documents appear much more uncontrolled and unpredictable if compared to publications on paper.

First of all, what is a web document? On the web, the boundary of a document is unclear. Is a web document a single file? If so, a frame composing a web page could be an autonomous web document. Or is it the individual web page? But then where is the core information in a web page? Can we identify it clearly? Web pages can be just navigational or both navigational and content bearing. How many autonomous texts can be found in a individual web pages? Maybe it is safer to identify the web document with a web site as a whole? Where then is the boundary of a web site?

It appears evident that on the web the granularity of documents cannot be kept implicit, because texts with different content and functions are tiled and connected together more tightly than on paper documents, where the physical pages act, sometimes, as “fences” that separate different contents and functions.

For instance, if we compare a daily newspaper like *The Times*, and its web counterpart, *Timesonline*,<sup>7</sup> we can realize that the “paper” gives a much more static status to the concept of “document”. On the paper too, a document can be interpreted at various degrees of granularity. For instance, a single text (like an editorial or a commercial advertisement) is a document; a page (like the newspaper frontpage) is a document; and a medium (like a newspaper or a book) is a document as well. But on the web, hyperlinking, search facilities, special features (like dynamic marquees), and other technicalities make the concept of documents much more dynamic and flexible. This is evident if we compare the same document granularity on the paper and on the web. Figure 1.1 shows an online frontpage (LHS) and a paper frontpage (RHS). Both the graphic appearance and the functionality associated with these documents differ. The basic idea of providing an entry point with snippets of the contents is maintained in both media,<sup>8</sup> but the online frontpage has also a corollary of interactive activities, such as menus, search boxes, and dynamic texts. Additionally, past editions or news articles are immediately available by clicking on the archive link. While the paper frontpage is a self-contained unity, with internal cross-references and occasional citations to external sources, the online frontpage has no boundaries, each web page or each section of a web page can be connected to both internal and external pages. Interactivity, multimodality and dynamic content make the online frontpage different from a paper frontpage. While the paper frontpage has the physical boundary of the first page in a newspaper, and one can dwell on it, the online frontpage is a gateway, i.e. a navigational page providing access to other pages. It becomes clear, then, that when working with web documents, although all levels of granularity are plausible, there is the need to spell out explicitly and justify the *unit of analysis*.

Essentially, web genres are composite functional types of web-based communication. For this reason, in order to make them an object of automatic classification we need to decide on the reference units of their manifestations. That is, we need

---

<sup>7</sup> Global edition: <http://www.timesonline.co.uk/tol/global/>, or UK edition <http://www.timesonline.co.uk/tol/news/>

<sup>8</sup> As noted by Bateman [9] functionality belongs to both paper and web documents.



Fig. 1.1 Frontpage of a web newspaper vs. its printed counterpart

to decide which document structures of the web are attributed to web genres: e.g., self-contained pages [78] or their constituents [74, 75, 88, 94], websites [57, 65] or even larger units such as, for example, domains consisting of several websites [15]. When it comes to modelling such web document structures as instances of web genres, we realise that the vector space approach (see Part III, this volume) is only one of many ways to model genre computationally. One reason is that if one had to choose a single characteristic of genres on the web, then the linkage of their instances by hyperlinks would be a prime candidate (see Part IV, this volume). Web genres are manifested by pages [78, 79] that are interlinked to create, in effect, larger units *above the level of single pages*. Thus, any decision on the manifestation unit of web genres should clarify the role of hyperlink-based structure formation as a source of attributing these units to the focal web genres.

With respect to web content mining, Menczer [67] observes that the content of a page is similar to that of the pages that link to it. We may vary this *link-content conjecture* by saying that you shall know a web genre (though not solely) by the link-based neighbourhood of its instances. Following this line of thinking we can distinguish three levels of modelling web documents as instances of web genres (cf. [62, 75]):

- On the *micro level* we analyse page-level [77] units and their constituents [88] as self-contained (though not necessarily the smallest) manifestations of web genres. These then enter into websites as more complex web genre units.
- On the *meso level* we deal with single or conglomerate websites and their web-specific structure formation which, of course, is hardly found beyond the web [15].
- On the *macro level* we deal with the web as a whole from the perspective of complex network analysis and related approaches [30].

In order to exemplify the differences of these three perspectives, take social software as an example: here, web genre analysis may focus microscopically on single weblogs [69] as instances of this genuine web genre or on networks of blogs which are interlinked by trackbacks and related means [42, 52]. From the point of view of a mesoscopic perspective we may analyse, more specifically, *blog sites* as sub-networks of networked blogs whose connection may result from their discussion of a common topic [52]. Last but not least, we gain a macroscopic perspective by taking into account blog network-external links which embed blogs into the web as a whole. Analogously, by analysing Wikipedia as an instance of web-based knowledge communication we may distinguish wiki-internal structures (e.g. in the form of portals) from wiki-external structures (by analysing links from wikis to pages of external sites) [61].

Genre research has focussed mostly on analysing micro and meso level units as instances of web genres (see, for example, the contributions of Björneborn [16] and Santini [80]). One might hesitate to consider macro level approaches under this perspective. However, by analogy to text genres we know of the existence of macro genres which are generated from instances of different (micro-level) genres [59]. In the web, this build-up of macro genres is more explicit on the instance level as authors make use of hyperlinks to interconnect micro or meso level units of the same macro genre. Further, the macro-level perspective opens the chance to study both the network of web genres as a network of hypertext types (which evolve as part of the same semiotic universe) as well as the network of their instances. This gives a bipartite perspective on networking on the level of hypertext types and their instances which is nearly inaccessible to text genre analysis.

Björneborn [15] (and in this volume) offers a rich terminology by distinguishing four nested levels of structure formation (i.e., pages, directories, domains and sites) together with a typology for the perspective classification of a link. A university website, for example, is described as comprising different websites of various genres (among other things, the difference between project homepages and personal academic homepages) whereas, together with other university websites, it forms the domain of academia. Thelwall et al. [92] generalise this model in terms of the *Alternative Document Model*. They do that by additionally distinguishing *web spaces* as sub-networks of web documents demarcated, e.g., by geographic criteria.

If we, on the other hand, look on the micro level of structure formation in the web, we see that the notion of *logical document structure* dominates the corresponding range of models. By analogy to text documents [72] the idea is that the attribution of a web document to a web genre is made more difficult by insufficiently explicit logical document structures. This can come as a result of, e.g., the abuse of tags [6] or the failure to use hyperlinks to connect functionally homogeneous, monomorphic document units [66]. Manifestations of webgenres are analysed, for example, as *compound documents* [31], as *logical domains* [54], as *logical documents* [55, 91] or as *multipage segments* [25].<sup>9</sup> Whatever is seen to be the exact unit of manifestation

---

<sup>9</sup> See also Tajima et al. [90], Cohn and Hofmann [23] and Chakrabarti et al. [22] for topic-related approaches in this line of research.

of a web genre – say on the page level, below or above – approaches to learning corresponding classifiers face the formation of hyperlink-based, network-inducing structures apart from purely hierarchical text structures. Notwithstanding these differences we have to state that whatever is seen to be the exact unit of manifestation of a web genre – say on the page level, below or above – the corresponding classifiers, in their approach to learning, face the challenge of forming hyperlink-based, network-inducing structures that are fundamentally different from [or more complex than] purely hierarchical text structures. It might be the case that more complex graph models (above the level of tree-like structures) are needed to bring into focus the web genre modelling of the future, which complete and complement the more traditional vector space approaches.

One obvious consequence of the composite and diversified characterisation of web documents is the necessity to devise *classification schemes* not constrained to the single genre class assignment. Intuitively, there is a high likelihood that many web documents (whatever their granularity) would fall into multiple genre classes, and many would remain unclassified by genre because of a high degree of individualisation or hybridism. Genre analysts also point out that the acknowledgement and usage of genres are subjective and depend upon membership in a discourse community (cf. Chapter 4 by Crowston et al., this volume). The flexibility of a classification scheme would then account also for the subjectivity of use and recognition of genres by web users. Since the web serves many communities and web users are exposed to innumerable contacts, it would be wiser to devise a classification scheme addressing this complexity in the future.

Importantly, the nature and the unit of analysis of web documents has not only repercussions on genre classification schemes, but also affects *genre evolution*. Genres are historical entities, they develop over time, and in response to social, cultural and technological contexts (e.g. see Chapter 13 by Paolillo et al., this volume). Existing genres may simply go out of fashion, or undergo transformation. Frequently, genres on the web evolve when they migrate from one medium to another (see Fig. 1.1). They can also be created from scratch, due to new web technologies or new contexts of interaction. The personal home page and blog genres are the classical examples of web genres whose existence cannot be imagined outside the web. The formation of new genres from an antecedent can also be monitored computationally [64]. For example, it is easily predictable that the recent booming of social networks – from Facebook to Twitter and LinkedIn – will presumably destabilise and change web genres like the personal home page and blog that were thought to be “novel” up to very recently. The technology offered by social networks in creating personal profiles, live feeds, blogging, notes and material of any kind at the same time are clear signs that new genres are going to materialise soon.

In summary, web documents would require a flexible genre classification scheme capable of making sense of (1) the composite structure of web documents at any level of unit of analysis; (2) the complexity of interaction allowed by web documents; (3) the subjective and differing naming conventions due the membership to different communities and finally (4) the tendency towards rapid change and evolution of genre patterns.

### 1.3.2 Corpora, Genres and the Web

According to John Sinclair, a corpus is “a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language” [85]. Criteria for selecting texts for a corpus can include information about the authorship, audience or domain of its constituent texts, but selection of texts by their genre is nearly always present as one of the main criteria for designing a traditional corpus. For instance, the Brown Corpus, the first computer corpus developed in the 1960s, was compiled using the following linguistic criteria [51]:

- it was restricted to texts written originally in English by native speakers of American English (as far as this can be determined);
- the texts were first published in the United States in 1961;
- samples of entire texts were selected starting from a random sentence boundary and ending by the first sentence boundary after an uninterrupted stretch of 2,000 words (this means that texts themselves had to be longer than 2,000 words);
- texts were selected from 15 text categories: (A) Press: reportage, (B) Press: editorial, (C) Press: Reviews, (D) Religion, (E) Skill and hobbies, (F) Popular lore, (G) Belles-lettres (biography, memoirs, etc.), (H) Miscellaneous: US Government & House Organs, (J) Learned (i.e., research articles), (K) Fiction: general, (L) Fiction: mystery and crime, (M) Fiction: science, (N) Fiction: adventure and western, (P) Fiction: romance and love story, (R) Humor.

As we can see from this specification, the only variation among samples present in the Brown Corpus concerns their text categories, which roughly correspond to genres (the only possible exceptions are Religion, Skills and Hobbies, but even they constitute distinct functional styles, which are normally associated with specific genres, i.e., sermons and DIY magazines).

Further development of corpora, e.g., creation of the Bank of English [84], the British National Corpus [5], or the American National Corpus [44], resulted in a greater variety of parameters for describing their constituent texts, but they nevertheless classified them into genres, even if the genres in each corpus were defined in various incompatible ways. For instance, the original release of the BNC classified the written texts into their publication medium (e.g., book or periodical), domain (commerce, social sciences or imaginative), and target audience. This provided an opportunity to specify some genres by restricting one or more BNC metadata tags, e.g., fiction corresponds to imaginative texts, research papers can be found by a combination of tags coding texts from natural, applied or social sciences, aimed at the professional audience, and not published as books. Since this situation was treated as less than adequate, David Lee developed a system of 70 genre tags for BNC documents [53], e.g., `W_ac_natsci` or `W_ac_socsci` for academic papers in the domains of natural or social sciences.<sup>10</sup>

---

<sup>10</sup> This is another example where a difference in the domain of a text contributes to a difference in its genre.

The situation with genres in web-derived corpora is a bit different. The majority of large web corpora have not been collected in any pre-planned way with respect to their target domains or genres. Collection of texts from the web normally involves taking publicly accessible documents from a list of URLs. This means it is driven by the availability of sources, which leaves many parameters of corpus collection, such as genres, unspecified.

Some web corpora are created by “focused crawling”, which, in its simplest form, involves selecting several websites containing a large number of texts which are of interest to the corpus collector, and retrieving the entire set of texts from these websites, e.g., the entire Wikipedia or webpages of major universities. More advanced methods of focused crawling involve starting with a seed set of links and then collecting links to other relevant websites, with the relevance assessed by keywords and/or hypertext links between pages, as similar pages tend to have more inter-connections with each other [21]. In all cases of focused crawling, the seed set of URLs used for collecting a web corpus restricts its range of genres, but does not define it precisely. For instance, articles retrieved from Wikipedia can be biographies, time-lines of events, introductions to academic theories, some subtypes of news items, etc., but they cannot include such genres as blogs, fiction, humour or memoirs.

Another method for corpus collection relies on making automated queries to a major search engine and retrieving webpages for the top N (10-20-100) URLs returned by it. The choice of keywords affects the composition of the resulting corpus to some extent. For instance, if a large number of specialised terms are used in queries, e.g., *amnesia*, *myoclonic*, *paroxysmal*, the resulting corpus will contain mostly highly technical medical texts and relatively few patient leaflets or news items. Using common words from the general lexicon, e.g., *picture*, *extent*, *raised*, *events*, results in a corpus with a variety of domains and text types [81]. On the other hand, queries using function words (*the*, *of*, *to*) result in a larger number of index pages [34].

Finally, web corpora usually contain a very large number of relatively small documents. The Brown Corpus contains 500 documents. The BNC, being 100 times bigger in terms of word count, contains just 4,055 distinct documents, many of which are composite texts collected from entire issues of newspapers, journals or radio programmes. Given a small number of texts in traditional corpora it was feasible to annotate them with respect to genres while they were collected. On the other hand, the number of documents in web corpora is considerably larger, e.g., exceeding two million webpages for Web-as-Corpus projects developed at the University of Bologna [7, 33]. Thus, their manual annotation is practically impossible. Their genre composition is usually assessed indirectly by studying samples of their texts or by comparing the frequencies of keywords extracted from them (however, see Part III, this volume for a variety of methods for automatic classification of texts by genre).

There are at least three factors that can influence the distribution of genres in web-derived corpora:

- some genres are not well represented on the web;
- a large number of documents are located in the “hidden web”, which is not accessible to crawling;
- the process of corpus collection usually puts restrictions on file types retrieved from the web.

The web is an enormous resource, with more and more texts appearing there in a variety of languages. However, many genres are still underrepresented. This primarily concerns copyrighted work aimed at a wider public audience, such as fiction and non-fiction recreational reading. Their authors expect to receive royalties for their effort, and their publishers do not normally provide free online access. Texts in these genres do appear on the web, for instance, many amateur science-fiction authors regularly publish their works electronically under a Creative Commons licence, and Project Gutenberg collects out-of-copyright fiction. However, the selection available on the web is significantly skewed in comparison to offline fiction.

The hidden web (also called Deep Web) consists of pages that are difficult to access by crawling. Some of them are dynamically generated in response to a user query, e.g., some archived news items are stored in a database and can be retrieved only by specifying their date or keywords. Some hidden webpages are ordinary webpages which are not linked to any visible webpage, or which are accessible only by a password (not usually available to the crawler) or via a mechanism requiring some kind of user interaction, e.g., Javascript-based selection. Some estimates put the total size of the hidden web to be 500 times bigger than the surface web accessible to major search engines [41]. The hidden web is particularly important for search engines, as their aim is to index every possible webpage. This concern is less important for corpus collection, as a corpus is only a sample of the totality of texts in a given language. However, understanding the composition of the hidden web is important as it affects the distribution of genres. For instance, short descriptions of a large number of resources, such as synopses of books in a library, are more likely to be in the hidden web (accessible by queries to book names), so they are more likely to be underrepresented in web-derived corpora.

Finally, some file types are inherently easier to deal with. For instance, it is easy to retrieve plain text content from HTML pages, so HTML pages are more often used for corpus collection in comparison to, say, Word documents, which need special tools for retrieving textual content. PDF and Postscript files are commonly used on the web to present publishable information, such as books, articles or brochures. However, in terms of their internal format they contain a sequence of drawing primitives, often, but not necessarily, corresponding to characters, so that it is difficult to reconstruct the flow of text, spaces between words or even the encoding of non-Latin characters. The situation with Flash objects (normally containing animation, but often presenting a large amount of text) is even worse, as their drawing primitives include motion of respective objects across the computer screen. In the end, many formats apart from plain HTML files are often omitted from web-derived corpora, skewing their genre diversity. In the modern web this is especially important for PDF

files, which are the preferred format for final typeset products, such as catalogues, published research results or white papers. Often these texts are not available in the form of HTML files.

In summary, although web corpora are designed to contain examples of texts in exactly the same way as traditional corpora are, they are different in some respects and there is no consensus on many important aspects.

In addition to the construction issues outlined above, there are also other controversial issues related to formatting and cleaning webcorpora. In many cases traditional corpora were produced by scanning hard copies of texts and applying OCR (optical character recognition) to the result. In other cases, texts were typed in from scratch. In either case, traditional corpora do not preserve much information about formatting, with the only possible exception of paragraph boundaries. In the end, a text stored in a traditional corpus often consists of a flat sequence of sentences with little typographic information preserved.<sup>11</sup>

On the other hand, Web corpora coming from HTML pages contain relatively rich markup. As far as corpus collection is concerned, this markup takes three different forms:

1. navigation frames enabling navigation on a complex website (topics/subtopics, pages on related topics, calendar links, etc); and
2. text-internal hyperlinks, when running text is enriched with hypertextual markup linking to other relevant documents or other sections of the same document;
3. non-hypertextual markup, such as explicit formatting of headings, lists, tables, etc.

When webpages are collected to be used as a corpus for linguistic studies, one approach to corpus collection pays more attention to selecting running text. In this approach extra efforts are devoted to cleaning webpages from unwanted navigation frames [8]. The rationale behind this “cleaning” approach is to make web-derived corpora useful for research in natural language processing, lexicography or translation, because expressions frequently occurring in navigation frames, such as *Current events*, *See also* or *Have your say*, can considerably distort the language model. Similarly, text-internal links are often discarded, while their text remains, so that web corpora become more similar to their traditional counterparts.

Some portions of non-hypertextual markup in the form of headings and lists are often preserved in the cleaning approach, since deletion of this information again distorts the language model by introducing incomplete sentences within standard running text. Finally, some markup present in many webpages is used for presentational purposes only. For example, web designers often introduce table cells to separate different parts of text, e.g., navigation frames from the main body, or a new reply message in a forum from a quote from a previous message, whereas

---

<sup>11</sup> After collecting texts, developers of traditional corpora often introduce their own set of annotation layers, such as POS tagging, semantic or metatextual markup, but such layers are not taken from original texts in the form they have been published.

from the viewpoint of the content, such elements can be considered as distinct paragraphs. Therefore, the cleaning approach normally discards information about tables or replaces them with paragraph boundaries.

This approach to collecting and distributing webcorpora is useful in some respects, since it makes web-derived corpora closer to their offline counterparts. However, it discards a lot of information and makes the study of unique features of web genres more difficult. This also makes it harder to detect web genres automatically, as some crucial information for genre detection is present in the form of discarded features, e.g., navigation frames are more common in particular genres, and, similarly, documents of the same genre are often cross-linked. As a matter of fact, many genre collections built for classification purposes maintain original webpages in their entirety without attempting to clean them artificially (e.g. see the KI-04 corpus and the 7-webgenre collections described in Chapter 5 by Santini, this volume; see also the super-genre collection used in Chapter 10 by Lindemann and Littig, this volume).

In summary, at the current stage of genre research no standards have been agreed for the construction of web genre corpora. Decisions, choices and operationalisations are made subjectively, following individual needs. However, projects are put forward to establish shared standards (see Chapter 16 by Santini et al., the concluding chapter of this volume).

### ***1.3.3 Empirical and Computational Models of Web Genres***

The approach dominating automatic genre identification research is based on supervised machine learning, where each document is represented like a vector of features (a.k.a. the vector space approach), and a supervised algorithm (e.g. Support Vector Machines) automatically builds a genre classification model by “learning” from how a set of features “behave” in exemplar documents (e.g. see Chapter 7 by Sharoff; Chapter 6 by Kim and Ross, this volume). Many different feature sets have been tried out to date, e.g. function words, character n-grams, Parts of Speech (POS), POS tri-grams, Bag of Words (BOW), or syntactic chunks. Most of these feature sets have been tested on different genre corpora, differing in terms of number and nature of genres, and in terms of number of documents per genre. Although some comparative experiments have been carried out, the absence of genre benchmarks or reference corpora built with shared and agreed upon standards makes any comparison difficult, because existing genre collections have been built with subjective criteria, as pointed out in the previous section. A partial and temporary remedy to this situation has been adopted recently, i.e. cross-testing (see Chapter 5 by Santini, this volume).

Although the vector space approach is, for the time being, the most popular approach, in this last section of the open issues, we would like to outline a more complex view of web genres as source of inspiration and food for thought in future research. In Section 1.3.1, we suggested locating instances of web genres on, above and below the level of websites. The decision on this *manifestation level* belongs to

a series of related decisions which have to be made when it comes to modelling web genres. In this section, we briefly describe four of these decisions when the focus is on *structure*.

- *Deciding on the level of web genre units as the output objects of web genre classification:* Chapter 10 by Lindemann and Littig (this volume) present a model of web genre classification at what they call the *supergenre* level. This concerns a level of functional units which are composed of one or more genre level units. Interestingly, Lindemann and Littig consider websites as manifestation units of these supergenres. From that perspective we get the level of supergenres, of genres themselves and of subgenres as candidate output objects of a web-genre-related classification. Note that we may alternatively speak of macro, meso and micro (level) genres as has been done above. Conversely, Chapter 5 by Santini (this volume) and all approaches reviewed by her consider generic units of a comparative level of abstractness, but focus on web pages as their manifestation units. This divergence opens the possibility of a many-to-many relation between the output units of classification, i.e., the types which are attributed, and the input objects of classification, that is, the instances to which these types are attributed. Thus, by opting for some micro-, meso- or macro-level web genres one does not automatically determine the manifestation unit in the form of websites, web pages or page constituents. From that perspective, a decision space is created in which any location should be substantiated to keep replicability of the model and comparability with related approaches. By looking for what has been done towards such a systematisation we have to state that it is like weeding the garden, and that we are rather at the beginning.
- *Deciding on the level of manifestation units as the input objects of web genre classification:* the spectrum of this decision has already been outlined above.
- *Deciding on the features to be extracted from the input objects as reference values of classification:* when classifying input objects (e.g. web pages or sites) by attributing them to some output units (as elements of a certain genre palette), we need to explore certain features of the input objects. Among other things, we may explore *distinctive features* on the level of graphemes [46, 57], *linguistic features* in a more traditional sense [17, 38, 49, 80, 83, 86], *features related to non-hyperlink-based discourse structures* [19] or *structural features induced by hyperlinks* [16, 26, 57, 64]. In Section 1.3.1 we put special emphasis on less-frequently considered structure-related features of web genres. This is done according to the insight that they relate to an outstanding characteristic of genres on the *web*.
- *Deciding on the classifier model to be used to perform the classification:* facing complementary or even competing feature models as being inevitable in web genre modelling, composite classifiers which explore divergent feature resources have been common in web genre modelling from the beginning [45]. In line with this reasoning we may think of web genre models which simultaneously operate on nested levels of generic resolution. More specifically, we may distinguish *single-level* from *multi-level* approaches, which capture at least two levels

of web genre structuring: that is, approaches which attribute, for example, genre categories to websites subject to attributing subgenre categories to their elementary pages (other ways of defining *two-level* genre models can be found in Chapter 5 by Santini; Chapter 15 by Bruce, this volume). Note that the majority of approaches to web genre modelling realize single-level models by mapping web pages onto genre labels subject to one or more bag-of-features models. For this reason, multi-level approaches may be a starting point for building future models in this area.

By analogy to Biber [13] we may say that the structure of a web document correlates with its function, that is, with the genre it manifests. In other words: different genres have different functions, so that their instances are structured differently. As a consequence, the structure of a web document, whether a site, page or page segment, can be made a resource of feature extraction in web genre tagging. We summarise five approaches focussing on structure in the following list:

- *Bag-of-Structural-Features Approaches*: A classic approach to using structural features in hypertext categorisation is from Amitay et al. [1] – see Pirolli et al. [71] for an earlier approach in this line of research. Amongst others, Amitay et al. distinguish up, down, side and external links by exploring directory structures as manifested by URLs. They then count their frequencies as structure-related features. The idea is to arrive at a bag-of-structural features: that is, to analyse reference units whose frequencies are evaluated as dimensions of corresponding feature vectors. A comprehensive approach to using structure-related features in line with this approach is proposed by Lindemann and Littig [57].<sup>12</sup> They explore a wide range of features, similar to Amitay et al. [1], by including features which, amongst others, are based on the file format and the composition of the URL of the input pages. See also Kanaris and Stamatatos [46] who build a *bag of HTML tags* as one feature model of web genre classification (see Santini [80] for a comparative study of this and related approaches).

Generally speaking, linguistics has clarified the fundamental difference between explicit layout structure, implicit logical (document) structure and hidden semantic or functional structure [13, 10, 72]. From that perspective one does not assume, for example, that URL-based features are reliable indicators of logical web document structures. Rather, one has to assume – as is done by Lindemann and Littig [57] – an additional level of the manifestation of web genres, that is, their *physical storage* (including file format and directory structures). In any event, it is important to keep these structural levels apart as these are different resources for guessing the functional identity of a website. This can be exemplified by Amitay et al. [1] who introduce the notion of a *side link*, which exists between pages located in the same directory (cf. Eiron and McCurley [31] for a directory-based notion of up, down and side links). It is easy to construct

---

<sup>12</sup> See Lim et al. [56] for a study of the impact of different types of features including structural ones.

an example where a side link, which in terms of its physical storage manifests a *paratactic* link, is actually a *hypotactic* down or up link when being considered from the point of view of logical document structure [62]. Thus, any approach which explores structural features should clarify its standpoint regarding the difference of physical storage, layout and logical document structure.

- *Website-Tree- and Page-DOM-related Models*: A bag-of-structural-features approach straightforwardly adapts the bag-of-words approach of text categorisation by exploring the link and page structure of a site. This is an efficient and easy way to take web structure into account [57]. However, a more expressive and less abstract way to map this structure is to focus on the hierarchical *Document Object Model* (DOM) of the HTML representation of pages [28] or, additionally, on the mostly hierarchical kernel of the structure of a website [32]. Starting from the tree-like representation of a website, Ester et al. [32] build a Markov tree model which predicts the web genre  $C$  of a site according to the probability that the paths of this tree have been generated under the regime of  $C$ . Tian et al. [93] build a related model based on a hierarchical graph model in which the tree-like representation of websites consists of vertices which denote the DOM tree of their elementary pages. See Diligenti et al. [28], Frasconi et al. [35] and Raiko et al. [73] for related models of web document structures. See Chakrabarti [20] for an early model which explores DOM structure for hypertext categorisation (however with a focus on topical categorisations). Further, see Wisniewski et al. [95] for an approach to transforming DOM trees into semantically interpreted document models.
- *Beyond Hierarchical Document Models*: The preceding paragraph has presented approaches which start from tree-like models of web documents. This raises the question for approaches based on more expressive graph models. Such an alternative is proposed by Dehmer and Emmert-Streib [26]. Their basic idea is to use the page or site internal link structure to induce a so-called *generalised tree* from the kernel document structure, say, a DOM tree. The former is more informative than the latter as it additionally comprises up, down and lateral edges [63] which generalise the kernel tree into a graph. Note that this approach is powerful enough to represent page internal and external structures and, therefore, grasps a large amount of website structure. However, it maps structured data onto feature vectors which are input to classical approaches of vector-based classifications and, thus, departs from the track of Markov modelling. See Denoyer and Gallinari [27] who develop a Markov-related classifier of web document structures which, in principle, can handle *Directed Acyclic Graphs* (DAG). See alternatively Mehler [64] who develops a structure-based classifier of social ontologies as part of the Wikipedia. Extending the notion of a generalised tree, this model generalises the notion of a DAG in terms of *generalised nearly acyclic directed graphs* in order to get highly condensed representations of web-based ontologies with hundreds and thousands of vertices.
- *Two-level Approaches to Exploring Web Genre Structures*: The majority of approaches considered so far have been concerned with classifying units of web documents of a homogeneous nature – whether pages, their segments or complete

websites. This leaves plenty of room for considering approaches which perform, say, a generic categorisation of websites, subject to the categorisation of their elementary pages. Alternatively, we may proceed according to a feature-vector approach by representing a website by a composite vector as the result of aggregating the feature vectors of its pages (cf. the “superpage” approach of Ester et al. [32]). However, such an approach disregards the structure of a site because it represents it, once more, as a bag of features. Therefore, alternative models are required. Such an approach has been proposed by Kriegel and Schubert [50] with respect to topic-related classifications. They represent websites as vectors whose dimensions represent the topics of their pages so that the sites are classified subject to the classification of the pages. Mehler et al. [66] have shown that web genres may be manifested by whole sites, single pages or page segments. Facing this variety, the genre-related segmentation of pages and their fusion into units of the logical web document structure is an important step to grasping macro, meso and micro level units of web genres in a single model. Such a segmentation and fusion algorithm is proposed by Waltinger et al. [94] for web pages. The idea is to arrive at *monomorphic* segments as manifestations of generic units on the sub-genre level. This is done by segmenting pages using their visual depiction – as a byproduct this overcomes the tag abuse problem [6] which results from using HTML tags for manifesting layout as well as logical document structures. A paradigmatic approach to a two-level website classification which combines the multi-level manifestation perspective with a tree-like structure model is proposed by Tian et al. [93], who build a hierarchical graph whose vertices represent the DOM structure of the page constituents of the corresponding site.

- *Multi-Resource Approaches – Integrating Thematic with Structural Features:* Almost all approaches discussed so far focus on structural features. However, it is obvious that one must combine structural with content-related features by considering the structural position of content units within the input pages. See, for example, Joachims et al. [45] who study combined kernels trained on bag-of-words and bag-of-links models, respectively. See also Tian et al. [93] who integrate a topic model with a DOM-related classifier, with a focus on thematic classification.

In summary, as already suggested in Section 1.3.1, more focus on structure is needed to enhance web genre modelling in the future. We conjecture that a closer interaction between vector space approaches and structure-oriented methods can increase our understanding of web genres as a whole, thus providing a more realistic computational representation of genres on the web.

## 1.4 Conclusions

In this introduction, we emphasised why the study of genres on the web is important, and how empirical studies and computational models of web genres, with all their challenges, are the cutting edge of many research fields.

In our view, modern genre research is no longer confined to philosophical, literary and linguistic studies, although it can receive enlightenment from these disciplines. Undoubtedly, Aristotle, with his systematic classificatory mind, can still be considered the unquestioned initiator of genre studies in the Western World.<sup>13</sup> However, modern genre research transcends the manual and qualitative classification of texts on paper to become a meta-discipline that contributes to and delves into all the fields grounded in digital media, where quantitative studies of language, language technology, information and classification systems, as well as social sciences play an important role.

In this respect, this volume contributes to the current genre discussion in six ways:

1. It depicts the state of the art of genre research, presenting a wide range of conceptualisations of genre together with the most recent empirical findings.
2. It presents an overview of computational approaches to genre classification, including structural models.
3. It focuses on the notion of genres “for the web”, i.e., for the medium that is pervading all aspects of modern life.
4. It provides in-depth studies of several divergent genres on the web.
5. It points out several representational, computational and text-technological issues that are specific to the analysis of web documents.
6. Last but not least, it presents a number of intellectually challenging positions and approaches that, we hope, will stimulate and fertilise future genre research.

## 1.5 Outline of the Volume

Apart from the introduction, the volume is divided into four parts, each focussing on a specific facet of genre research.

PART II (*Identifying the Sources of Web Genres*) includes three chapters that analyse the selection and palettes of web genres from different perspectives.

Karlgren stresses how genre classes are both sociological constructs and stylistically observable objects, and how these two views can inform each other. He monitors genre variation and change by observing reader and author behaviour.

Crowston and co-workers report on a study to develop a “bottom-up” genre taxonomy. They collect a total of 767 (then reduced to 298) genre terms from 52 respondents (teachers, journalists and engineers) engaged in natural use of the Web.

Rosso and Haas propose three criteria for effective labels and report experimental findings based on 300 users.

---

<sup>13</sup> There are indeed many other scholars in other parts of the world, such as the Mao school in ancient China, who have pondered about the concept of genre.

PART III (*Automatic Web Genre Identification*) presents the state of the art in automatic genre identification based on the traditional vector space approach. This part includes chapters showing how automatic genre identification is needed in a wide range of disciplines, and can be achieved with a wide range of features.

In computational linguistics, Santini highlights the need for evaluating the generality and scalability of genre models. For this reason, she suggests using cross-testing techniques, while optimistically waiting for the construction of a genre reference corpus.

Kim and Ross present powerful features that perform well with a large number of genres, which have been selected for digital library applications.

In corpus linguistics, Sharoff is looking for a genre palette and genre model that can permit comparisons between traditional corpora and web corpora. He proposes seven functional genre categories that could be applied to virtually any text found on the Web.

Stein and co-workers present implementation aspects for a genre-enabled web search. They focus on the generalisation capability of web genre retrieval models, for which they propose new evaluation measures and a quantitative analysis.

Braslavski studies the effects of aggregating genre-related and text relevance rankings. His results show moderate positive effects, and encourage further research in this direction.

PART IV (*Structure-oriented Models of Web Genres*) focuses on genres at the website or network level, where structural information play a primary role.

Lindemann and Littig propose a vector-space approach for the automatic identification of super-genres at website level with excellent results.

Dehmer and Emmert-Streib discuss a graph-based perspective for automatically analysing web genre data by mining graph patterns representing web-based hyper-text structures. The contribution emphasises how an approach entirely different from the vector space model can be effective.

Björneborn outlines an exploratory empirical investigation of genre connectivity in an academic web space, i.e., how web page genres are connected by links. The pages are categorised into nine institutional and eight personal genre classes. The author builds a genre network graph to discuss changes in page genres and page topics along link paths.

PART V (*Case Studies of Web Genres*) focuses on the empirical observation of emerging web genres.

Paolillo and co-workers apply the social network approach to detect genre emergence in the amateur Flash community by observing social interaction. Their results indicate that participants' social network positions are strongly associated with the genres of Flash they produce, and this contributes to the establishment of genre norms.

Grieve and co-workers apply Biber's multi-dimensional analysis to investigate functional linguistic variation in Internet blogs, with the goal of identifying text types that are distinguished linguistically. Two main sub-types of blogs are identified: personal blogs and thematic blogs.

Bruce first reviews approaches to the notion of genre as a method of categorisation of written texts, leading to the presentation of a rationale for the dual approach

of social genre and cognitive genre. This approach is used to analyse 10 sample texts of the participatory journalism genre. The author concludes by saying that an adequate operationalisation of a genre as a category of written texts, including a web genre, should be able to account for the socially-constructed cognitive, organisational, and linguistic elements of genre knowledge.

The book ends with a view of possible future directions.

## References

1. Amitay, E., D. Carmel, A. Darlow, R. Lempel, and A. Soffer. 2003. The connectivity sonar: Detecting site functionality by structural patterns. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, 38–47. University of Nottingham, UK.
2. Andersen, J. 2008. The concept of genre in information studies. *Annual Review of Information Science & Technology* 42:339, 2007.
3. Andersen, J. 2008. Bringing genre into focus: Lis and genre between people, texts, activity and situation. *Bulletin of the American Society for Information Science and Technology* 34(5): 31–34.
4. Askehave, I., and A.E. Nielsen. 2005. Digital genres: A challenge to traditional genre theory. *Information Technology & People* 18(2):120–141.
5. Aston, G., and L. Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
6. Barnard, D.T., L. Burnard, S.J. DeRose, D.G. Durand, and C.M. Sperberg-McQueen. 1995. Lessons for the World Wide Web from the text encoding initiative. In *Proceedings of the 4th international World Wide Web conference "The Web Revolution"*. Boston, MA.
7. Baroni, M., and A. Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Companion Volume to Proceedings of the European Association of Computational Linguistics*, 87–90. Trento.
8. Baroni, M., F. Chantree, A. Kilgarriff, and S. Sharoff. 2008. Cleaneval: A competition for cleaning web pages. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech.
9. Bateman, J.A. 2008. *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*. London: Palgrave Macmillan.
10. Bateman, J.A., T. Kamps, J. Kleinz, and K. Reichenberger. 2001. Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics* 27(3):409–449.
11. Biber, D. 1988. *Variation across speech and writing*. Cambridge, MA: Cambridge University Press.
12. Biber, D. 1989. A typology of English texts. *Linguistics* 27(3):43–58.
13. Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, MA: Cambridge University Press.
14. Biber, D., U. Connor, and T.A. Upton. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: Benjamins.
15. Björneborn, L. 2004. Small-world link structures across an academic web space: A library and information science approach. PhD thesis, Royal School of Library and Information Science, Department of Information Studies, Denmark.
16. Björneborn, L. 2010. Genre connectivity and genre drift in a web of genres. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
17. Braslavski, P. 2010. Marrying relevance and genre rankings: An exploratory study. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.

18. Bruce, I. 2008. *Academic writing and genre: A systematic analysis*. London: Continuum.
19. Bruce, I. 2010. Evolving genres in online domains: The hybrid genre of the participatory news article. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
20. Chakrabarti, S. 2001. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proceedings of the 10th International World Wide Web Conference*, May 1–5, 211–220. Hong Kong.
21. Chakrabarti, S., M. van den Berg, and B. Dom. 1999. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the 8th International World Wide Web Conference*. Toronto, ON.
22. Chakrabarti, S., M. Joshi, K. Punera, and D.M. Pennock. 2002. The structure of broad topics on the web. In *Proceedings of the 11th International World Wide Web Conference*, 251–262. New York, NY: ACM Press.
23. Cohn, D.A., and T. Hofmann. 2000. The missing link – a probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS)*, eds. T.K. Leen, T.G. Dietterich, and V. Tresp, 430–436. Denver, CO: MIT Press.
24. Condamines, A. 2008. Taking genre into account when analysing conceptual relation patterns. *Corpora* 3(2):115–140.
25. Craven, M., D. DiPasquo, D. Freitag, A.K. McCallum, T.M. Mitchell, K. Nigam, and S. Slattery. 2000. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence* 118(1–2):69–113.
26. Dehmer, M., and F. Emmert-Streib. 2010. Mining graph patterns in web-based systems: A conceptual view. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
27. Denoyer, L., and P. Gallinari. 2004. Un modèle de mixture de modèles génératifs pour les documents structurés multimédias. *Document numérique* 8(3):35–54.
28. Diligenti, M., M. Gori, M. Maggini, and F. Scarselli. 2001. Classification of HTML documents by hidden tree-markov models. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 849–853. Seattle, WA.
29. Dillon, A. 2008. Bringing genre into focus: Why information has shape. *Bulletin of the American Society for Information Science and Technology* 34(5):17–19.
30. Donato, D., L. Laura, S. Leonardi, and S. Millozzi. 2007. The web as a graph: How far we are. *ACM Transactions on Internet Technology* 7(1):4.
31. Eiron, N., and K.S. McCurley. 2003. Untangling compound documents on the web. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, 85–94. Nottingham.
32. Ester, M., H.-P. Kriegel, and M. Schubert. 2002. Web site mining: A new way to spot competitors, customers and suppliers in the world wide web. In *KDD '02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 249–258. New York, NY: ACM Press.
33. Ferraresi, A., E. Zanchetta, S. Bernardini, and M. Baroni. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *The 4th Web as Corpus Workshop: Can We Beat Google? (At LREC 2008)*. Marrakech.
34. Fletcher, W.H. 2004. Making the web more useful as a source for linguistic corpora. In *Corpus linguistics in North America 2002: Selections from the 4th North American Symposium of the American Association for applied corpus linguistics*, eds. U. Connor, and T. Upton. Editions Rodopi: Amsterdam/New York.
35. Frasconi, P., G. Soda, and A. Vullo. 2002. Hidden Markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems* 18(2–3):195–217.
36. Freund, L. 2008. Exploiting task-document relationships to support information retrieval in the workplace. PhD thesis, University of Toronto.

37. Freund, L., and C. Nilsen. 2008. Assessing a genre-based approach to online government information. In *Proceedings of the 36th Annual Conference of the Canadian Association for Information Science (CAIS)*. University of British Columbia, Vancouver.
38. Grieve, J., D. Biber, E. Friginal, and T. Nekrasova. 2010. Variation among blogs: A multi-dimensional analysis. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
39. Gunnarsson, M. 2010. Classification along genre dimensions. PhD, Inst. f. Biblioteks- och Informationsvetenskap, Göteborgs Universitet.
40. Gupta, S., H. Becker, G. Kaiser, and S. Stolfo. 2006. Verifying genre-based clustering approach to content extraction. In *Proceedings of the 15th International Conference on World Wide Web*, 875–876. New York, NY: ACM Press.
41. He, B., M. Patel, Z. Zhang, and K. Chen-Chuan Chang. 2007. Accessing the deep web: A survey. *Communications of the ACM* 50(2):94–101.
42. Herring, S.C., I. Kouper, J.C. Paolillo, L.A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu. 2005. Conversations in the blogosphere: An analysis “from the bottom up”. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS’05)*. Big Island, Hawaii.
43. Heyd, T. 2008. *Email hoaxes: Form, function, genre ecology*. Amsterdam: Benjamins.
44. Ide, N., R. Reppen, and K. Suderman. 2002. The American National Corpus: More than the Web can provide. In *Proceedings of the 3rd Language Resources and Evaluation Conference*, 839–844. Las Palmas.
45. Joachims, T., N. Cristianini, and J. Shawe-Taylor. 2001. Composite kernels for hypertext categorisation. In *Proceedings of the 11th International Conference on Machine Learning*, 250–257. San Francisco, CA: Morgan Kaufmann.
46. Kanaris, I., and E. Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’07)*, Washington, DC: IEEE Computer Society.
47. Karlgren, J., and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, vol. 2, 1071–1075. Kyoto.
48. Kessler, B., G. Nunberg, and H. Schütze. 1997. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. 32–38. Madrid, Spain.
49. Kim, Y., and S. Ross. 2010. Formulating representative features with respect to genre classification. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
50. Kriegel, H.-P., and M. Schubert. 2004. Classification of websites as sets of feature vectors. In *Databases and applications*, ed. M.H. Hamza, 127–132. Anaheim, CA: IASTED/ACTA Press.
51. Kucera, H., and W.N. Francis. 1967. *Computational analysis of presentday American English*. Providence, RI: Brown University Press.
52. Kumar, R., J. Novak, P. Raghavan, and A. Tomkins. 2004. Structure and evolution of blogspace. *Communications of the ACM* 47(12):35–39.
53. Lee, D. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3): 37–72.
54. Li, W.-S., O. Kolak, Q. Vu, and H. Takano. 2000. Defining logical domains in a web site. In *Proceedings of the 11th ACM on Hypertext and Hypermedia*, 123–132. San Antonio, TX.
55. Li, W.-S., K.S. Candan, Q. Vu, and D. Agrawal. 2002. Query relaxation by structure and semantics for retrieval of logical web documents. *IEEE Transactions on Knowledge and Data Engineering* 14(4):768–791.

56. Lim, C.S., K.J. Lee, and G.C. Kim. 2005. Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management* 41(5):1263–1276.
57. Lindemann, C., and L. Littig. 2010. Classification of web sites at super-genre level. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
58. Marshman, E., M.-C. L’Homme, and V. Surtees. 2008. Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora* 3(2):141–172.
59. Martin, J.R. 1994. Macro-genres: The ecology of the page. *Network* 21: 29–52.
60. Martin, J.R., and D. Rose. 2008. *Genre relations: Mapping culture*. London & Oakland: Equinox Pub.
61. Mehler, A. 2008. Structural similarities of complex networks: A computational model by example of wiki graphs. *Applied Artificial Intelligence* 22(7&8):619–683.
62. Mehler, A. 2010. Structure formation in the web. A graph-theoretical model of hypertext types. In *Linguistic modeling of information and markup languages. Contributions to language technology*, eds. A. Witt and D. Metzger, Text, Speech and Language Technology, 225–247. Dordrecht: Springer.
63. Mehler, A. 2009b. Generalised shortest paths trees: A novel graph class applied to semiotic networks. In *Analysis of complex networks: From biology to linguistics*, eds. M. Dehmer and F. Emmert-Streib. Weinheim: Wiley-VCH.
64. Mehler, A. 2010. A quantitative graph model of social ontologies by example of Wikipedia. In *Towards an information theory of complex networks: Statistical methods and applications*, eds. M. Dehmer, F. Emmert-Streib, and A. Mehler. Boston, MA/Basel: Birkhäuser.
65. Mehler, A., M. Dehmer, and R. Gleim. 2006. Towards logical hypertext structure: A graph-theoretic perspective. In *Proceedings of the 4th International Workshop on Innovative Internet Computing Systems (I2CS ’04)*, eds. T. Böhme and G. Heyer, Lecture Notes in Computer Science, vol. 3473, 136–150. Berlin/New York, NY: Springer.
66. Mehler, A., R. Gleim, and A. Wegner. 2007. Structural uncertainty of hypertext types. An empirical study. In *Proceedings of the Workshop “Towards Genre-Enabled Search Engines: The Impact of NLP”*, September, 30, 2007, in Conjunction with RANLP 2007, 13–19. Borovets, Bulgaria.
67. Menczer, F. 2004. Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology* 55(14):1261–1269.
68. Montesi, M., and T. Navarrete. 2008. Classifying web genres in context: A case study documenting the web genres used by a software engineer. *Information Processing and Management* 44:1410–1430.
69. Ounis, I., M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. 2006. Overview of the trec 2006 blog track. In *Proceedings of the Text Retrieval Conference (TREC)*. NIST.
70. Päiväranta, T., M. Shepherd, L. Svensson, and M. Rossi. 2008. A special issue editorial. *Scandinavian Journal of Information Systems* 20(1).
71. Pirolli, P., J. Pitkow, and R. Rao. 1996. Silk from a sow’s ear: Extracting usable structures from the web. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing*, 118–125. New York, NY: ACM Press.
72. Power, R., D. Scott, and N. Bouayad-Agha. 2003. Document structure. *Computational Linguistics* 29(2):211–260.
73. Raiko, T., K. Kersting, J. Karhunen, and L. de Raedt. 2002. Bayesian learning of logical hidden Markov models. In *Proceedings of the Finnish AI Conference (STeP-2002)*, 64–71. Finland.
74. Rehm, G. 2002. Towards automatic web genre identification – A corpus-based approach in the domain of academia by example of the academic’s personal homepage. In *Proceedings of the Hawaii International Conference on System Sciences*. Big Island, Hawaii.
75. Rehm, G. 2010. Hypertext types and markup languages. The relationship between HTML and web genres. In *Linguistic Modeling of Information and Markup Languages. Contributions to*

- Language Technology*, eds. A. Witt and D. Metzger, Text, Speech and Language Technology, 143–164. Dordrecht: Springer.
76. Rosso, M.A. 2008. Bringing genre into focus: Stalking the wild web genre (with apologies to euell gibbons). *Bulletin of the American Society for Information Science and Technology* 34(5):20–22.
  77. Rosso, M.A., and S.W. Haas. 2010. Identification of web genres by user warrant. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
  78. Santini, M. 2007a. Characterizing genres of web pages: Genre hybridism and individualization. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. Big Island, Hawaii.
  79. Santini, M. 2007b. Automatic identification of genre in Web pages. PhD thesis, University of Brighton, Brighton.
  80. Santini, M. 2010. Cross-testing a genre classification model for the web. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
  81. Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working Papers on the Web as Corpus*, eds. M. Baroni and S. Bernardini, 63–68. Bologna: Gedit.
  82. Sharoff, S. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.
  83. Sharoff, S. 2010. In the garden and in the jungle. Comparing genres in the bnc and internet. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
  84. Sinclair, J. ed. 1987. *Looking up: An account of the COBUILD project in lexical computing*. London and Glasgow: Collins.
  85. Sinclair, J. 2003. Corpora for lexicography. In ed. P. van Sterkenberg, *A practical guide to lexicography*, 167–178. Amsterdam: Benjamins.
  86. Stein, B., S. Meyer zu Eissen, and N. Lipka. 2010. Web genre analysis: Use cases, retrieval models, and implementation issues. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
  87. Stewart, J.G. 2008. Genre oriented summarization. PhD thesis, Carnegie Mellon University.
  88. Sun, A., and E.-P. Lim. 2003. Web unit mining: Finding and classifying subgraphs of web pages. In *CIKM '03: Proceedings of the 12th International Conference on Information and Knowledge Management*, 108–115, New York, NY: ACM Press.
  89. Swales, J.M. 1990. *Genre analysis: English in academic and research settings*. Cambridge, MA: Cambridge University Press.
  90. Tajima, K., Y. Mizuuchi, M. Kitagawa, and K. Tanaka. 1998. Cut as a querying unit for WWW, netnews, e-mail. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 235–244. New York, NY: ACM Press.
  91. Tajima, K., and K. Tanaka. 1999. New techniques for the discovery of logical documents in web. In *International Symposium on Database Applications in Non-traditional Environments*. IEEE, 125–132.
  92. Thelwall, M., L. Vaughan, and L. Björneborn. 2006. Webometrics. *Annual Review of Information Science Technology* 6(8):81–135.
  93. Tian, Y.H., T.J. Huang, W. Gao, J. Cheng, and P. Bo Kang. 2003. Two-phase web site classification based on hidden Markov tree models. In *WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*. IEEE Computer Society, 227, Washington, DC.
  94. Waltinger, U., A. Mehler, and A. Wegner. 2009. A two-level approach to web genre classification. In *Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST '09)*, March 23–26, 2007. Lisboa.

95. Wisniewski, G., F. Maes, L. Denoyer, and P. Gallinari. 2007. Modèle probabiliste pour l'extraction de structures dans les documents web. *Document numérique*, 10(1):151–170.
96. Wodak, R. 2008. Introduction: Discourse studies – important concepts and terms. In *Qualitative Discourse Analysis in the Social Sciences*, eds. Wodak, R. and Krzyzanowski, M., 1–29. Palgrave.
97. Yates, S.J., and T.R. Sumner. 1997. Digital genres and the new burden of fixity. In *Proceedings of the 30th Hawaii International Conference on System Sciences*, vol. 6. Maui, HI.



**Part II**  
**Identifying the Sources of Web Genres**

# Chapter 2

## Conventions and Mutual Expectations

### Understanding Sources for Web Genres

Jussi Karlgren

#### 2.1 Genres Are Not Rule-Bound

A useful starting point for genre analysis is viewing genres as artifacts.<sup>1</sup> Genres are *instrumental categories*, useful for author and reader alike in forming the understanding of a text and in providing the appropriate intellectual context for information acquired through it. Genre distinctions are observable in terms of whom a text<sup>2</sup> is directed to, how it is put together, made up, and presented.

Recognising genres or detecting differences between genres is typically done by identifying *stylistic differences* with respect to any number of surface characteristics: presence or preponderance of linguistic items, treatment of topical entities, organisation of informational flow, layout characteristics, etcetera. This type of stylistic or non-topical variation can be observed on many levels, and is by no means orthogonal or independent to topical variation – quite to the contrary, it shows strong dependence on subject matter as well as on expected audience and many other contextual characteristics of the communicative situation. Given a communicative situation, systematical and predictable choices made by the author with respect to possible stylistic variation eases the task of the reader and helps organise the discourse appropriately for the conceivable tasks at hand. Guidelines for stylistic deliberation can similarly function as a support for authors, as an aid for making some of the many choices facing an author: giving defaults where no obvious alternatives are known and granting preferences where many alternatives seem equivalent.

Genres need not and cannot be understood without understanding their place in communication, in terms of usefulness for readers and authors. Their function is to act as a frame for informing and conceptualising the communication at hand. Their utility for us as providers of new technology or researchers in human communicative behaviour is that of a tool for describing instances of behaviour in an appropriate

---

J. Karlgren (✉)  
Swedish Institute of Computer Science (SICS), Stockholm, Sweden  
e-mail: jussi@sics.se

<sup>1</sup> Cf. Santini: "... cultural objects created to meet and streamline communicative needs" [12].

<sup>2</sup> Or other information object: the obvious generalisations are to be assumed in the following.

bundle, in appropriate chunking of reality. Genre is a vague but well-established notion, and genres are explicitly identified and discussed by language users even while they may be difficult to encode and put to practical use. While we as readers are good at the task of distinguishing genres, we have not been intellectually trained to do so and we have a lack of meta-level understanding of how we proceed in the task. To use reader impressions profitably in further research, they must be interpreted or analysed further in some way.

In recent years, genre analysis has been extended to typologies of communicative situations beyond the purely textual [e.g. 1, 15], with genres ranging much further than the analysis of simple linguistic items: the range of possible human communicative activities is much wider than the range of possible texts. The demarcation of genre to other variants of categorisation may be difficult – how does it relate to individual, author-conditioned, variation or to topically dependent variation in textual character? One perspective – *genre as a characteristic of text* – would be to understand genre as yet another variant dimension of textual variation, in addition to topical variation, individual variation, temporal, and even stylistic non-genre variation. Another – *text as instance of a communicative genre* – is to understand it as a dimension of variation or a categorisation scheme on another level of abstraction, encompassing topic, individual variation, and other facets of textual variation that can be observed.

The latter view underlies the explorative studies given in this chapter, taking as a starting point the view that genres have a reality in their own right, not only as containers or carriers of textual characteristics. Genres have utility and a purpose in that they aid the reader in understanding the communicative aims of the author; they provide a framework within which the author is allowed to make assumptions on the competence, interest, and likely effort invested by the reader. On the contrary, human communicative spheres can be understood to establish their conventions as to how communicative action can be performed. When these conventions bundle and aggregate into a coherent and consistent socially formed entity, they guide and constrain the space of potential communicative expressions and form a genre, providing defaults where choices are overwhelming and constraints on choices should be given.

Most importantly, those of us who work with identifying genre-based variation must keep in mind that genres are not a function of stylistic variation. However solid stylistically homogenous groupings of information items we might encounter, if they do not have a functional explanation, they are not a genre, and the intellectually attractive triple {*content, form, function*} is misleading in its simplicity: two categories of text are not different genres solely by virtue of difference in form. It may be a requirement, for practical purposes, that the genre can be algorithmically and computationally recognized (see e.g. Chapter 3 by Rosso and Haas, this volume), but the quality of being discriminable is not, by itself, sufficient to label a category of texts a genre. When the character of text in a typical communicative situation is formed by or based on the bidirectional flow of authors' expectations on their audiences and that or those audiences' expectations on likely behaviour on the part of the authors they are reading, those items, or that family of items, constitute a genre.

While the vagueness of the term “genre” may be vexing to the language technologist seeking to formulate recognition algorithms, it poses no difficulty to the individual reader or consumer – genres are readily and intuitively understood and utilised by most people. This does not mean that readers find it easy to make sharp and disjoint categorisations of information objects presented to them: genres appear, overlap, evolve, and fall into disuse but are not regularly defined nor delimited against each other [e.g. 12]. Prototypical central examples, rather than borderline definitions are the best ways of describing genres.

Genres are recognisable and identifiable by their readers on several levels: conventions range from the abstract and general to the detailed and concrete, from spelling conventions to pragmatic conventions to informational organisation of the discourse. Differences between situations can be very fine-grained: Sports news items have a different character than business news items in spite of many extraneous similarities. On certain levels of linguistic practice prescriptive rule systems are less dominant than others, and allow for more genre-bound practice to emerge; on others, grammar rules or general conventions govern the decision space.

Stylistic variation can be governed by more or less fixed rules and conventions, but is with regard to many observable features open to individual or idiosyncratic choice. The span between individual variation and genre-bound convention has not been systematically probed in stylo-statistic studies, but some indications can be found that the differences between the two can be described to some extent in terms of feature variation [8] and the underlying mechanisms of divergence from genre standards can be described as one of individualisation [13] in genre systems where conventions are weak.

What then, influences the life cycle of genres on new media such as those that can be found on the web? How might we find traces of what expectations readers and web users have on the material they encounter? What motivates the emergence of new genres and new stylistic techniques on the web? This chapter will by giving three examples of simple explorative studies which examine the interface between text and convention, discussing in turn the experience of readers of web information, the editorial effort of commercial information specialists at the Yahoo! directory, and observations from search logs. None of the studies are intended to provide the last word on analysis of reader experience, librarianship or searcher behaviour, but they all contribute to an understanding of genre as a carrier of convention, agreed upon by author and reader community in concert.

## 2.2 So, Let’s Ask the Readers

A useful and frequently utilised resource to better understand web genre is to turn to web users and readers. Readers (with the obvious generalisations to usage of other media) can provide information about their understanding of genres in a variety of ways. As noted in Chapter 4 by Crowston et al., is volume, genres only exist in use. How information resources are used – assessed, read, viewed, examined, cited, recycled – varies from type of resource to type of resource and user group to user group. Genre defining differences in use can be established through observing

user actions, asking users to behave as they would normally and observing them in a near-realistic situation e.g. in a think-aloud study (again, as in Chapter 4 by Crowston et al., this volume), by modelling links between users in various ways (as Chapter 13 by Paolillo et al., this volume) or asking them explicitly to rate samples. Many of these methods are labour-intensive for the researchers – setting up the study conditions and interpreting the results require a significant effort on the part of the researcher.

An alternative – less demanding, but obviously less controlled – method is using questionnaires, allowing users to formulate their impressions of what genres they are aware of [e.g. 11]. In April and May of 1997 we used an e-mail questionnaire to solicit responses from engineering students to the question of what genres they thought were available on the internet – the study was published in 1998, the following January [5]. The students were not given any tutoring on what would constitute a genre – the intention was to gain an understanding of what categories inform the behaviour of information technology users at the time. The study was sent to all active students at the time, and we received 67 responses, a rate of just over 10%, which we thought disappointing at first. We later accepted the fact that thinking about genre constitutes a challenge for most readers – genres are accepted as an unobtrusive aspect of reading, not as an abstract quality open for discussion and deliberation.

We found that answers ranged from very short to extensive discussions. Some of the answers given are shown in Table 2.1 which is excerpted from the 1998 report. Many readers conflated genre and form on the one hand with content and topic on the other: “tourism”, “sports”, “games”, “adult pages”; many (but not all) used paper genres as models for the analysis of web genres; many referred to the intention of the of the information provider showed up as a genre formation criterion in several responses: “here I am”, “sales pitches”, “serious material”; or, as an alternative formulation of the same criterion, the type of author or source of information: “commercial info”, “public info”, “non-governmental organisation info”; or intended usage environment or text ecology: “public documents”, “internal documents”, “personal documents”; many explicitly mentioned quality of the information as a categorisation criterion: “boring home pages”.

This last aspect was especially gratifying for the purposes of the study at the time, since it was motivated by the desire to build a better search engine. The conclusions of the study were that internet users have a vague but useful sense of genres among the documents they retrieve and read. The impressions users have of genre can be elicited and to some extent formalised enough for automatic genre collection. The names of genres in an information retrieval setting should be judiciously chosen to be on an appropriate level of abstraction so that mismatches will not faze readers. This, of course, is a non-trivial intellectual and editorial effort and involves interpretation and judgments on audience, readership habits, and textual qualities.

The results of the 1997–1998 study provided the genre palette given in Table 2.2 which was later used to construct a genre-aware front end to a search engine, which was evaluated in separate studies – while the genre palette met with the approval of users then it obviously reflects the usage and standards of its time. The usage and

**Table 2.1** Some translated excerpts of response to the 1998 study

---

Science, entertainment, information
Here I am, sales pitches, serious material
Home pages
Data bases
Guest books
Comics
Pornography
FAQs
Search pages
Reference materials
Home pages
Public info
Non-government organisation info
Search info
Corporate info
Informative advertisements
Non-informative advertisements
Economic info
Tourism, Sports, Games
Adult pages
Science, Culture, Language
Media
Public documents, internal documents, personal documents
“Check out what a flashy page I can code”
“I guess we have to be on the net too”

---

**Table 2.2** The genre palette, as given by the 1998 study

---

Informal, private
Personal home pages
Public, commercial
Home pages for the general public
Searchable indices
Pages with feed-back: customer dialogue; searchable indexes
Journalistic materials
Press: news, reportage, editorials, reviews, popular reporting, e-zines
Reports
Scientific, legal, and public materials; formal text
Other running text
FAQs
Link collections
Other listings and tables
Asynchronous multi-party correspondence
Contributions to discussions, requests, comments; Usenet news materials
Error messages

---

the expectations of users are necessarily formed by their backgrounds and by the rapidly changing technological infrastructure.

A similar questionnaire was again sent to engineering students and non-technical mailing list participants in various subjects in January 2008. The answer rate was

again on the order of 10%, giving us 31 answers. The answers from the two studies were quite similar as to their content, with notable addition of social and networking sites as a new genre, and shopping sites, neither which were in much evidence 10 years earlier. A sample of the answers in high-level categories is given in Table 2.2.

We find here, 10 years later, the same mix of quality, source, content and similar concerns as in the first study: games, news, erotica feature prominently as genre labels – in some cases with subcategorisation – as well as the distinction between *commercial* and *non-commercial* sites which cuts through most answers (Table 2.3). In the previous study, responses centered on the distinction as if it were clear-cut, in this second edition the responses reflect the fact that there are marketing pages that may not appear to be marketing at first glance: “viral marketing” or “sham games” designed to lead the game player to link farms rather than to entertain.

The model web users appear to have for information on the web centers on the function of the pages, best summarised by the response of one respondent “I classify the internet in two top categories. One is information, the other non-information. . . . For me there are only two kinds of genres on the internet.”

The new distinctions we find are firstly made between *media of different types* – which reflects the new technology available for the web users of today, and secondly of *specific services* developed during the period: “downloads”, social sites, and user-contributed information. Some genres have acquired a more concrete presence – while there were sites dedicated to computer mediated communication, journal keeping, and on-line discussions in year 1997, they catapulted to public awareness and achieved an accepted status as a medium of communication only after a broader wave of uptake occurred around year 1999, when the term “blog” first came into use. Others are less salient. While “radio” and “video” were mentioned, no respondent mentioned “Tv”.

There are numerous more mentions of *special interests and special topics* – reflecting the appearance of more information, not limited to technology. A more cross-cutting distinction made explicit by relatively few of the respondents but which is inferrable in several of the responses is that of *temporality* or timeliness – pages that change, versus pages that stay the way they are: “. . . home pages where the text doesn’t change radically over time.” and *pages without interactivity!*. These are characteristics of pages with direct ramifications on the usefulness and usage of it.

There is clearly a limit to the usefulness of questionnaire or other user-elicitation methods for understanding web genre. Firstly, the respondents are bound to their personal perspective and experiences, which may not be general or even generalisable. Their responses are biased towards the function and the source of the information they usually peruse. To help formulate the distinctions we wish to make or to capture the generalities we wish to work with, we will need a large number of users from various walks of life. We will then face a daunting task of bringing order to the responses given by them. Secondly, the web, the information items which form it, and thus the communicative situations they engender are fluid and ill-defined. Data sources of various types and services are compared to reproduced traditional media of, again, various types and sources. The publishing threshold is

**Table 2.3** Selection of categorised answers given in the reproduced study*Conversations*

Personal blogs, “Serious” blogs, discussion lists

*Social networks*

General, niched to special interest groups

*User-contributed data*

Tagging, media

*Static pages*

*Commercial:* Products, sales, corporate, advertising, viral advertising, sham games

*Non-commercial:* “Old fashioned home pages”, academia, technology, programming, standards, technical documentation

*Special interest information:* Encyclopedias, wikipedia, learning, topical, schools, FAQ (only one mention!), music, lyrics, automobiles, religion, sports, fashion, travel, retro/history, geography, stats, photography, recipes, comics

*Advice:* How-to, DIY, health (self-diagnosis, hypochondria)

*Propaganda:* Activists, nuts

*Portals news*

Newspapers, gossip, web news sites, radio, video

*Services and web applications*

General search engines, niche search engines

Games

Buy-and-sell, downloads (Legal, illegal, torrent pages, clearinghouses),

office sites, webmail, price comparison sites, banks, tests, diagnostics, databases

low, leading to a large amount of material in imperfect states of publication and thence to questions of versioning. The inclusion of e.g. database reports and similar dynamic services on the web can lead to a discussion of whether web material which is compiled and served on demand is a document or a service; passage retrieval, data mining, summarisation, and extraction services will compose documents out of raw materials that may not have existed before they were demanded. The ease of including supporting extraneous materials in documents may in some sense change their genre; the possibility of splitting material into several documents for convenience of use might lead to a coherent whole being experienced as several items of different style. The advent of new types of services and innovations will make intractable the formulation any stable genre typology in the near future [cf. 14].

Thus, any genre palette established by survey studies of this type will encounter overlaps, contradictions, and imperfect definitions among the views expressed by the respondents. The results will show change, may capture evolution, and may inform us better of which features can characterise a genre and which cannot. General lessons found from the responses given to these two studies are

- Previous important distinction between commercial vs. non-commercial has blurred, both from the introduction of useful commercial services, and from the advent of adversarial and viral marketing mechanisms.
- The previously mentioned fairly simple category of “interactive forms” has developed into a large space of services of various types.

- New mechanisms of computer-mediated communication tools and publishing platforms have given rise to genres of communication, outside previous classification schemata.
- Previous static or approximatively static pages are now by users distinguished per their temporal and dynamic qualities.
- Previous preponderance of technical topics has given way to numerous niche or special topics, often viewed as separate genres by readers.
- User needs, formulated as quality still is main criterion for classification.

Forming the understanding of these potentially new genres is to a large extent a fairly demanding intellectual and editorial task. Explorative qualitative studies such as the ones briefly presented are a useful basis for such tasks, but do not in themselves build new knowledge: the knowledge is built from the refinement and structuring of reader impressions. Eliciting such impressions can be done methodologically much more stringently – as has been done and discussed in several recently published studies and discussions [e.g. 6]. These studies are intended to demonstrate that the readership is conscious of genre, and that readers expect genre to be based on both their previous experiences and on technological developments.

New media and modes of communication bridge synchronous highly interactive spoken communication and less interactive and asynchronous written communication modes [7]. By blurring the distinction between spoken and written, new media and new types of communicative situations are created. New forms of communication, while initially patterned on traditional, established, and well-conventionalised genres (such as e.g. the genre “Poetry” identified by the subjects studied in Chapter 3 by Rosso and Haas, this volume) gradually evolve new conventions, new stylistic and formal characteristics and eventually emerge as genres in their own right. Identifying new genres require understanding of what characteristics occasioned their emergence: not only determined as combinations of “... observable physical and linguistic features” [16] but of additional features of function, interactional characteristics [14] and, based on responses such as the ones presented above, *temporal* qualities of the information.

### 2.3 An Editorial, Third Party, View of Genres on the Web

A slightly more external source of information on understanding genres is that of an editorial board, attempting to organise information produced by some for the use of some others. An example of such an effort is the Yahoo! directory, which has been one of the most visited web information resources since 1994 when the directory first was launched. During this time, the user base and the content of the web itself has grown and changed from a primarily technology- and engineering-oriented tool to a communication system for the general public. Examining the make-up of some of the categories in the Yahoo! directory we find that most of the categories correspond well to established paper genres, many or even most primarily topical. The top level of the directory, together with four of its subcategories are examined to

**Table 2.4** Categories in the Yahoo! directory

<code>dir.yahoo.com</code>	Top	Entertainment	Reference	Health	News and media
Subcategories 2008	14	38	41	49	69
Web-related					
subcategories 2008	1	6	4	2	3
Changes 2000–2008	–	+7, –5	+3, –2	+2	+5, –4

find if the categories are web-specific, and whether they have potential to be called “genres” in any realistic sense of the term. The data are collected in February 2008, and a comparison with the status of the directory in year 2000 is done by retrieving a version of the respective pages from a web archive<sup>3</sup> which collects and stores periodic snapshots of the web.

An overview over the categories examined is given in Table 2.4.<sup>4</sup> In the top level directory, none of the 14 top level subcategories have been changed since 2000. This reflects well on the stability of the categories, which are all well-established from traditional libraries and document collections, recognised by readers and information specialists alike. Labels such as “Reference”, “Education”, and “Science” are well anchored in everyday experience of printed matter, and conform to some extent with our sense of genre. The only web-specific subcategory is that of “Computers and Internet” which would most likely not be promoted to top level in a paper based library, but neither will it merit being called a genre in the sense discussed in this volume. Most of the Yahoo! categories have web-specific subcategories related to web activities such as “Web directories” or “Searching the web” and interactive and communicative subcategories such as “Blogs” (renamed from “Weblogs” in 2007) and “Chats and forums” or “Ask an expert”. These are familiar to us from other studies of web genres and web services.

In addition to these we find categories of web-specific information, such as “FAQs” and “How-to guides” under “Reference”. These are written knowledge sources, written by internet users of some expertise for other internet users, and are a direct product of the lowered publishing threshold afforded by information technology and web publishing. They resemble traditional engineering notes, and have rapidly gained take-up in non-engineering fields and are a clear emerging genre with distinctive linguistic characteristics. We also find “Entertainment” categories such as “Randomized Things”, “Webisodes” and “X of the Day, Week, etc” which are based on internet technology. Most interestingly, we find some category churn motivated by technology in that the “News and Media” category has lost the subcategory “Personalised News” during the past few years and that the “Entertainment” category has lost its subcategories “Cool links” and “Virtual Cards”.

<sup>3</sup> The Wayback Machine – <http://web.archive.org>

<sup>4</sup> Of terminological but somewhat tangential interest for this examination is the subcategory “Genres” under “Entertainment”, which consists only of the four entertainment subsubcategories “Comedy”, “Horror”, “Mystery”, “Science Fiction and Fantasy”.

What can an examination of a web information resource tell us? It can safely be assumed that a commercial resource such as the directory in question will neither strike categories from its hierarchy nor add categories to it without deliberation and study of user habits. The categories given in the hierarchy will be useful, in the sense that they in fact are used: the links are followed by site visitors.

Even this cursory glance at the hierarchy tells us three things: first, that most categories found useful by web users are topical; second, most are grounded in traditional media and learning, in categories that are well agreed upon by authors and readers alike; and third, that technologically based innovative genres which appear to cut across topical categories are not necessarily stable even after being recognised by an obviously thoughtfully and conservatively built static resource. “Virtual postcards” is a good example of a genre based on traditional paper-based media and realised as a server-based solution. After an initial period of enchantment by web users, it has been supplanted by point to point messaging. The new genres that seem to lead a stable existence in the hierarchy are “Blogs”, “Chats and Forums” and various variants of “Searching the Web”. These are genres that can be given an analysis beyond their immediate technology and implementation – they introduce new communicative situations and new services, transcending the constraints of paper-based and spoken media. We can expect new genres to emerge when similar qualitative developments in communicative technology become wide-spread and stable; a mere new widget or transplant of previous media will not in itself provide new necessities for conventionalisation.

## 2.4 Data Source: Observation of User Actions

A further source to understanding data variation is to investigate actual user behaviour. What genres do users believe they can retrieve? By examining three months of queries made by web search engine users released for academic research in 2006 by America Online [9] we find several expressions of genre-related preference.

The collection consists of about 20 million web queries collected from several hundred thousand users over 3 months. The users are primarily home users and the queries reflect this fact, both with regards to topic area and user expertise. The data incorporate information about whether the user in question pursued reading any of the retrieved documents through clicking on the link, and in that case gives the rank of the item.

Understanding, or, more correctly, inferring user aims, plans, and needs from observing search queries issued by the user is naturally fraught with risk. The behaviour of search engines discourages users to be overly specific – there are practically always many ways of understanding a brief and often very general expression of information need; search system users perform their actions for very various reasons. The information need that prompts users to use a search engine may result in various types of changeable or interleaved information seeking strategies [e.g. 2], and the behaviour of the user is strongly influenced by factors such as feedback,

the conceptual framework presented to the user, and various factors in the interface itself [e.g. 3]. In current web search interfaces, very little information is volunteered to users, who in effect are left to their own devices and need to form their conceptual framework unaided, by browsing and perusing the information sources at hand. Genre is not a facet the search engines encourage users to specify, and no indication that the search engine might be competent in judging genre is given. This means that the variety of user queries with respect to form and content alike must be interpreted with some care – the learning process of users in new topical areas is likely to influence the query history crucially. But given these caveats, what might we find in web search engine query logs, and how might we understand what we find?

The log entries are of the following form:

```
2178 hepatitis b vaccine safety infants 2006-05-09 19:43:37 2
http://www.vaccinesafety.edu
```

which tells us that some user posed a query on vaccine safety at a certain date and clicked on the second item in the returned list of web sites. Queries can be of many types. An early, simple, and useful classification of information needs into *Navigational*, *Informational*, and *Transactional* queries was made by Broder [4]; later elaborated by Rose and Levinson [10] to *Navigational*, *Informational*, and *Resource*, with a number of sub-classes for the second two. Navigational queries are “look-up” queries that are issued to find a specific item of knowledge on the web: a corporation, an address, a personal web page, or some other specific web location. Informational queries attempt to locate some information assumed to present on some web page or web pages. Transactional – or Resource, using Roses and Levinsons terminology – queries reflect the intention of the user to perform some activity served by some mechanism found on the web. This typology can be matched reasonably well to the results from questionnaire studies such as the one given above. In the material here at hand, many or even the majority of searches appear to be navigational searches: attempting to find a specific site or a specific product. Examples are e.g. “Avis” or “Northrop Grumman Corporation”. These are less interesting for our present analysis purposes, as are the Transactional/Resource queries: we will here take a closer look at Informational queries.

To examine whether genre indication made a difference for the information request the query logs were investigated for presence of explicitly genre-indicating terms. Various food-related queries were collected and checked for presence of the word “recipe”; queries searching for information on Eminem, the rap artist, were checked for presence of the word “lyrics”; queries searching for information on wiring were checked for presence of the word “instructions” or “schema”. In all, about 20,000 queries were tabulated, as given in Table 2.5. Most queries in the sample resulted in a click through to some retrieved result, which is to be expected – this holds true for the entire query log collection. For each of the three experimental topics, the queries with genre-indicating terms delivered a lower average rank score for click-throughs, which would seem to indicate adding genre adds useful information to a query.

**Table 2.5** Examples of genre mentions found in queries

Type of target	Non-genre query	Genre-indicating query
Food recipes		“Recipe”
Paella, risotto, turkey etc	<i>Grilled turkey wings</i>	<i>Recipe sourdough bread</i>
Average rank of click-through	7.5	6.7
Number of click-through queries	6,575	505
Number of non-click queries	3,047	124
Looking for advice or self-help		
Technical: wiring		“Instructions”, “schema”
Examples	<i>Jeep grand cherokee radio wiring</i>	<i>kl800 wiring instructions relays solenoids keyless</i>
Average rank of click-through	12.1	6.7
Number of click-through queries	2,956	93
Number of non-click queries	1,491	62
Musical: eminem		“Lyrics”
Examples	<i>When i’m gone eminem</i>	<i>When im gone eminem lyrics</i>
Average rank of click-through	6.4	3.1
Number of click-through queries	1,645	397
Number of non-click queries	2,002	113

There are several reasons to be cautious in drawing too far-reaching conclusions: we cannot say for sure what the users were after; longer queries (which the genre-enhanced queries are, on average) often are more successful; queries without the genre indicator may in fact be searching for other genres; some queries might not be informational. In spite of this, given the reasonably large numbers of several thousand queries given in the table, and the fact that the difference in rate of click through is significantly higher<sup>5</sup> for the genre enhanced queries we can identify the fact that users do refer to genres in posing queries, and that the effect of doing so appears to be beneficial. This means, since the queries are mediated by a genre-unaware search engine, that the explicit mention of genre in the query is matched by a likewise explicit mention of genre in the target text – harking back to the discussion on bidirectionality between reader expectation and author model of audience given in the first section of this chapter.

<sup>5</sup> Tested by  $\chi^2$ , significantly higher click-through rates for the “food” and “eminem” queries separately ( $p > 0.999$ ) as well as for all three examples taken together.

## 2.5 Conclusions

The argument given by the three investigations presented in this chapter is to strengthen the claim given in the introduction that genres are a form of implicit agreement between readership and authorship. Questionnaire responses indicate that readers have clear in their mind their own needs, and sometimes the needs they believe others have. They do not explicitly model genres by their character – the content and topic is much more salient than the style and form, even while readers use style and form to identify adequate content. The new genres they mention and acknowledge can be claimed to be of two types.

Firstly, new genres not based on traditional media but based on new technology, technology that bridges earlier distinctions between e.g. written and spoken discourse transcend old categories and while they sometimes borrow from it are likely to create new genres and new conventions eventually.

Secondly new genres that delve deeper into special interests and topics, based on the larger uptake of information technology as a mass communication mechanism for the general public. Sports, Motoring, Celebrities – these are genres or subgenres known from traditional media of today and represent the “normalisation” process information technology is undergoing at present.

If we, as we investigate the character and form of computer-mediated communication, information access systems, social media, or human-machine dialogue are to postulate new genres, it is not enough to discover new surface features, new turns of phrase, or new forms of expression. Not unless they are in some way related to the audience, its communicative needs, and to author understanding of the same. This can most reliably be discovered through study of actions people make, by the information needs engender. Genres are a behavioral category which can be *described* by content analysis, but not *explained* by it.

**Acknowledgments** The author of this paper gratefully acknowledges the help of Jan Pedersen, ConnieAlice Hungate, and Adrienne DeiRossi at Yahoo!, of Viggo Kann, Leif Dahlberg, and Oscar Sundbom at KTH for recruiting informants, and of course the generous contribution of the participants in the questionnaire studies. Thank you! Part of this research was conducted while visiting Yahoo! Research in Barcelona.

## References

1. Bakhtin, M.M. 1986. The problem of speech genres. In *Speech genres and other late essays* (Trans: McGee, V.W.). Austin, TX: University of Texas Press.
2. Belkin, N.J., P.G. Marchetti, and C. Cool. 1993. Braque: Design of an interface to support user interaction in information retrieval. *Information Processing and Management* 29(3):325–344. doi: doi: 10.1016/0306-4573(93)90059-M.
3. Bennett, J.L. 1971. Interactive bibliographic search as a challenge to interface design. In *Interactive bibliographic search: The user/computer interface*, ed. D. Walker, 1–16. Montvale, NJ: AFIPS.
4. Broder, A. 2002. A taxonomy of web search. *SIGIR Forum* 36(2):3–10. ISSN 0163-5840. doi: <http://doi.acm.org/10.1145/792550.792552>

5. Dewe, J., J. Karlgren, and I. Bretan. 1998. Assembling a balanced corpus from the internet. In *Proceedings of 11th Nordic Conference of Computational Linguistics*. Copenhagen. URL [http://eprints.sics.se/63/01/Dropjaw\\_korpus.html](http://eprints.sics.se/63/01/Dropjaw_korpus.html)
6. Freund, L., and C. Ringlsetter (Guest Editors of Special Section), eds. 2008. Bringing genre into focus, volume 34 of *Bulletin of the American Society for information science and technology*. The American Society for Information Science and Technology.
7. Karlgren, J. 1992. The interaction of discourse modality and user expectations in human-computer dialog. Master's thesis, Licentiate thesis at Department of Computer and Systems Sciences, Stockholm University, Stockholm. URL <http://eprints.sics.se/53/01/lic.ps>
8. Karlgren, J. 2005. The whys and wherefores for studying textual genre computationally. In *Proceedings of AAAI Fall Symposium on Style and Meaning in Language, Art and Music*. Arlington, VA. URL <http://eprints.sics.se/46/01/FSS804JKarlgren.pdf>
9. Pass, G., A. Chowdhury, and C. Torgeson. 2006. A picture of search. In *InfoScale '06: Proceedings of the 1st International Conference on Scalable Information Systems*, 1, New York, NY: ACM. ISBN 1-59593-428-6. doi: <http://doi.acm.org/10.1145/1146847.1146848>
10. Rose, D.E., and D. Levinson. 2004. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th International Conference on World Wide Web*, 13–19, New York, NY: ACM. ISBN 1-58113-844-X. doi: <http://doi.acm.org/10.1145/988672.988675>
11. Roussinov, D., K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu. 2001. Genre based navigation on the web. In *HICSS '01: Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, Washington, DC: IEEE Computer Society. ISBN 0-7695-0981-9.
12. Santini, M. 2006. Interpreting genre evolution on the web: Preliminary results. In *Proceedings of the Workshop on New Text: Wikis and Blogs and Other Dynamic Text Sources*, ed. J. Karlgren. Trento: European Association of Computational Linguistics.
13. Santini, M. 2007. Characterizing genres of web pages: Genre hybridism and individualization. In *HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Washington, DC: IEEE Computer Society. ISBN 0-7695-2755-8. doi: <http://dx.doi.org/10.1109/HICSS.2007.124>
14. Shepherd, M., and C. Watters. 2004. Identifying web genre: Hitting a moving target. In *Proceedings of the WWW 2004 Conference Workshop on Measuring Web Search Effectiveness: The User Perspective*. New York, NY: ACM.
15. Swales, J. 1990. *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.
16. Yates, J., and W. Orlikowski. 1992. Genres of organizational communication: A structural approach to studying communication and media. *Academy of Management Review* 17:299–326.

# Chapter 3

## Identification of Web Genres by User Warrant

Mark A. Rosso and Stephanie W. Haas

### 3.1 Introduction

Genre is seen by many as a promising enhancement to the process of web search [4, 12, 16, 23]. The capability to specify or exclude certain types of web pages during a search is intuitively appealing. Historically, document type has proven to be a useful tool for document retrieval (e.g., [6]).

Figure 3.1 graphically depicts how the use of genre in the web search engine interface could enhance web search at two points in the search process: formulation/reformulation of the search query and browsing of the search results.

A genre recognized as relevant to the user's information need could be part of the user's query formulation. For example, a user could specify that only documents of that genre be included in the search results; or, a user might decide to exclude from the search results documents of a genre deemed not to be useful. In either case, document genre is being used to constrain the search space, with the intent of improving the search results. In essence, part of the users' task of filtering search results would be taken on by the system.

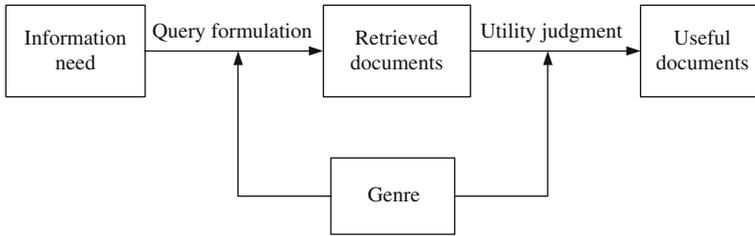
The second point at which document description by genre could be helpful is in viewing the search results. Labeling each document description with document genre could help the user to make faster and more accurate relevance judgments, and omit the viewing of some documents' full-text, thus shortening the time needed to assess the documents' relevance. Genre information in the search results could also be useful for query reformulation. For example, a user searching for detailed information on a medical condition, may notice a preponderance of advertisements for products in the search results, and could choose to exclude that genre from future results.

Also, it has been suggested that presentation of search results could be based on the characteristics of the genre of the documents that the results represent: *genre oriented summarization* [8]. For example, the summary of a product review might

---

M.A. Rosso (✉)

School of Business, North Carolina Central University, Durham, NC 27707, USA  
e-mail: mrosso@nccu.edu



**Fig. 3.1** Two points where genre could impact the web search process: query formulation/reformulation and judging search results

mention price and features while a movie review could describe the plot and include the running time of the film.

In addition to these explicit uses of genre to improve the search process, Braslavski (Chapter 9, this volume) suggests an implicit use: incorporating genre automatically into the improvement of relevance ranking of search results.

Thus, genre would seem to be of great use for enhancing search. However, the implementation of information retrieval by genre is problematic on the web – an immense, heterogeneous collection of documents from disparate sources. The pages are generally not labeled with genre metadata, there are far too many for manual classification, and it is unclear how the incredible diversity of the collection will allow for the development of effective automatic classification algorithms.

In addition to these thorny issues, a more fundamental complication exists. Before any classification (manual or otherwise) can take place, the actual genres to be used in the classification must be decided upon. What set of genres can adequately describe to users the contents of the web? What methods can be used to determine (or discover) the members of the set? Given the diversity of pages, the method for reaching this initial decision is far from obvious.

The goal of this chapter is to address methodological considerations in the selection of genre labels to be used to describe web pages indexed by web search engines. For the purposes of this chapter, we will consider that the specification of a genre includes a description or definition (intensional, extensional, or hybrid) of the documents that fall into the genre category, and a name or label that users can recognize as identifying the genre. We first propose criteria for the identification of web genres, and the types of methodologies that are implied by those criteria. We then discuss in detail how the concept of genre applies to the web, and identify the resulting implications for the development of web retrieval by genre. A series of user studies designed to create a genre “palette” are used as examples to illustrate the issues involved in developing a methodology for the identification of genres for enhancing web search, based on the concept of user warrant.

A fundamental difficulty in designing genre studies is how to incorporate the context of search in a realistic manner. Aspects of context come into play at multiple points. The user’s context of search includes what has already been seen, both prior to the search and in reviewing retrieved pages. Pages that seem similar to ones that have already been dismissed as not useful may be examined only briefly, if at all.

Pages of a genre type that the user has found useful in the past may be of more interest, at least initially; the opposite is also true. Web pages do not exist in isolation; the pages that link to them and that they link to form a context. A page may make little sense without seeing its predecessor or parent page. So the user's willingness to explore the surrounding pages could affect his judgment of the target page. To simplify the task of web genre identification, we consider here that a label applies to a genre instance of a single web page, as opposed to applying to a website, or other multi-page instantiation. This is consistent with the reality that current search engines deliver search results as individual pages, but we acknowledge that it is an artificial constraint.

### 3.2 Criteria for the Identification of Web Genre

We propose three criteria for the identification of genres to be used in web page retrieval. First, the users of the system (or some portion of them) must possess sufficient knowledge of the genre to have some understanding or expectations of what it is. Users unfamiliar with a genre will receive no benefit from encountering its label in the search process. The genre must be recognizable to the searcher. Second, searchers must be able to relate the genre to their information needs or tasks, that is, be able to predict if it is likely (or unlikely) to contain useful information. Otherwise, the genre label will not be meaningful to them in the context of their search. Third, the genres must be predictable by a machine-applied algorithm. Because of the size and rate of growth of the web, automatic categorization with a reasonable level of accuracy is crucial. So, in summary, genres for web retrieval must be recognizable by searchers, useful for searchers' information needs, and predictable by machine.

These criteria for the identification of web genres suggest types of methodologies that web genre researchers could employ in their work. For example, to identify typical genres of a specific user group, one might sample members of the group and ask about their typical web usage and what genres come to mind. To show that a specific genre is recognizable, one might ask members of that genre's user group to name, describe or define the genre of specific web page instances in order to see if they agree – thus showing that the user group does possess the shared knowledge that genre theory normally espouses. Thus, measures of participant agreement are used to estimate the strength (in terms of recognizability) of a genre. Any characteristics of form mentioned by the users could also be noted for use as potential features in the development of an automatic classifier.

For insight into genres' usefulness, the users could also be asked to talk about their search needs and what genres they search for. Ideally, to demonstrate the usefulness of the retrieval by genre concept, one would want to directly compare search systems with and without genre augmentation. Then, any number of search evaluation criteria could be used to show the difference between the two. However, the methodology is problematic in that it would require that the automatic genre classifiers to already be developed. (See Chapter 8 by Stein et al. elsewhere in this book for work in this area.)

Alternatives to compensate for the lack of genre classifiers would necessarily detract from the ability to generalize the studies' results, but could still provide useful experimental data. For example, one could pre-label a limited corpus and restrict the queries used by participants in their search sessions. Also, one might observe and record users' search sessions to uncover relationships between searchers' judgments of search result or document relevance, and the corresponding document genre.

Regardless of the specific methodologies chosen, the criteria of recognizability, usefulness, and machine predictability are necessary for the successful integration of genre into the web search engine.

### **3.3 Operationalizing Traditional Genre Theory for the World Wide Web**

We consider a genre on the web to be a pragmatic type (with corresponding form and substance), that is recognized by the genre's "user group", those with a common or shared knowledge of the genre [19]. The genre classification scheme is derived from user instincts, experiences, and preferences, not any given theoretical framework, much like folksonomies and other user-derived or -generated schemes. These contrast with expert-imposed classifications, like the difference between a zoological taxonomy and a lay distinction between pets and wild animals (see Chapter 7 by Sharoff, this volume, for an example of an expert-imposed classification). Regardless of the derivation of the classification, it is necessary to validate the definitions, labels, and application of definitions to web pages by members of the target group, in order to ensure a reasonable level of recognition and agreement. Without the recognition and agreement by the user group, a page type (i.e., a proposed genre) is not necessarily a genre.

The preceding paragraph encapsulates the challenge of transforming the theoretical construct of web genre into an operational definition (as embodied in labels and definitions of web genres) that could be used by content authors/designers, classifiers (human or automatic), and end users of a variety of applications such as information retrieval or content management. We briefly discuss the issues associated with the transformation in order to provide context for the decisions made in the studies described later in this chapter, and the implications these decisions have for the experimental results.

#### ***3.3.1 A Genre's User Group***

Traditional genre theory almost always includes the notion of a "user group" whose members share some knowledge about the genre, and thus have expectations about its intended use, form, and substance (e.g. [14, 22]). User groups may vary in cohesiveness or restrictiveness of membership criteria. For example, the primary user group of the letter of recommendation for an applicant to graduate school

contains writers and readers of such letters, typically faculty members at colleges and universities. They share knowledge of the purpose of the letter, and what the reader expects in terms of its content, formality, and even legal status. The genre may have somewhat limited circulation, and those not in the primary user group are less likely to encounter it or need to use it. If they do, they may use it for uses other than its intended function. For example, a new faculty member may use such a letter as a guide for writing for his/her first letter of recommendation for a student, or a biographer may glean information about a person's life from it.

In contrast, the user group of the newspaper editorial is varied, with the most salient shared characteristic being that they are readers of newspapers, and are likely to understand the difference between an editorial and a news article (although not necessarily). Level of education (beyond some level of literacy), vocation, and other characteristics are not part of the "membership criteria".

It is important to recognize that any one individual is a member of multiple user groups, both broad and specific, and can view a single page from the multiple vantage points the groups afford. Although the purpose of a search is likely to derive from a user's membership in one group, he/she can switch hats rapidly if something of interest to his/her role in another group appears (serendipity).

When we focus specifically on web genres, this view of a genre's user group does not change substantially. There are still cohesive user groups who work with specialized web pages, and have clear expectations of what they contain; these expectations are not widely shared outside the group. Indeed, the web may provide the means for even more specialized groups to exist: profession-based groups, hobby and fan groups, employees of a single company, and so on. They may recognize more specific genres, or have more accurate expectations as to their form and content, even though they are not the only web users to encounter it. For example, anyone can find a university department's home page on the web, but a faculty member at a university may have stronger expectations of what information should (and shouldn't) be there, and how it should be organized than, for example, a high school sophomore. Thus, the "web user", like the non-web "newspaper reader" will have shared experience and expectations about genre-related characteristics of commonly encountered types of web pages.

However, searching the web greatly increases the likelihood that someone from outside of a genre's primary user group will encounter an instance of that genre. Pages from relatively esoteric user groups may turn up in search results, or someone may deliberately seek information that is outside their usual information environment, e.g., a consumer searching for expert health information. Thus, although a page may be created by and for a specific user group as an instance of a familiar genre, the page may be viewed by "outsiders" to whom the genre is foreign. This is a characteristic of web search that genre augmentation may not improve. Another such characteristic of the web is the existence of pages that are not the results of recurring situations, i.e., not recognized by any user groups as belonging to any genre. Ideally, such pages would remain unlabeled by an automatic genre classifier. Complete coverage of a collection (suggested elsewhere in Chapter 4 by Crowston et al., this book) is neither possible nor desirable.

An issue that is related to the concept of user group is the level of abstraction of a genre. Broader genres, such as article and home page, are hypothesized to typically be recognized by larger or more diverse user groups than narrower genres like an SEC filing or copyright transfer agreement [24]. A question for further research is whether the characteristics of the “typical” web user tend to be associated with broad, large-grained genres. In other words, how specific are the genres recognized by most everyone, and how strong are the expectations about the genres? Is the concept of “web user as user group” useful for our purposes? How useful are the broad genres that they may recognize for improving web search? Some researchers have questioned the utility of considering web users as a whole to be a relevant group in terms of retrieval by genre [15, 16].

Despite the issues of pages with unrecognized or unknown genres, one can conclude that the concept of a genre’s user group is very much applicable to web genres, and not materially different from documents in other media. What are the implications of this for web genre research? Operationalizing the “user group” as a group of people with obvious shared characteristics, such as profession or workplace, thus also characterizes to some extent the websites they frequent, specifically those associated with the shared characteristics. This is likely to make some parts of the research easier. The limitation provides some justification for limiting the sample of web pages used in the research by domain or organization. The participants are likely to have more shared knowledge (e.g., what an academic department does), familiarity with the work and work documents, and thus be able to recognize more specific genres and have more accurate expectations as to their intended use, form, and content. Because of these expectations, they may also be able to see the utility of using genre as part of information seeking.

However, generalizing research findings to other user groups, or to web users as a whole, will be problematic. Some groups may work with more specific genre that support stronger expectations than others. Some specific genres may have characteristics that are more easily usable by people outside the primary user group for some purposes, e.g., finding links to relevant information.

### ***3.3.2 Genre: Function, Form and Substance***

In discussing the individual aspects of the genre pragmatic type and how they apply to the web environment, we set up a sense of distinctness among them that does not exist in reality. In use, the distinction between function, form, and substance blur: form shapes substance, substance entails function, and so on.

*Function.* The “function” of a web genre could be viewed from two perspectives: that envisioned or intended by the creator of an instance of the genre, and that perceived or acted upon by the user. For a genre used by members of its intended user group, the two perspectives will generally be aligned. Non-members’ actual uses of the page may be in alignment, or be entirely different. The common phenomenon of using a genre as a container of needed information, rather than for its intended purpose, frequently occurs on the web. In some of our studies, participants would

commonly judge a web page as useful not because of its content, but because it contained a link to the desired content. In this scenario, recognition of the utility of a genre means recognizing evidence of where it might lead, overlaying a directory or referral-type function on top of its intended function (what Chapter 4 by Crowston et al., elsewhere in this book, describe as “borrowed purpose”).

Adding to the difficulty is the fact that search engines return individual pages, isolated from related pages that may provide needed context – potentially important pages whose existence may not even be known to the searcher. The function or purpose of a genre is traditionally seen as a shared understanding among creators and users of the genre as to its role in actions and communications. The shared understanding is based on knowledge of the context in which it is used. On the web, the originally intended context of pages can be more elusive: users may come to a page deep in a website from a Google search, and may have little interest in looking beyond it. Any guess as to the purpose of the page is based on face evidence, not an understanding of its context. This type of situation could increase the difficulty for even a genre “insider” to recognize a page’s genre. Thus, single-page genre validation methodologies (such as the one described later) could underestimate users’ recognition of a genre.

For research into the use of web genre for information retrieval, these observations suggest that asking subjects to rate the utility (or relevance or whatever construct is used) of a genre instance could be misleading. A page could be judged useful because the user views it as supporting a function that is unrelated to the definitional functions of its genre. This does not mean that the user hasn’t recognized its genre, or has no expectations associated with it, rather that the user associates different (or additional) functions with it. For example, someone looking for the title of an article written by a faculty member may judge a department home page to be useful, because from there, he can find the faculty member’s personal page, which is likely to link to his CV, which should have the article listed. Asking subjects to articulate the reasons for their judgments is more likely to reveal their view of the functions supported by the web page.

*Substance.* By “substance” we mean the content (which may include topic) of a genre. When experiment participants are asked to name a web document’s genre, they often conflate topic and genre. Theoretically, genre labels should be as topic-neutral as possible. In practice, some genres are more closely tied to topic than others. For example, the substance of a newspaper article is a description of an event or situation, usually including information about the people and places involved, and often carrying an aspect of timeliness. Within this substance, however, the range of topics is vast; elections, war, weather, tennis, fashion, or just about anything else. Substance and topic are relatively independent. In contrast, the genre of university course listing is inherently about courses. The substance includes course numbers and titles, and often a brief description. The topic could be broader or narrower, for example, listing only chemistry or sociology courses, but the distinction between the topic and substance is fuzzy.

The substance may be communicated by a series of moves (e.g. [2, 22], or types of information that are typically included in a genre instance. In a letter of

recommendation, expected moves include a greeting, a description of how the writer knows the applicant, and reasons why the writer recommends the applicant. The kinds of reasons cited, or how they are framed may differ according to the type of application. For a job, the letter may discuss past educational accomplishments, while a recommendation for an award may discuss why the applicant is worthy. Consideration of the substance of web genre follows the same pattern as the previous consideration of function: a user may use elements in unexpected ways.

A fundamental difference among web genres, traditional genres implemented on the web, and non-web genres, is the presence of hyperlinks. This expands the notion of substance: the link text itself can be substance, but is also a reference to another page. The target page forms some part of the context of the initial page; users may consider its substance to be a part of the initial page's substance. The implications for experimentation are similar to those for function: a user may consider a genre instance useful because of its links and the pages it links to, rather than the page itself. Hyperlinks are also responsible for the research decisions over what constitutes a genre instance: an individual web page, an entire website, or a multi-page document (e.g., an FAQ (frequently asked questions) that spans multiple web pages).

The URL is another web-specific element: users may pick up clues as to the genre of a web page, and therefore trigger expectations of its utility, by words or abbreviations contained there, e.g., "home", "interview", or "syll".

*Form.* Form is the most obvious difference between traditional and web genres. Web genres do not provide the same physical cues (weight, size, material, etc.) as their traditional counterparts. Nonetheless, research has shown that people can recognize genre [23] and elements of specific genre [5] in digital environments.

Form is the vehicle through which genre function and substance are expressed. Returning to the letter of recommendation, the expected moves may be expressed in casual, formal, or extremely formal language. Form includes whether a letter is typed or handwritten, and even the kind of paper used. On the web, means of expression are practically unlimited, including sound and images, color, escape from the normal (for western languages) top-down, left-to-right scanning, and even form that changes as the user watches. The form of a web page can be indicative of the context of the page: the home page of a university department will use different design elements than a children's game website, although both may embody the directory genre. As page design conventions have coalesced over the past decade, the web user can expect some common elements on most, though not all, pages. The form of some genres and genre instances may be exactly the same as their non-web counterparts, as is often the case with a .pdf document. Other specifically-web genres, such as the home page and the blog, have developed their own conventions. The appearance of the substance elements is as informative of genre as the actual words or pictures themselves. For example, a list of questions at the top of the page that are links is highly suggestive of a FAQ, as opposed to an interview, which typically has alternating questions and answers.

### ***3.3.3 Genres on the Web: Further Implications for Research***

In many respects, traditional genre theory transfers easily to the web environment. The aspects of function, substance, and form are still integral to the definition and expression of a genre. The user group is also essential to the core definition of a genre, but the digital, accessible, and linked nature of web documents provides more opportunities for people outside of a genre's primary user group to view instances of the genre. These considerations affect both the selection of experiment participants, and the construction of the sample of web pages.

The presence of hyperlinks is the other important distinction between traditional and web genres, which impacts research design decisions. The perception during a web search that a document may link to something useful may have nothing to do with the document's genre: it's simply functionality added to a document (and not its genre) by the existence of hyperlinks. In other cases, the linking expands the context in which a page is viewed. For example, organization home pages can link to individual person's home pages.

These distinctions suggest that genre researchers who observe users' web search behavior, must gather more information about page utility from users than just a bare rating. The reasons for the judgment may reveal that the page itself isn't useful except as a starting point: the links to related pages (i.e., the page's context), and expectations about the related pages may be the reason for a "useful" rating. Further, a genre instance may be implemented on the web to span multiple pages. For example, a frequently-asked questions page (FAQ) may have the questions on one page, and answers on separate pages, yet users may perceive them as a single "document". Researchers must decide if subjects should be allowed to follow links when making utility judgments, and if so, how far afield they may go.

## **3.4 Developing a Web Genre Palette**

As web genres are recognized by their respective user groups, the collection of terminology to describe web genres would, ideally, directly involve the users. At a minimum, proposed genre terminology (labels and descriptions) would be validated by users in order to show that the identified labels do indeed represent genre. Thus, the genres are identified by user warrant, meaning that the appropriateness of the terminology is affirmed by the users' actual use of the terms.

A series of three user studies [19] was undertaken with the purpose of developing a genre palette for use in web retrieval. In order to start on a more manageable problem, pages to be examined by participants were limited to the edu domain, as in Rehm [16]. The web pages in the terminology studies were collected by interval sampling the Google search results obtained from one-word queries consisting of the most frequently used English words [7]. As discussed earlier, the choice to restrict the user group not only limited the pages that could be included in the sample, but also limited the generalizability of the results. The choice was made partly to avoid problems that earlier studies attributed to a web-wide focus: that it leads

**Table 3.1** Overview of the studies

	Methodology	Product
Study #1 Survey of user terminology	3 participants individually separated 100 webpage printouts into stacks according to genre, assigning names and definitions to each genre	A collection of 48 genres names with definitions
Study #2 User-based refinement of terminology into a tentative genre palette	10 participants individually classified 100 webpages (same as in the previous study) using the 48 genres (plus a “suggest your own”) category	A palette of 18 genre names and definitions
Study #3 User validation of the genre palette	In an online experiment, 257 participants each classified a new set of 55 webpages using the 18-genre palette	Validation of participants’ ability to classify pages using the palette
Study #4 Measurement of user relevance judgments of genre annotated search results	32 participants performed 4 tasks. In each task, participants judged the usefulness of 20 search results and 20 web pages according to an assigned task scenario	Comparison of participants’ performance with and without genre annotated search results

to vague and unusable results. We also desired to minimize the size of the resulting genre palette so that if the palette were used in search engine query formulation, the choice of genres available to the user in the search interface would be a manageable number. Finally, a fourth user study was conducted to gauge the usefulness of the genres identified in the first three user studies for the purpose of web retrieval. See Table 3.1 for an overview of the four studies.

The intended user group, people who share genre knowledge of web pages in the edu domain, was operationalized as college graduates. Arguably, a college graduate is most likely not as aware of the workings of an academic department as a departmental staff member would be. It is recognized that this experimental design choice, obviously made for convenience, could impact the validity of the results.

### ***3.4.1 Collecting Genre Terminology in the Users’ Own Words***

In the first study, three participants (an information technology professional, an organ transplant social worker and a computer science professor), in separate sessions, were given a stack of 102 web page printouts, and were asked to separate the pages into piles according to genre. They were also asked to name the genres by writing the names on sticky notes and placing them on the piles. After the piles were complete, participants were asked to provide a short, one or two sentence, description of each genre, and then to describe the page characteristics that led them to place a page in that genre. Participants were also asked to identify the most and

least representative pages in each pile, and to explain those choices. At any time during their explanations, they were allowed to move pages between piles, and to explain these moves.

Major experimental design decisions made here include how to present the pages to subjects, and how to allow them to name and group the pages. Certainly, allowing participants to interact with the pages in a web browser would establish a more realistic context for their experience of the pages. In addition to the fact that perusing 8.5" by 11" pieces of paper is not the natural way to view web pages, other compromises had to be made as a result of the printing. Page backgrounds were not printed because that inhibited the readability of many pages, as well as using a lot of ink. Web pages consisting of multiple printed pages were stapled together in the upper left margin of the printed pages. Long web pages (i.e., in excess of 10 printed pages) had middle pages (mostly with repetitious content and/or formatting) excluded from the printing. As genre is characterized by specific types of content and format, we hypothesized that these omitted pages should not have materially impacted the subjects' assessments. Some pages were omitted from the final sample for various reasons. Some pages looked radically different in print (often because of the missing background). Some pages just would not print properly. Despite the use of color printing, in some cases, it was hard to discern what text represented links. Thus, it is possible that participants' terminology did not fully take into account the importance of hyperlinks noted earlier in this chapter.

Despite the obvious limitations of using printed web pages, the printouts provided the participants with tangible things to place in piles (which they could name, give definitions to, and move pages between, easily and whenever desired). We did not have the resources to construct a software-based alternative that could have provided this much functionality for implementing a "card-sorting" process (e.g. [17]) with web pages, and it could be argued that users unfamiliar with the software would not find the online "piles" as hospitable to rearrangement as physical piles.

The session lengths ranged from 1.75 to 2.5 h, and still some genre names, definitions, and sorting decisions were left unexplored. It is our perspective that this was an effective, albeit time-consuming, method for gathering the desired genre terminology. Thus, we made the design decision to limit the sample size of this first study to three participants.

A danger of using such a limited participant sample is overfitting the results to this specific sample. Our experimental design reduces this possibility by filtering the resulting genres through the two subsequent studies. In the second study, a new participant sample gives their input on the genres named in the first study, and a refined set of genres is created. This refined set of genres is then given to a third set of participants for describing an entirely new set of webpages.

In this first study, the three participants used similar wording or concepts for their piles' names and descriptions, in many cases. For some pages, participants grouped them at different levels of abstraction (e.g., one had separate piles for FAQ and Help, while another had a combined FAQ/Help pile). In addition to the genre names and definitions collected, the page characteristics (in [19]) that participants associated with specific genre could be helpful in building automatic genre classifiers.

Note that the card-sorting process does not allow participants to place web pages in more than one pile. For example, if a home page contained a search box, the participant was forced to choose between the two genres, home page and search engine. This is clearly not a realistic categorization. Many researchers have noted that web pages can contain elements of multiple genres (e.g. [9]). However, given that the purpose of this first study was to collect genre terminology (i.e., names and definitions), the particular categorization of any given page was of secondary importance. We do acknowledge that this restriction could have affected the names and definitions that the participants generated.

The principle of user warrant requires a generation stage in the development of a genre palette. The card-sorting technique clearly demands a lot of effort from the participants, but the method used here allowed them to find similarities among pages first, and then name them. Thus, their genre definitions were based on several instances of what they viewed as a genre. In contrast, the method used in Chapter 4 by Crowston et al. (elsewhere in this book) asked users to generate a genre name as they viewed each individual page, which may be a more difficult task. Either way, the generation stage must be followed by a refinement stage, to group and normalize genre names.

Genres names elicited from the participants included familiar document types such as article, abstract, bibliography, course description, job listing, newsletter, etc. We crafted the terminology from this study's three participants into a list of 48 genre names and definitions, keeping the terminology as similar as possible to the original, while combining definitions which were nearly identical in wording. Many of the genres left in the list were still quite similar (e.g., product for sale, and shopping). The rationale for this is that genres, if expressed in user-generated terminology, should theoretically be more easily recognized by members of the genres' user group. For the complete list of the 48 genres, (see [19]).

Given the frequently synonymous and overlapping definitions in the list resulting from this study, the goal of the next study was to help refine the terminology into a smaller set of mutually exclusive genres.

### ***3.4.2 Users Choose the Best of the Collected Genre Terminology***

In this second user study, the extent of user agreement would once again be used to determine the most natural terminology, but this time with a different set of users who would vote on the terminology collected in the first study. Each of ten participants was given the list of genre name/definition pairs, the same stack of 102 printed web pages (arranged in a different random order for each participant), and a data collection form to record a genre for each web page. For each of the 102 web pages, the participant wrote a number from the list corresponding to a genre/definition pair which best described the page; or suggested his/her own genre name and definition, if none of those in the list seemed adequate.

The participants were drawn from a convenience sample of approximately 10 college graduates of various occupations. The ten sessions ranged from 65 to 120 min, for an average of 90 min per session overall. From a list of 49 genres (including the addition of the “none of the above” option), many of which were extremely similar in nature, the resulting level of agreement is quite acceptable: half or more of the participants agreed on one genre for a given page in 60% of the instances. This result is particularly notable, given that each of the 10 participants was voting on terminology from three other people, all collected independently from each other.

Another factor that might be detrimental to the agreement level here is that the definitions shown to the participants in this second study were presented out of context. In the previous study, each genre definition was part of a participant’s constructed genre palette. If we think of a palette as a collection of genres, each genre definition not only describes a single genre but also impacts the boundaries of other genres in the palette. That quality was lost in this study in which several palettes had been combined. Unlike a genre definition in a genre palette, each definition in this study had to stand on its own. These genre definitions can also be considered to be out of context because the participants in the previous study did not necessarily intend for their definitions to be understood by a public audience.

Of course, as in the first study, web pages presented individually are automatically out of context, devoid of the links to other pages, and pages that link to them. The fact that shared genre knowledge is based on understanding the context in which it is used, makes the level of agreement on genres here seem even more robust.

Another limitation in these studies is the use of the same set of 102 pages in the first two studies. This could work to reduce the generalizability of the resulting palette to other sets of pages. The decision to use the same set of pages again was based on convenience, and may have worked to increase the level of agreement observed.

After the 10 participant sessions were completed, we then developed a set of five principles [19] for creating a genre palette from individuals’ sortings. Based on those principles, the original list was trimmed down to 18 genres (see Table 3.2).

Note that the genres in Table 3.2 seem to be at varying levels of abstraction. There are broad genres such as Article and Welcome/Homepage, and more specific genres like job listing and course description. Certainly, the genres named by participants were influenced to some extent by the specific pages in the 102 page sample. Regardless, genres’ varying level of abstraction raises research questions for each of the three proposed criteria for genres to be used in search.

First, as noted earlier, what are the levels of abstraction of genres that the “typical” web user recognizes? Does targeting all web users for the user group (i.e., the “lowest common denominator”) limit the palette to broad genres? It is obvious that targeting a narrower user group (e.g., people familiar with higher education) does not limit the palette to sub-genres. They recognize all the broad genres that the larger group understands (like article), and even more specific ones like “job listing” that are not specific to the edu domain.

Second, is there a general relationship between genres’ level of abstraction and their usefulness for searching? For example, the concept of product review has

**Table 3.2** Palette of 18 genres

Genre	Description
Article	Something about a topic, often with supporting facts or opinions
Course description	What's covered in a course; syllabus
Course list	Page that lists courses
Diary, weblog or blog	A personal narrative or time log of activities (not a biographical article)
FAQ/Help	Frequently asked questions, or assistance in helping you perform a task; questions may be links to answers, or topics may be links to assistance; not interactive like a forum
Form	Page primarily for entering and submitting information (other than a search engine)
Forum/interactive discussion archive	One or more messages and/or responses that are viewable by an audience
Index/table of contents/links	A page which is primarily a list of links or text items ordered (usually alphabetically) so that a list item can be found easily, AND the page does not belong to any of the other categories
Job listing	Describes one or more jobs that are available
Other instructional materials	Materials (other than a syllabus) used in teaching courses, including but not limited to tests, quizzes, assignments, answer keys, etc.
Personal website	Page (possibly a home page) that somebody writes about oneself (but not a biographical article)
Picture/photo	Page primarily containing a picture or pictures with few or no words (other than captions)
Poetry	Contains poetry or similar wordplay
Product for sale/shopping	For purchasing products (not a product review article)
Search start	Page primarily to enter key words and search a database; a search engine
Speech	Text of a speech
Welcome/homepage	Starting page (does not have to be the "top" page in a site); may contain introductory information about a specific organization, department, program, etc. and a table of contents
NONE OF THE ABOVE	Page that definitely does not fit into any of the above categories

more distinguishing characteristics than that of article. Does that mean that users could more easily relate product review to their information needs than article? Certainly, it depends on the task and the document collection. In general, though, it makes intuitive sense that broader genres may not be as useful for searching as those sub-genres with greater number of distinguishing features. Lee [13] provides an in-depth discussion about genres' level of abstraction. In a project to label the British National Corpus (BNC), Lee asserted that the level of abstraction does not matter as long as the categories are found to be useful. However, his statements were made in the context of researchers selecting texts from the BNC for linguistic study. For our purposes, this remains an open research question.

The article genre is an interesting case in point. The name can refer to wide variety of documents, from a research article to a newspaper article. One can further subdivide these, recognizing distinctions between a hard news article and a fashion article, or a biochemistry research article and a literary theory research article. The interplay with the user group suggests that multiple levels of specificity might be useful. If a user is in his/her role as general web user, then the ambiguity of "article" may be helpful in making a broad distinction between an article and a FAQ or job listing. The finer distinctions between different subgenres of research articles are not likely to be meaningful to the general web user, whereas they may be important to a researcher. The researcher user group can recognize the characteristics of a typical biochemistry research article. If both broad and narrow genres are useful at different points to different user groups, this suggests that a palette with hierarchical structure would be more adaptable.

Finally, how does a palette containing genres of varying levels of abstraction affect the ability of automatic classifiers? Some researchers have suggested this to be a problem, (e.g. [20]). It makes intuitive sense that a mix of broad and narrow genres could cause problems for automatic classification.

We will re-visit the issue of varying levels of abstraction of the genre palette derived from user terminology. For now, we will turn to the third study. Its' objective is to validate the palette by measuring the agreement among a new set of participants using the palette to label a completely different set of web pages.

### ***3.4.3 User Validation of the Genre Palette***

The first proposed criterion for genres to be used in search is that of recognizability by the community of persons (the user group) that create and use the genre in the context of a recurring situation. We operationalized recognizability in this study as the level of agreement between participants in classifying a set of web pages into the genre palette. Agreement is measured on a page-by-page basis by a simple percentage of all the participants' votes. We based this decision on the principle of user warrant. Historically, user warrant was used as the justification for including a term in an indexing system because the users used it to search for documents (e.g. [1]). Although the genre names were not derived from actual searches, it was derived from users' classification activities, which is essentially what people do

when specifying a search query: produce terms that describe a document. Thus, we believe this analogy is appropriate. The next decision was to define a threshold of agreement that would represent a sufficient level of recognition. We propose that if 50% or more of the participants say that a page is an instance of a specific genre, then it is. The rationale behind choosing 50% is that it guarantees that the genre is the most frequently cited for that page. In most cases, the genre garnering the second highest level of agreement had much lower agreement than the highest one. We would consider our genre palette as a whole to be “validated”, thus satisfying the first proposed criterion for web genres, if the majority of the pages reached or exceeded the 50% threshold. At a minimum, we hoped that the palette contained at least some genres that met the threshold in order to “certify” them as true genres.

A new set of 55 web pages was collected using a method similar to that for collecting the 102 pages used in the first two studies. We created a website to collect demographic data and participants’ genre choices for the 55 web pages. After completing the study, participants had the option of giving feedback about their classification experience and/or leaving contact information if they wanted to talk about their experience.

Again, the intended user group was people familiar with the higher education environment. This time, it was operationalized as faculty, staff and students at 4-year institutions. Two hundred fifty-seven people participated in the study.

A flaw in the experimental design was in not collecting enough demographic information regarding the academic disciplines that the participants were associated with. We were not able to determine if the results from this self-selected sample were from a representative cross-section of the intended user group, or biased toward those who may be especially interested in web pages, e.g., people in information technology and information science-related fields.

In any case, the results were quite good. Eighty-seven percent (48 of 55) of the pages reached the 50% recognizability threshold. The average agreement for the most frequently genre assigned for a page was 71.9% for all 55 pages. Inter-participant agreement was 58.3%, with a Cohen’s kappa of 0.55. We used two measures to estimate the strength of the individual genres’ recognizability.

First, for each genre, we looked at the average agreement for that genre over the pages that were determined, according to our threshold, to be of that genre. The higher this percentage, the more frequently a page of this genre was recognized as being a page of this genre.

The second measure can be thought of as a measurement of “false hits”. This was the percentage of votes for a particular genre, across the subset of 48 pages in which this genre was not the threshold-exceeding genre. (Remember that only 48 of the 55 pages received votes exceeding the 50% threshold for any single genre.) The lower this percentage, the less frequently a particular genre was confused with the other genres. In other words, this measure shows how well participants recognized that pages were NOT of this particular genre. Note that this measure of recognizability is imprecise in that all false hits are not created equal: confusion between two similar genres like syllabus and course description is not as severe as confusion between two more dis-similar genres like poetry and job listing

For an example of how the two measures were used, the genre *job listing* scored high in recognizability on both measures: average participant agreement on job listing pages was 82.1%, while false hits were just 0.0%. Together, these two measures gave a more complete picture of the strength of the genres in the palette. An open question is how to combine these measures into one measure. Using these two separate measures, it is not possible to rank these genres according to the single construct of recognizability. Some genres had high levels of agreement, but also more false hits, and vice versa. For example, “course description” had the highest consensus of all the genres at 94.2%. However, it had one of the worst false hit rates at 2.4%. See Rosso [19] for additional details.

Using the two separate measures, we attempted to derive general ranges of recognizability for the genres in our palette. Highly recognized genres included picture/photo, job listing, poetry, product for sale/shopping, FAQ/Help, “diary, weblog, or blog,” and search start. Personal web site, forum/interactive discussion archive, and form fell into the medium range of recognizability. Genres with low recognizability were article, index/table of contents/links, other instructional materials, and none of the above. Genres with disparate scores on the two measures were course description, course list, welcome/homepage and speech. These are harder to place in a range, but course description and course list would likely fall into the high or medium range, and welcome/homepage and speech into the medium or low range.

What jumps out from this list of rankings is that the broadest genres (e.g., article) received the lowest recognizability scores. If this finding is corroborated in future research, it has important implications for the future direction of research in web retrieval by genre. We have already said that the usefulness of broad genres for retrieval is an open question. If typical web users are not clear on these broad genres (i.e., there is not strong shared understanding), then it seems more unlikely that they will be useful for search. If that is the case, are there enough narrower genres recognized by the typical web user to make web search by genre feasible? It is possible: in this study, most of the better-recognized genres are narrow, but not specific to the educational domain.

In addition to participant agreement, an abundance of detailed “de-briefing” comments written by participants provided a rich lens through which to interpret the results. Some comments noted the general ease of the task, but participants also noted several difficulties that have implications for the design of future studies.

Some pages fit into more than one category, for example, a home page with a search engine on it. As mentioned earlier, the operational decision to force participants into a one genre per page classification simplified the calculation of participant agreement. However, it made the task less natural. Agreement might be higher if multiple genre assignment was allowed, but it is unclear how that agreement should be calculated.

Another problem noted was that some pages didn’t seem to fit any of the categories. Participants suggested many names for these types of pages. This could be an artifact of using a different sample of web pages in this last study. Studies similar to the first two studies may need to be repeated to capture as many of the commonly used genres as possible. Participant comments also suggested that several

of the broad genres such as article, other instructional materials, and form should be broken down into more specific categories.

Also, some labels for web page types may not represent a single shared understanding – in other words, the label means different things to different groups of people. For example, one participant made the following comment:

I found the welcome/homepage a bit disconcerting. Many pages seemed that they were welcome pages, but definitely not homepages, whereas [sic] others were in fact homepages. [18, p.115]

In an email exchange with this participant, it became clear that he considered a welcome page to be a top-level entry point to a website, and a home page to be a personal Web site. A search of home page definitions using Google uncovered both definitions for home page. (Perhaps Dillon and Gushrowski's [5] "personal home page" would work better in the palette than personal Web site.) The point is that some commonly used labels may appear to be genres but that a single shared meaning for the label has not yet crystallized within the user group. Blog is another label that is commonly used to refer to pages with vastly different functions [11].

In summary, although several genres with high levels of agreement were identified in these studies, further user studies are necessary to collect additional genres and to refine genre names and descriptions already in the palette. Questions remain regarding the identification of broad genres with good recognizability, and the decomposition of broad genres down into narrower ones. Methodological issues such as allowing users to assign pages to multiple genres, and how to measure agreement in these cases, as well as the creation of a single measure of recognizability, also deserve attention.

There is still cause for optimism regarding the genre approach to web search. Interestingly, the genres in this palette, although developed independently, are similar to 7 of 8 Internet-wide genres based on user input reported in Stein and Meyer Zu Eissen [21], and similar to 8 of 11 Internet-wide genres as reported in Karlgren et al., [12]. Based on these observations, one might infer that some substantial amount of genre knowledge exists among users, even from different cultures (in this case, the United States, Germany, and Sweden). See Rosso [19] for a side-by-side comparison of the palettes.

#### ***3.4.4 A Fourth Study: Determining the Genres' Usefulness for Web Search***

Having identified a palette of fairly recognizable genres in the first three studies, the next step was to investigate whether using genre to augment web search could produce a noticeable improvement. The final study compared participants' ability to make relevance judgments of web page search results with and without the pages' genre label included in each search result. Thirty-two participants (college faculty and staff) performed 4 tasks in random order. In each task, participants judged the usefulness of 20 search results and 20 web pages according to an assigned task

scenario. The stability of each judgment from search result to actual (the “gold standard”) was measured. Search results were labeled with the genre of the web page in two of each participant’s four tasks.

Overall, genre-annotated search results did not produce faster or more stable relevance judgments. However, many users preferred having the genre of the web page available in the search result to help them in the evaluation process [18].

What do these results mean for this line of research? There are many possible reasons for not finding a measurable difference in performance between genre-annotated search results and “standard” ones. Certainly, tasks, users, collections, and their interactions are all complex variables. In these experiments, the user tasks were assigned, and they weren’t real search tasks – each task was a series of judgments of single surrogates followed by a series of judgments of web pages. The set of tasks was long – an average of 1.75 h. Also, participants were not informed that genre labels would be present in half of their tasks; over half of the participants reported that they didn’t remember seeing any of these labels!

But comments from two participants of the study described in Section 3.4.3 may yield some insight into how to improve the design of this type of study.

The category of a page is hardly a consideration when “Where’s the information?” is the purpose of the visit.

Normally, I wouldn’t seek to classify web pages in order to know whether they were relevant to my interests or objectives; either the information would interest me or not, continue to inform me or not, and I’d move on to the next search technique. [18, p. 116]

These comments echo our earlier discussion about the function of “web page as a container of information” being overlaid onto the function of a page as expressed by page’s genre. This study required participants to make relevance judgments on a scale of 1–4, without taking into account the reasons behind the judgment. Relevance judgments may have nothing to do with a page’s genre, and everything to do with the presence of the sought-after information. The point is that the influence of genre on the evaluation process cannot be teased out of the experimental results unless it is determined which judgments were made on the basis of genre (and which were not).

Thus, experiments hoping to measure the effect of genre on the evaluation of search results need to include some method for getting this information from the user, while at the same time minimizing the disruption of the user’s decision-making process. Methods could include a think-aloud procedure, or a debriefing immediately following the experimental procedure.

### 3.5 Conclusion

We have described the issue of identifying genres on the web for the purposes of web retrieval. Through the examination of genre theory and the literature on web genres, we have attempted to document the methodological considerations necessary for this research area to progress. More user studies need to be done to collect

appropriate genre names and definitions, and to refine those that have already been collected. The issue of broad versus narrow genres, and their usefulness for search need to be explored. Finally, techniques for accurately predicting the genres identified need to be developed.

Several important questions remain. Is there enough social agreement on web page types for this endeavor to be feasible for a web-wide audience, for the “typical” web user? If not, how could this be implemented for smaller user groups? Certainly, corporate intranets with their more homogeneous sets of users, tasks, and pages would be excellent places to start. However, other than corporate intranet users, is there some subset of users that could benefit from search by genre, and if so, what would that solution look like? Would it involve narrow genres of web pages that just these users understand and use? Or would it involve categorizing only websites of interest to this group?

It is worth noting that in finding a web-wide solution, the pages that would be annotated with genre labels are most likely only a small segment of the web: search engines only return the most popular pages. If the solution is built on top of an existing search engine, then only those pages need be annotated by genre. Does the practice of only returning the most popular pages affect what genres are available through major search engines? Would we see other genres if we could find the less popular pages? This is not to be taken as a criticism of the major search engines. They are in business to help people meet their information needs, not to provide equal opportunity for every web author’s pages to be found. But, if academic researchers are to make progress in this, or any area of web search, we may need the help of commercial search engines. Others have expressed this concern:

The commercialization of web search has caused a significant shift in the balance of knowledge between industry and academia; large web search engines have Web data, user data, and computer hardware that researchers cannot begin to reproduce, raising concerns about the quality and relevance of some areas of academic research [3].

Finally, this research area is not the only one held back by the annotation problem. Well-respected experts [10] have called for the establishment of “Annotation Science” to help solve the widespread need of several disciplines for labeled corpora, including developing methods for determining what the labels are. Researchers in web genre should be part of this effort.

## References

1. Anderson, J., and J. Perez-Carballo. 2005. *Information retrieval design*. St. Petersburg, FL: Ometeca Institute.
2. Bhatia, V. 1993. *Analysing genre: Language use in professional settings*. London and New York, NY: Longman.
3. Callan, J., J. Allan, C. Clarke, S. Dumais, D. Evans, M. Sanderson, and C. Zhai. 2007. Meeting of the MINDS: an information retrieval research agenda. *SIGIR Forum* 41:25–34.
4. Crowston, K., and B. Kwasnik. 2003. Can document-genre metadata improve information access to large digital collections? *Library Trends*, 52:345–361.
5. Dillon, A., and B. Gushrowski. 2000. Genres and the Web: Is the personal home page the first uniquely digital genre? *Journal of American Society for Information Science* 51:202–205.

6. Fidel, R. 1991. Searchers' selection of search keys: I. The selection routine. *Journal of the American Society Information Science* 42:490–500.
7. Francis, W., and H. Kucera. 1982. *Frequency analysis of English usage*. New York, NY: Houghton Mifflin Co.
8. Goldstein, J., G. Ciany, and J. Carbonell. 2007. Genre identification and goal-focused summarization. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 889–892. New York, NY: ACM Press.
9. Haas, S., and E. Grams. 2000. Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of American Society for Information Science* 51:181–192.
10. Harman, D. 2007. Meeting of the MINDS: Future directions for human language technology executive summary. <http://www.itl.nist.gov/iaui/894.02/minds.html>
11. Herring, S., L. Scheidt, S. Bonus, and E. Wright. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 38th Annual Hawaii International Conference on Systems Sciences*. IEEE Computer Society Press.
12. Karlgren, J., I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. 1998. Iterative information retrieval using fast clustering and usage-specific genres. In *Eighth DELOS workshop – user interface in digital libraries*, 85–92. Stockholm, Sweden, October 21–23, 1998.
13. Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through The BNC jungle. *Language, Learning & Technology* 5:37–72.
14. Miller, C. 1984. Genre as social action. *Quarterly Journal of Speech* 70:151–167.
15. Nilan, M., J. Pomerantz, and S. Paling. 2001. Genres from the bottom up: What has the Web brought us? In: *Proceedings of the American Society for Information Science and Technology Annual Meeting*, 330–339. Washington, DC, November 2–8, 2001.
16. Rehm, G. 2002. Towards automatic Web genre identification. In: *Proceedings of the 35th Annual Hawaii International Conference on Systems Sciences*, 1143–1152. Los Alamitos, CA: IEEE Computer Society Press.
17. Rugg, G., and P. McGeorge. 1997. The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems* 14:80–93.
18. Rosso, M. 2005. Using genre to improve Web search. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, NC. [http://ils.unc.edu/~rossm/Rosso\\_dissertation.pdf](http://ils.unc.edu/~rossm/Rosso_dissertation.pdf).
19. Rosso, M. 2008. User-based identification of web genres. *Journal of the American Society for Information Science and Technology* 59:1053–1072.
20. Santini, M. 2006. Common criteria for genre classification: Annotation and granularity. In *Proceedings of the Workshop on Text-Based Information Retrieval Held in Conjunction with the European Conference on Artificial Intelligence*.
21. Stein, B., and S. Meyer zu Eissen. 2004. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*. Ulm, Germany.
22. Swales, J. 1990. *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.
23. Toms, E., D. Campbell, and R. Blades. 1999. Does genre define the shape of information: The role of form and function in user interaction with digital documents. In: *Proceedings of the 62th American Society for Information Science Annual Meeting*, 693–704. Washington, DC, October 31 – November 4, 1999.
24. Yates, J., and W. Orlikowski. 1992. Genres of organizational communication: A structural approach to studying communication and media. *Academy of Management Review* 17: 299–326.



# Chapter 4

## Problems in the Use-Centered Development of a Taxonomy of Web Genres

Kevin Crowston, Barbara Kwaśnik, and Joseph Rubleske

### 4.1 Introduction

Web search engines such as Google or Yahoo determine relevance of Web pages according to the occurrence of words in the pages indexed by the engine (additional information is then used to rank these results). Unfortunately, such searches are not always sufficient to solve information needs since task-driven searchers often must distinguish between documents that share a set of keywords (i.e., a topic) but assume a different form to serve a different purpose or function. For example, before purchasing a digital camera, an individual may want to read reviews from online magazines and see the blogs in which people who have used this camera express their opinions and personal stories. Using a query term such as “Canon Powershot G6” could yield the bulk of results referring to digital-camera sellers, not magazines, discussion forums or blogs. A renewed search with a more refined query might prove incrementally more effective, but might just as easily yield mixed results. Efforts to locate a current, trustworthy and pertinent discussion forum might require considerable manual searching through search results.

One way to improve the precision of a search and to ensure a better match of the results to a user’s task is to utilize additional metadata to distinguish or group relevant and irrelevant documents. We focus in particular on the role of document genre. Document genre can be defined as “essentially a document type based on purpose, form and content” [21, p. 1053]: e.g., a digital-camera *advertisement*, a digital-camera *review*, or a *schematic drawing* of a digital camera.

Genre is useful in information tasks because it makes documents more easily recognizable and understandable to recipients, thus reducing the cognitive load of processing them [2]. As well, knowledge of the genre can be exploited in a number of tasks because genre provides some fixity to otherwise infinitely variable texts [30]. Genre acts as a template of attributes that are regular and can be systematically identified. Most important, genre reflects the purpose of documents. Therefore, if a Web search could use genre metadata, it might be possible to use it to specify the

---

K. Crowston (✉)

School of Information Studies, Syracuse University, Syracuse, NY, USA  
e-mail: crowston@syr.edu

desired information more precisely and find a document whose purpose matches the user's. Indeed, [10] reports that searches on America Online that included a genre term such as *recipe* or *lyrics* seemed to yield more precise search results. Towards this end, researchers from the fields of information science, communications and linguistics have tried during the past decade to demonstrate the efficacy and viability of tools that group Web documents – as search results or contained in hierarchical directories – in terms of Web genre (see, e.g., [3, 4, 6, 8, 18, 19]).

#### 4.1.1 What Is the Purpose of a Genre Taxonomy?

At the core of building applications that apply the notion of genre to information-provision tasks is the fundamental problem of identifying, defining, labelling and organizing the genres in a useful structure – that is, a genre taxonomy [22]. Such a structure enables several functions:

- First, it provides a controlled vocabulary that resolves the issue of variation in labelling and meaning: synonyms, acronyms, variant spellings, grammatical variants such as plurals, and so on.
- Second, a taxonomy can arrange the entities in a meaningful structure – typically a hierarchy or a faceted scheme – where the scope and definition of each entity is further described by its relationship to other entities. Thus, we can say a *digital-camera review* is a kind of *product review*, the *product review* being a more inclusive term. Other structures are possible as well, such as part-whole arrangements in which entities can be described by their componential parts. For example, an *abstract* is part of a *scholarly article*. In this case an *abstract* is a genre that is typically part of another genre, each sharing part of the functional properties that make knowing the genre of something so useful.
- A well-designed taxonomy is useful at both ends of the information-provision process. From the user's perspective it allows for a more cognitively efficient way of choosing terms for a query. Rather than “thinking of a genre off the top of your head”, a user can choose from an organized array. The organization further allows expansion of the search to more general terms, or conversely a narrowing of the search for more specificity. For retrieval, a taxonomy allows for gathering terms with similar meaning together under one label, allowing for adjustments in the granularity of the results.

Unfortunately, our review of the literature reveals a lack of consensus about the Web genre taxonomy on which to base such systems. Furthermore, our review of efforts to develop such taxonomies, reported below, suggests that consensus is unlikely. As many researchers have found, reaching consensus on genre terms, their attributes, or their relationships to each other is not easy. This difficulty applies both to genre information gleaned from “genre use in the wild” and to reaching intercoder consistency for manually marked-up genre palettes in research studies. As [25] comments, there seems to be “a gap between genre theory and the practice of average users”. The purpose of this chapter is to support this claim by first briefly reviewing

prior work on developing taxonomies of Web genres and second to describe the problems we encountered in a study aimed at developing a genre taxonomy from a user study.

## 4.2 Why Is It Hard to Develop a Web Genre Taxonomy?

As noted above, document genre can be defined as “a document type based on purpose, form and content” [21, p. 2]. A fundamental question that must be addressed to develop a satisfactory taxonomy concerns the origin of genre terms in the taxonomy. Simply put, where should genre terms come from? (A second question to address is the organization of terms, an issue we addressed in prior work [14].) We note two problems that arise in generating such terms: the difficulty of defining genres precisely and the difficulties in generating a collection of genre terms that cover a collection of documents.

### 4.2.1 Difficulties in Defining Genres

A first challenge in studying genre is that there never has been, nor is there presently, a consensus on what a genre is, what qualifies for genre status, how genres “work,” how we work with genres, how genres work with each other, or how best to identify, construe, or study genres. Genres are a way people refer to communicative acts that is understood by them, more or less, but which is often difficult to describe in its particulars. Thus, genres are recognized and used, but not so readily described and defined.

The definition of document genre we quoted above includes both socially recognized form and purpose, and it is possible to make a logical division between intrinsic genre attributes (i.e., form and content) and the extrinsic function that genre fulfills in human activities. Many studies focus on the first aspect, that is, the nature of the document genres themselves or on the attributes of the documents that will allow them to exploit genre for knowledge-representation functions. From studying non-digital genres we know that the roles of content and form inform each other. For example, if we are presented with only the empty framework of the format of a letter (heading, salutation, body, and closing) most people can identify the genre. Similarly if we are presented with the content without the form – just the text – we can still recognize it as a letter [28]. For some genres, the content is more important, but for some the form is equally so. In studying digital genres we rely not only on traditional indicators of a genre, such as specific content and form, but also new and different cues for both identifying and then analyzing and making sense of them. Above all, we recognize that any approach to attribute analysis must deal with the problem of a genre’s intrinsic multifaceted nature, that is, the cues that not only identify the genre as an artefact, but also as a medium for participation in a communicative act [14].

What has changed from formal genre models, though, is that today we recognize that an exhaustive identification of attributes, even if that were possible, may not be sufficient for a full understanding of a document's genre (as also argued by [10]). This recognition is because we have come to understand the power and primacy of the document's actual implementation in a life situation in addition to its content and technical attributes. In the realm of print documents, genres have evolved over the centuries, often slowly and gradually, occasionally suddenly, and while there may be lively discussion about when, say, a novella becomes a novel, genres in general have been relatively stable. A play remains an essentially recognizable genre despite genre-bending endeavors at various points in the history of drama. We can still easily identify the prototypical limerick, the tempo of a rousing march, or an office memo. As documents have migrated to the Web, their identity as examples of genres has also evolved. New document genres have emerged [3, 5], while older ones have blended, changed, and been incorporated into different social endeavors. Print-document genres adapted to the Web, and new electronic genres emerging frequently, appear to be shuffled, disassembled and then put together again, in a seemingly chaotic manner. Many researchers, and indeed the public at large, assume that there are significant and fundamental differences in how these adapted and new genres will now function and be used. As with many new technologies, there are fond hopes that these genres will be socially transformative, enabling better communication, as well as more flexibility and expressiveness.

The lack of a one-size-fits-all solution when it comes to Web-genre taxonomies is, in our opinion, a result of the fact that genres are frequently not construed the same way across varied communities of users. In addition, even if some are more-or-less "universally" understood (such as a home page), there is still some debate about boundaries, granularity, and definition. In other words, genres may not be as generic as we would like in terms of implementing them in applications. This is not surprising, since the very essence of what makes a genre powerful is its intimate connection to the circumstances in which it is enacted. A genre only exists in use.

Emerging from these discussions is the broader question of whether technology leads human activities or follows it. In terms of genres of digital documents, the questions that arise are whether digital genres emerge from what people do on the Web, or whether the technology itself affords ways of doing things that people can then discover and exploit. This is by no means an easy question to answer, since people have always found ways to repurpose technologies, and digital technologies are no different. What is even more difficult in the electronic environment is that many technologies are converging – voice, image, text, databases, computing-creating opportunities for combining and recombining genres of many different kinds in inventive ways and for unexpected purposes.

So, a discussion of genre is challenging for a number of reasons – among them the differences in the concept's role in various domains and the contextual nature of genre in action. Still, we find genre a useful concept because in identifying and labeling genres we try to capture the gestalt of the various components of the communicative act. This is all the more important for digital genres on the Web, since

so many socially agreed-upon cues present in traditional print documents and oral communication are no longer available to us.

### ***4.2.2 Difficulties in Developing the Scope and Expressiveness of the Taxonomy***

Beyond the issues involved in defining the boundaries of a single genre are the problems involved in developing a collection of genres to comprise a genre taxonomy that is sufficient to describe a collection of documents. There are several benchmarks of a robust taxonomy: first and foremost is the attribute of reflecting the structure of the domain, but also very important is the ability of a taxonomy to be sufficiently expressive. This means that the taxonomy comprises genres that are able to adequately represent the documents to which it will be applied. As [14] have noted, there are two basic approaches to this task of genre term production: top-down and bottom-up.

#### **4.2.2.1 Top-Down**

Many attempts to develop a categorization of genres have been top-down, that is, they analyzed a set of documents based on theoretical principles or according to a priori classifications. In a top-down approach, the researcher draws from an existing set of genres and also from knowledge and understanding of Web genres of that domain. In one study, for example, each of two researchers “add[ed] new genres to the list” where “none of the already defined genres were appropriate . . . [The] two raters agreed completely on the coding for 68%” of the documents [3, p. 205].

A key difference in these efforts is the number of genre categories distinguished. Many studies of Web pages have used fewer, broader categories: for example, [18] used only eight genres (*help; article; discussion; shop; portrayal, non-private; portrayal, private; link collection; and download*). At the other extreme, [7] offered a catalog of some 2,000 genre (or text type) terms intended to be an exhaustive list of the terms used in English. Somewhere in between, [16] categorized documents in the British National Corpus (BNC) into 70 genres or subgenres (with some document assigned more than one genre). He notes, however, that the genre terms used were “meant to provide starting points, not a definitive taxonomy”, for example grouping *textbooks* and *journal articles* as *academic texts* that can be further distinguished by medium.

In studies where taxonomy developers start with (but ultimately modify) a palette of Web genres proposed in a prior study, there is the question of which “starter palette” to use. At least two studies [17, 26] made initial use of [4]’s genre taxonomy, for example, while [3]’s taxonomy of document genres was based on the Art and Architecture Thesaurus [20] and used by [23]. This question is important methodologically because the use of any starter palette frames how Web documents in a corpus will be viewed. A researcher may end up with a new taxonomy that does

not much resemble the one she started with, but that was almost certainly influenced by the form and shape of the taxonomy. In other words, a researcher might have created a completely different taxonomy had she used a different starter palette or no starter palette at all.

Very few of these top-down studies include a discussion of the role that personal attributes (e.g., experience or expertise) play in this process, or precisely how multiple researchers reach agreement on Web genre terms. In another study, for instance, the authors tell us only that “page descriptions evolved through the course of the analysis into a system of page types” [8, p. 183].

#### 4.2.2.2 Bottom-up

In a bottom-up approach, Web users who have volunteered to participate in a study do the same thing – draw to the extent possible (and sometimes aided by tutorials) from their understanding of Web genres – to produce Web genre terms for the taxonomy (see for example [10, 22] for examples). Such an approach seems desirable because it avoids imposing an a priori vocabulary with which users may lack familiarity. As [21, p. 1054] put it, “a good genre candidate for document descriptor should be recognizable to searchers”. However, this approach relies on the ability of the users surveyed to adequately recognize and label documents by genre, which is problematic for the reasons surveyed above.

As [18] notes, “An inherent problem of Web genre classification is that even humans are not able to consistently specify the genre of a given page.” Web documents are often ambiguous, and may not resemble the exemplar of a certain genre closely enough. Crowston and Williams [3] point out that some Web documents did not have a “recognizable genre;” others seemed to instantiate an emerging genre that does not yet have a name. Indeed, the intended purpose of many Web documents is unclear, in part because of the “increasingly wide range of uses to which the Web can be put” [8]. Alternately, multiple genre terms may seem appropriate to describe a particular document. Web documents may instantiate multiple genres [3, 8]. As [24, p. 6] puts it, “genres are not mutually exclusive and different genres can be merged into a single document, generating hybrid forms.” As well, more or less specific terms may be available. For example “. . . scholarly material can be seen as a super-genre that covers help, article and discussion pages” [18]. Which do we choose and how do we decide on the granularity? Finally, many lay users are unfamiliar with the formal genre concept and, as a result, some tend to conflate genre with topic, perceived document quality (e.g., “boring pages”) or intended audience (e.g., “internal documents”) [4].

In the face of the difficulties noted above, researchers may intervene by explaining the genre concepts to participants (e.g., [18]) and/or modifying the genre terms supplied by participants (e.g., [4]). As a result, most ostensibly bottom-up taxonomy development efforts may actually incorporate elements of both top-down and bottom-up approaches. In one such study, for instance, researchers “proposed ten genre classes” then asked interviewees to “specify up to three additional genre classes” [18, p. 4].

Other recent attempts at developing a genre classification aim at discovering relevant attributes automatically, rather than identifying them a priori [1, 9, 11, 13]. These attributes are then used to cluster documents into genres. This line of research assumes that genre attributes may be too unwieldy and slippery to identify “from the top,” and that there may be too many genres in a rapidly growing and expanding field of digital documents and their implementations [5, 12, 14, 29].

### **4.3 A Use-Centered Development of a Taxonomy of Web Genres**

We turn now to describing our own efforts at building a taxonomy of genres based on a user study of Web document use. We first describe the research design and data elicitation and analysis methods we adopted before briefly discussing the results of our study. We then present the main challenges we faced in the study and its resulting limitations as the basis for a genre taxonomy.

#### ***4.3.1 Research Design: Naturalistic Field Study***

Our goal was to develop a better understanding of the use of genre in information-access tasks and then to develop a human-centered taxonomy of genres for use in subsequent phases of the overall research plan (a full description of the projects is beyond the scope of this chapter). Because genres are situated in a community’s language and work processes, we felt it was important to learn about genres from people engaged in real tasks, and in their own words. We considered a top-down approach using a researcher-generated or standard list of genres as problematic for two reasons. First, genres are socially constructed, so different social groups using documents with similar structural features may think about them and describe them differently. A document may be unfamiliar and difficult to understand for someone outside of the community in which the genre is used, so it is important to capture the users’ own language and understanding of these genres. Second, it is imperative to extend any investigation to genres that are not necessarily vetted by traditional schemes, such as those that come out of domain-specific work (e.g., block-scheduled curriculum plans). As pointed out by [5, p. 202], genres are no longer necessarily “slow-forming, often emerging only over generations of production and consumption”. Thus, we assumed that a traditional typology of genre or document forms would not be sufficient to describe the emerging and dynamic genres identifiable by users in general and our study community in particular.

#### ***4.3.2 Research Informants***

Knowing that we could not study the universe of Web genres or searchers, our first task was to identify respondents who would, in the course of their daily work, need to search on the Web, and who most likely would want to distinguish between

**Table 4.1** Our source of genre information: three groups of respondents

Respondents	No.	Typical tasks	Typical genres	Comments
Teachers	15	Preparing and revising lesson plans	Lesson plan Story page Resource page	Teachers from four public and private schools; most grades from K-12 are represented
Journalists	20	Developing a story or article: generating ideas; searching for other stories on the same topic; collecting new information; fact-checking	News story Directory Press release	18 print journalists, 2 television journalists
Engineers	20	Searches for tutorials, detailed information about products and tools, new or updated “knowledge” about a topic	Manual page Commercial page Product page	Includes 20 aeronautical and software engineers from one multinational firm

one type of Web page and another. That is, we tried to identify people for whom genre information might be useful – indeed necessary – for determining whether a given Web page might be relevant to their needs. (Because we recognized that the genre terms elicited would likely be somewhat specific to the groups studied, we planned to use the same communities in later phases of the research plan.) Our study solicited information about genre from three groups of respondents: K-12 (kindergarten through grade 12, i.e., primary school) teachers, journalists and engineers, as summarized in Table 4.1. We chose these three groups because the members of each share a discourse community in which a set of identifiable tasks and genres may play a role, and in which the identification of the genre of a document was thought likely to be important for their tasks.

Respondents were recruited via a snowball-sampling approach, chosen to fit our goal of collecting a wide range of tasks, genres and genre attributes. (A more systematic sample would have been required for making inferences to a population, e.g., for documenting the relative frequency of use of terms, but that was not our purpose in this study.) All respondents were working full-time in one of these three professions and had the required educational background to do so, making them qualified to identify genres relevant to their work. Ages ranged from early twenties to late fifties; 40% were female and 60% male.

### 4.3.3 Data Elicitation

In general, our data-elicitation goal was to identify, for a collection of Web pages, the genre (or genres) of the page, the clues each respondent used to recognize the genre (or genres), and the usefulness of the page for a task, all in the words of the respondents. We used think-aloud technique to understand the search goals and

general strategy, but then followed it with a debriefing. These interviews were carried out in the respondents' offices, using their own computers. Respondents were asked to carry out a Web search for a real task of their own choice (e.g., a journalist searching for background information on an interview subject; an engineer looking for software documentation). During the interview, for every page visited we asked four questions:

1. What is your search goal?
2. What type of Web page would you call this?
3. What is it about the page that makes you call it that? (If they did not understand the question, we would ask, "Which features/clues on the page make you call it that?")
4. Was this page useful to you? How so (or why not)?

At the conclusion of the debriefing, and with permission from the respondent, we copied the URLs of the Web pages visited and the sequence in which they were visited. These data were used to re-create the search process. From this re-creation, screenshots were taken of each Web page visited by the respondent, and a Web-based slide show (with accompanying URLs) of the entire sequence was created for each session. We are able to use this for coding and analysis, and intend to draw from these slide shows to develop a corpus of Web pages that a subsequent set of respondents can view and evaluate. We have nearly 1,000 screenshots of Web pages visited by respondents, each accompanied by its original URL and digital audio recordings of the sessions with transcripts, or detailed field notes for those interviews where recording was not permitted.

#### **4.3.4 Data Analysis**

Content analysis was employed for identifying genre terms. We analyzed:

- The captured Web pages.
- Transcripts of audio files from the debriefing for the 32 respondents – 19 journalists and 13 teachers (3 of the original transcripts were corrupted by problems with the digital recorder and could not be used).
- We also content analyzed the detailed field notes for 20 engineer respondents where audio recording had not been permitted.

First, we collected the terms used in answer to the question: "What type of Web page would you call this?" We transcribed the terms as given to us, without making a judgment about whether it was a legitimate "genre" or not. In other words, we allowed the respondent to identify the candidate genre terms for the analysis. Respondents had the option of offering multiple terms for the same page.

Before calculating the frequency, we made a few changes to some genre terms which we call "trimming." This included merging terms with inflectional differences or derivational forms of a word. For example, class note was merged to class notes,

and governmental page with government page. As well, we considered both list of stories and list of articles as simply a list for frequency analysis.

Using the following rules, we further reduced the list of terms, bearing in mind that our goal was not so much to compile an exhaustive list or a taxonomy that represented a particular domain, but rather to build a taxonomy to use in subsequent stages of the research with these groups. We also wanted the taxonomy to be used eventually with a general audience. Thus we needed genre terms that we believed would be understood by our future study participants, who might not be from the same exact discourse communities as the participants in this study. Thus, we eliminated:

- Terms that had only a personal meaning to the respondent, e.g., “good page.”
- Terms that were so situation- or domain-specific that they would not be understood in any other context, e.g., an “uncontrolled resource page” from an engineer.

## 4.4 Results

We collected 226 genre terms from 20 engineers, 404 from 19 journalists, and 137 from 13 teachers for a total of 767 genre term tokens from the 52 subjects. The total of genre types (unique terms ignoring repetitions) was 522 (167 from engineers, 262 journalists, and 93 teachers). The count of genre terms is shown in Table 4.2. Table 4.3 shows the final number of genre terms following the trimming of variants and the elimination of terms we deemed not useful for the purposes of our study. Common genre terms across the populations studied are shown in Table 4.4, while Table 4.5 lists terms that were unique to particular groups.

**Table 4.2** Raw numbers and averages per respondent of candidate genre terms

	Engineers	Journalists	Teachers
Respondents	20	19	13
Genre term tokens	226 (11.3)	404 (21.26)	137 (10.53)
Genre term types	167 (8.35)	262 (13.78)	93 (7.15)

The numbers in parentheses indicate average genre terms per respondent.

**Table 4.3** Results of trimming and selection

	Original token	Genre terms type	Trimmed token	Genre terms type	Selected token	Genre terms type
Engineers (20)	226	167	226	131	127	104
Journalists (19)	404	226	404	209	191	150
Teachers (15)	137	93	137	70	62	44
Total	767	522	767	410	380	298

**Table 4.4** Examples of common genres

Common to E J T	Common to E J	Common to J T	Common to E T
Article	About us page	Education page	Book
Government page	Advertising page	Front page	Commercial
Home page	Blog	Gateway	Page
Index	Company home page	How-to page	Journal article
Information page	Corporate page	Link page	Magazine
List	Definition page	Newspaper	Resource page
Main page	Entry page	Organization page	Organization
Search engine	FAQ	Full story list/list of page/Stories	
Search page	Letter		Organization
Search results	List of links	Magazine/magazine Article	Home page
Site map	Navigation page		
Summary	Organization home page		
Table of contents	PDF		
Magazine/magazine	Press release		
Article	Question and answer		
	Terms and conditions		
	Archive of abstracts/archives		
	Executive overview/Overview		
	Magazine/magazine Article		
	Meeting notes/Minutes		

E = Engineers, J = Journalists and T = Teachers.

## 4.5 Discussion

Even though we learned a great deal about studying genres in the field and about the differences in genre use by our three respondent groups, in the end, we were disappointed with the results of our study with respect to its usefulness in building a taxonomy of genre terms for further application. We discuss these challenges briefly here and in more detail in [15]:

1. Difficulties with identifying the genre unit. A Web page can be composed of one or more elements, each of which can be construed as a stand-alone genre by itself. For example, a Web page was described as both an *article* and a *newspaper*. In these cases, it was sometimes difficult to ascertain from the interviews which part of the page had the genre that was being described. For example, homepages were often described as both a *homepage* and an *index page*, presumably because homepages often have a list or an index of links embedded in the Web page. One Web page that consisted of a search box, search directory and other related links was described as both a *search engine* and *search directory*, these labels being dependent on the emphasis of a different element of the page.

**Table 4.5** Examples of unique genres

Engineers	Journalists	Teachers
Change summary page	Editorial (2)	Activity
Coding manual	Fact box	Lesson plan (3)
Compiler listing page	Gray page	Lesson resource
Compilers home page	Index of news coverage	List of course
Data (3)	Index to the news stories	Offerings
Datasheet	Interview	List of lesson plan
Directory to white papers	List of headlines	Outline of a Textbook
Explanation of the code	News blog	
Library (2)	News entry	
License	News page (2)	
Man page (3)	News portal	
Manual	News release	
Online manual	News story	
Software description page	News summary page	
Software test document	Press release	
Standards	Press resources page	
Technical committee report	Story (2)	
Technical paper	Story list	
Test plan (2)	Transcript of an interview	
White paper		

2. Difficulty with eliciting unambiguous genre labels. We learned that the genres of some types of Web pages are more difficult than others for respondents to articulate. For example:
  - Multiple genre terms were applied to one document. Several genre terms (both conceptually similar and different), might be suggested for one Web page as respondents struggled to find an appropriate term. For example, one page was described as a “first-search-step” page, “navigation page”, and “menu” with the comment “I don’t know if I have the vocabulary to describe it.”
  - Different types of pages were labeled with same genre term. In the iterative process of asking for genre terms, respondents had a tendency to use some words repeatedly. One respondent described a page as a *highlights* page since she saw the word “highlights” on it. Later, she used the same term to describe what to us seemed to be a *memo*, a *news release*, a *calendar page*, and so on.
  - The respondent lacked a term for a given genre. When respondents could not easily name a genre, it was either because they could not think of the term or because they didn’t know if a term exists. In the first case, a respondent may just describe the page based on a personal feeling, such as calling it a “frustrating page”, or admit to not having a word for the page.
  - Terms were too general or unspecific. When a genre term does not come readily to mind, respondents often provide a general or vague term such as, a “page with information”.
3. Difficulties with identifying genre attributes. We wanted the respondents to identify the criteria by which an entity (in our case a Webpage genre) is aggregated

with like entities or differentiated from unlike ones. We expected respondents to identify genre based on document attributes of form, content and purpose. However, participants were often vague about clues to these attributes. For instance, they might refer to a page as having a “look and feel” but not specifying in what way. Since journalists are very familiar with the format of a *news story* page, for instance, they are good at identifying that genre; however, they may have difficulty specifying the clues that helped them identify it because such clues have become implicit and they barely pay attention to them.

4. Challenges in distinguishing form and content. In coding we first flagged the genre term applied to a Web page, and then tried to mark the clues the respondents identified in establishing their concept of that genre. Marking clues in a consistent manner according to the tripartite definition of form, expected content and purpose has not been easy, however. The first two aspects are often convolved in the participants’ utterances where it is difficult to ferret out both what they mean or what is in their minds when they invoke a genre term. This convolution of form and content has three manifestations:
  - Identifying aspects of key page elements that signify a page belongs to a genre. For example, one participant invoked a *municipality* genre, and using the municipality’s seal as a clue. How much of a simplified seal “form” would have been enough to qualify it as a *municipality* page? Or, was she looking at the particular “content” of the seal that made it specific to a municipality of interest?
  - The mixture of form and content in total that establish a page as part of a genre. For example, a participant readily assigned a genre term based on the presence of tabs that allowed for presentation of categories and subcategories. Was it the form of the page, with spatial separation of categories and less visual emphasis given to the subcategories that mattered to him? Or, was it the contextual relationships among the written material on the Web page to which he was referring?
  - Our own preconceived notions of what these “form” and “content” concepts mean. Achieving consistent coding for clues has been difficult when coders bring different conceptions to the task. For example, in deciding on whether an image represented form or content, one coder interprets the meaning of the image and calls it “content,” while the other coder, interprets an image as pure “form.”
5. Challenges in identifying purpose. One of the key ways in which genre provides context is by incorporating an understanding of the genre’s purpose or function. While most of the respondents can identify the purpose of the Web page for their own work it is not always clear whether the task requires a particular genre or whether the genre identified happens to be useful (but another one could have been just as useful).
6. Borrowed purpose. Another situation that causes some confusion is the difficulty in assessing whether the purpose of a genre is generated by the respondents’ situation, or whether they recognize the purpose others have for that genre. The

*homepage* of a university that is described as an *institutional* page has several purposes depending on the perspective of the user. The purpose of the page from the institution's perspective is to "get its message out," while from the perspective of students and their parents, its purpose is to provide different kinds of information about the university.

7. Granularity of tasks. We are finding that people's tasks, as well as the genres that are useful for them are at various levels of specificity. Some are expressed broadly, such as "double-checking facts," while some are narrowly defined, such as "finding the phone number of Joe Smith."

## 4.6 Conclusions

In summary, in our study we discovered how difficult it is to study genres "naturalistically." At the same time, we also learned that this is an area of great promise. Rather than trying to study the genres themselves, researchers can instead study human activity through genres, especially those activities that focus on communication [27]. This is, obviously, not new. We have studied diaries and letters for many hundreds of years for what they reveal about their writers and the times they lived in. Others have looked at epitaphs, songs, and political slogans. These texts are useful because they can be studied not only at the level of what they say, literally, but what they convey at many other levels. Genres are consensually created and thus they capture not only the meanings of the individual, but also the meanings of the community in which that text is used.

As a result, genre provides an excellent lens for discourse analysis – that is the analysis of language in use in a given community. This type of analysis strives to understand not only the words, per se, but the contexts in which those words acquire meaning. So, for instance, a discourse-based study of rap-music lyrics reveals the culture in which they are created, as well as the values held by the artists and fans. The *rap-music* genre captures this culture and reveals it simultaneously.

In this vein, we have noticed that several factors that may determine the identification and use of Web genres as well as their place in an overall conceptual map of genres, which our taxonomies try, but fail, to capture. Among these are such factors as the professional affiliation of the person identifying the genre as well as their familiarity with the function for which the genre was created. Most interestingly, though, we have picked up hints – no proof – that perhaps a strong correlation can be made between tasks and genre. That is, perhaps we could structure our Web-genre taxonomies in part by the types of tasks for which a given genre might be useful.

There are many unanswered questions, of course. At the top of the list is the big question of whether a searcher can identify the type of task he or she is contemplating, and second, is the question of whether there is a way of mapping the genres onto the task types in such a way that there is some flexibility and room for individual search strategies. Nonetheless, even a small improvement in the effective use of genre information would be welcome.

**Acknowledgment** This research was partially supported by NSF IIS Grant 04-14482. We thank John D'Ignazio and You-Lee Chun for their contributions to this research project.

## References

1. Bagdanov, A., and M. Worrington. 2001. Fine-grained document genre classification using first order random graphs. In *Document analysis and recognition*. Seattle, WA: IEEE Computer Society.
2. Bartlett, F. 1932/1967. *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.
3. Crowston, K., and M. Williams. 2000. Reproduced and emergent genres of communication on the world wide web. *Information Society* 16(3):201–215.
4. Dewe, J., J. Karlgren, and I. Bretan. 1998. Assembling a balanced corpus from the internet. In *11th Nordic Conference of Computational Linguistics*, 28–29 Jan 1998. Copenhagen, Denmark.
5. Dillon, A., and B. Gushrowski. 2000. Genres and the web: Is the personal home page the first uniquely digital genre? *Journal of the American Society for Information Science* 51(2):202–205.
6. Freund, L., C.L.A. Clarke, and E.G. Toms. 2006. Towards genre classification for IR in the workplace. In *Proceedings of the 1st International Conference on Information Interaction in Context*, 30–36. Copenhagen.
7. Görlach, M. 2004. *Text types and the history of English*. Trends in Linguistics: Studies and Monographs 139. New York, NY: Mouton de Gruyter.
8. Haas, S.W., and E.S. Grams. 2000. Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science* 51(2):181–192.
9. Karjalainen, A., T. Päiväranta, P. Tyrväinen, and J. Rajala. 2000. Genre-based metadata for enterprise document management. In *Proceedings of the 33rd Hawaii'i International Conference on System Sciences*. Wailea, Maui, Hawaii.
10. Karlgren, J. 2010. Conventions and mutual expectations: Understanding sources for web genres. In *Genres on the Web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
11. Karlgren, J., and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto.
12. Kennedy, A., and M. Shepherd. 2005. Automatic identification of home pages on the web. In *Proceedings of the 38th Hawaii International Conference on System Sciences*. Waikoloa, Hawaii, HI.
13. Kessler, B., G. Nunberg, and H. Schuetze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*, 32–38. Madrid: Morgan Kaufmann Publishers.
14. Kwaśnik, B.H., and K. Crowston. 2004. A framework for creating a faceted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Hawaii'i International Conference on System Science (HICSS)*. Big Island, Hawaii'i.
15. Kwaśnik, B.H., Y.-L. Chun, K. Crowston, J. D'Ignazio, and J. Rubleske. 2006. Challenges in creating a taxonomy for genres of digital documents. In *2006 ISKO Conference*. Vienna.
16. Lee, D.Y.W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language, Learning & Technology* 5(3):37–72.
17. Lim, C.S., K.J. Lee, and G.C. Kim. 2005. Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management* 41(5):1263–1276.

18. Meyer zu Eissen, S., and B. Stein. 2004. Genre classification of web pages: User study and feasibility analysis. In *Proceedings of the 27th Annual German Conference on Artificial Intelligence (KI 04)*, eds. S. Biundo, T. Frühwirth, and G. Palm, 256–269. Ulm: Springer.
19. Nilan, M.S., J. Pomerantz, and S. Paling. 2001. Genres from the bottom up: What has the Web brought us? In *Proceedings of the American Society for Information Science and Technology Conference*, 330–339. Washington, DC.
20. Petersen, T. 1994. *Art and architecture thesaurus*. New York, NY: Oxford University Press.
21. Rosso, M.A. 2008. User-based identification of web genres. *Journal of the American Society for Information Science & Technology* 59(7):1053–1072.
22. Rosso, M.A., and S.W. Haas. 2010. Identification of web genres by user warrant. In *Genres on the Web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
23. Roussinov, D.G., and H. Chen. 2001. Information navigation on the web by clustering and summarizing query results. *Information Processing and Management* 37(6):789–816.
24. Santini, M. 2008. Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing and Management* 44(2):702–737.
25. Sharoff, S. 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
26. Stubbe, A., C. Ringlstetter, and K.U. Schulz. 2007. Genre as noise – noise in genre. In *Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*. Hyderabad.
27. Swales, J.M. 1990. *Genre analysis: English in academic and research settings*. New York, NY: Cambridge University Press.
28. Toms, E.G., D.G. Campbell, and R. Blades. 1999. Does genre define the shape of information? The role of form and function in user interaction with digital documents. In *American Society for Information Science; ASIS '99*. Washington, DC: Information Today.
29. Watters, C., and M. Shepherd. 1999. Cybergenre and web functionality. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences*. Maui, HI: IEEE Press.
30. Yates, S.J., and T. Sumner. 1997. Digital genres and the new burden of fixity. In *Hawaiian International Conference on System Sciences (HICCS 30)*. Wailea, HA: IEEE Computer Press.



**Part III**  
**Automatic Web Genre Identification**

# Chapter 5

## Cross-Testing a Genre Classification Model for the Web

Marina Santini

### 5.1 Introduction

The main aim of the experiments described in this chapter is to investigate ways of assessing the robustness and stability of an Automatic Genre Identification (AGI) model for the web. More specifically, a series of comparisons using four genre collections are illustrated and analysed. I call this comparative approach *cross-testing*. Cross-testing exploits existing genre collections that are publicly available,<sup>1</sup> which have been built for individual needs and shared by their creators, thus allowing constructive comparative experiments. Thanks to these, big steps forward have been made in the last few years, regardless the absence of official genre benchmarks and test collections. Yet, the current state of AGI is one of fragmentation and tentativeness, and automatic genre research still lingers in the starting phase.<sup>2</sup>

The lack of benchmarks and test collections is only one of the reasons behind AGI cautious progress. Other reasons are well summarized in the five points listed by Sharoff (in this book) and discussed, from different perspectives, by all the other authors contributing to this volume, namely (1) the lack of an established genre list, (2) the unclear relation between traditional and web genres, (3) the need to classify large quantities of web documents quickly, (4) the design of the genre inventory, and (5) the identification of emerging genres.

From the outside, one might wonder why it is so difficult to harness and create consent among textual categories that we habitually employ in our everyday

---

M. Santini (✉)  
KYH, Stockholm, Sweden  
e-mail: marinasantini.ms@gmail.com

<sup>1</sup> Many of them are available through the WEBGENREWIKI, at <<http://purl.org/net/webgenres>>. Copyright of genre collections built with web material may vary according to national laws. The copyright of the web pages contained in the genre collections used in this chapter is held by the author/owner(s) of the web pages. These web pages are used for research purposes only.

<sup>2</sup> A quite long initial phase, if we consider that this research field was initiated in 1994 with Karlgren and Cutting's extensively cited paper based on the Brown Corpus and discriminant analysis [18].

life. Who is not familiar with one or more of the following genres: EDITORIALS, INTERVIEWS, LETTERS TO THE EDITOR, CLASSIFIED, WEATHER REPORTS etc. in newspapers and magazines; or BLOGS, FAQs, HOME PAGES, PERSONAL PROFILES, ACADEMIC PAPERS, SUGGESTIONS, HINTS, DIY GUIDES, HOW-TOS, NARRATIVES, INSTRUCTIONS, ADVERTISING, etc. on the web or other digital environments? Surprisingly, findings show that agreeing on the genre labels to be applied to documents is not so straightforward as one could imagine [23, 30].

From a terminological point of view, there is a great variation in the use of genre labels. There are problems with synonyms, with similarity between and across genres, with the level of generality or specificity that genre labels represent and so on. In these conditions it is very difficult to make decisions about the definition of a genre palette. Experience from Meyer zu Eissen and Stein [22], Rosso [24] and Chapter 4 by Crowston et al. (this book) shows the problems related to the definition of genre taxonomies. Connected to the terminological elusiveness is the problem of genre evolution. New genres are spawned continuously in web communities (e.g. see Chapter 13 by Paolillo et al., this book), and social networks have certainly novel genres in store for us. Facebook WALL or LinkedIn's PUBLIC PROFILE seem to be good candidates in this respect. Some ideas on how to detect new genres have been mentioned, e.g. by Shepherd et al. [34] who suggest adaptive learning. However, human acknowledgment of new or evolving genres might be slower than their automatic detection, since genres require social recognition, at least within the community where a new genre is envisaged (e.g. cf. the historical analysis of BLOG creation in Blood [4]).

From a perceptive point of view, experiments have shown that individuals have a differing perception and recognition of genres [24, 30]. The low understanding of how genres are perceived and how their labels are used by humans is a major drawback for genre annotation tasks, since raters tend to disagree when deciding on the genres to assign to documents. In this respect, the experience by Berninger et al. [1] and Chapter 7 by Sharoff (this book) are very instructive. Both experiences show that it is both difficult to instruct people consistently and to get strong agreement. Only the good will of the corpus builders, lots of dedicated time, financial resources and, last but not least, resolute and clear-cut decisions led to the finalization of KRYIS-01 and Sharoff's English-Russian genre collections.

There is also a problem of sheer classification. The genre classes mentioned above cannot be levelled to one dimension. There are hierarchical relations and horizontal relations (cf. [14]). For instance, we can deal with supergenres (e.g. ADVERTISING), subgenres (e.g. WEATHER REPORTS), genres at basic level (e.g. EDITORIALS),<sup>3</sup> etc. In which way these different levels of generality affect an AGI classifier? Some experiments have shown that an AGI classifier performs better when the level of classes is consistent. However, we currently know very little about the relation between the granularity of genre classes and classification performance.

---

<sup>3</sup> Cf. Lee [19] for the application of these three levels following the prototype theory to genre.

Another compelling issue concerns the ontological nature of genre. It would be intriguing to delve into the special traits that distinguish genre from other textual categories, such as topic, domain, style, registers or sentiment. Although a good attempt to shed some light on these relations has been made by Lee [19] for the genre annotation of the British National Corpus (BNC), one practical solution is to conflate all textual categories into the catch-all term “text categories”, in line with the Lancaster-Oslo/Bergen Corpus (LOB)<sup>4</sup> or Brown corpus<sup>5</sup> built about 50 years ago.

It is undoubtedly true that the term “genre” is loosely applied in everyday life to a number of conceptually heterogeneous classes, as highlighted in Chapter 2 by Karlgren (this book) underpins through the analyses of Yahoo! directory. His claim is also supported by booksellers’ catalogues. For instance, under the tab “Browse Genres”,<sup>6</sup> Amazon UK lists many disparate categories, from genres (i.e. BIOGRAPHIES AND MEMOIRS) to mere descriptive labels (i.e. Subjects within Arts). As a matter of fact, one trend in AGI experiments to date has been the separation of genre from other textual categories and, above all, from topic. Many authors argue that topic and genre are orthogonal to each other (see Chapter 8 by Stein et al., this volume). However, others, like Vidulin et al. [41] experiment with mixed textual categories, from genres like FAQs or ERROR MESSAGES to subjects like “Childrens”, to functions like “Gateway”, to less transparent labels like “Content delivery” (see Table 5.18 for the description of these categories). Some correlations between genres and other activities have been explored, e.g. the one between genre and tasks [11] through an ad-hoc built corpus. Conflating genre and topic into the anodyne label of “document types” is common practice in IR (e.g. see [44–47]), although the collections they used are mostly topical. Interestingly, though, corpus linguists studying language variation have shown recently that subject categories are not clearly well defined on linguistic grounds, as shown by the findings by Biber and Kurjian [3] who have applied multi-dimensional analysis to two Google categories, i.e. Home and Science.

In short, as this abbreviated list of issues shows, there is an ongoing heated discussion in AGI research. One practical shortcoming of this never ending debate is the absence of common and shared genre resources. This lack hinders the creation of agreed upon genre framework, thus affecting the progress of AGI research.

A temporary remedy to this lack is the practice of cross-testing. This practice has been possible because some researchers have shared their own collections within the genre community. This has allowed a number of comparative experiments (some of them are listed in Section 5.5) that provide insights into AGI problems.

---

<sup>4</sup> See <[http://en.wikipedia.org/wiki/LOB\\_Corpus](http://en.wikipedia.org/wiki/LOB_Corpus)>, retrieved April 2009.

<sup>5</sup> See <[http://en.wikipedia.org/wiki/Brown\\_Corpus](http://en.wikipedia.org/wiki/Brown_Corpus)>, retrieved April 2009.

<sup>6</sup> See <[http://www.amazon.co.uk/Books-Categories/b/ref=sv\\_b\\_1?ie=UTF8&node=1025612](http://www.amazon.co.uk/Books-Categories/b/ref=sv_b_1?ie=UTF8&node=1025612)>, retrieved April 2009.

In this chapter, I will leverage on the practice of cross-testing to assess the robustness of a simple genre classification model<sup>7</sup> that will be described in the next sections. This model is provocatively simple and is used maieutically here to show that genre can be captured with high accuracy (i.e. 86–96%) with any kind of features (from high-level linguistic attributes to low-level byte n-grams) and any kind of algorithms (from the elementary inferential/rule-based approach described in this chapter to sophisticated statistical/mathematical methods) when genre models are evaluated in a restricted and relatively clean *in vitro* settings. When one attempts to approximate the population of web genres by introducing lot of noise and many different characterizations of genre classes, it is difficult to understand the significance of the results. In short, the experiments described here show that the diverse definitions of the concepts of genres have a strong bearing on the characterization of genre classes, thus affecting the generability of AGI models as a whole. Ultimately, this chapter is nothing more than a strong encouragement to investigate more extensively the robustness of AGI models for the web in less conventional experimental settings in the future.

The chapter is organized as follows: Section 5.2 lists and describes the genre collections employed to cross-test the model; Section 5.3 briefly explains the genre palette and the features; and Section 5.4 presents the model and its motivation. In Section 5.5 – the most articulated part of the chapter – results are reported. First, the model is cross-tested on four genre collections independent from each other on a single label (Sections 5.5.1, 5.5.2 and 5.5.3). Then, the accuracy of the model on multilabelling is tested (Section 5.5.4). Section 5.6 contains the discussion of the findings and Section 5.7 concludes the chapter.

## 5.2 Approximating Genre Population on the Web

Since the web is in constant flux, it is almost impossible to compile a representative corpus/sample of the web as a whole (the multi-lingual web), or only of a single language, like the English web. There are estimates of the number of indexed web pages,<sup>8</sup> which is a daily growing number, but we do not know anything about the proportions of the different types of text on the web. Interesting approaches have been proposed to automatically create corpora from the web, but these methods are biased towards the construction of corpora having topic or domain as priority, rather than genre. From a statistical point of view, when the composition of a population is unknown, the best solution is to extract a large random sample and draw inferences from that sample. However, deciding the size of this random sample is not a trivial issue. In this chapter I temporarily override this problem by using some available genre collections to cross-test the model performance. Although the total amount of

---

<sup>7</sup> This model has already been presented to the genre community with a partial evaluation in Santini [27, 29, 33].

<sup>8</sup> In April 2005 – when the genre model described in this chapter was designed and built – Google could search 8,058,044,651 web pages.

the web pages of the combined genre collections used here is only 6404 (virtually a drop in the web ocean), this amount is the largest ever used in AGI experiments with one exception, namely the CMU genre corpus [6]. This corpus, containing 9705 documents divided into seven genres without any noise, is no longer available and, as far as I know, it has never been used in other experiments. I conjecture that the composite corpus of 6404 web pages described in this chapter well represents a noisy environment like the web, where documents come from disparate communities, enact different genre conventions and classification schemes, and not necessarily belonging to a recognized genre.

### 5.2.1 Noise

The impact of noise<sup>9</sup> on genre classification results has been little explored in AGI. The only explicit investigation was carried out by Shepherd et al. [34]. They compared the performance of two classifiers on three subgenres (i.e. 93 PERSONAL HOME PAGE, 94 CORPORATE HOME PAGE and 74 ORGANIZATIONAL HOME PAGES) with and without noise (i.e. 77 non-home pages). Predictably, they results show that there is a deterioration of performance when noise is introduced. In their case, the “noise” was represented by documents that did not fall into the three subgenres to be identified, but belonged to other genres. In their experiment, Shepherd et al. [34] conflated into one single class all the genre classes not being “home pages”. Decisions about the size and the proportion of this class were not underpinned. I will call this type of noise *structured noise* because this noise is represented by well-defined genre classes that should always be a negative for a classifier.

A slightly different approach to structured noise is used Chapter 6 by Kim and Ross (this book) and Vidulin et al. [41]. Kim and Ross considered the 24 classes of KRY5-01 as noise with respect to the performance of their classifier on the 7-webgenre collection. In their case, noise is represented by 24 well defined genre classes, each of them represented by a relatively small number of documents (at most 90), while the 7 web genres are represented by 190 web pages each. Noisy classes represent around 60%, and the 7 web genres about 40%. The size and the proportion of this structured noise are not underpinned by any hypotheses. But accuracy results on the 7-webgenre collection is very good (see Table 5.5).

Interesting information on the impact of the proportion of structured noise can be derived also from Vidulin et al. [41], though their corpus is quite small with respect to the number of classes (see the description of MGC below). Their genre palette is supposed to represent all the genres on the web (but their proportions seem to be arbitrary). They build 20 individual subclassifiers and perform a binary classification, i.e. one class against the remaining 19. These 19 classes can be considered as

---

<sup>9</sup> The concept of “noise” can be applied to different situations. For example, while in Stubbe et al. [37] “noise” refers to orthographical errors, in the present study “noise” refers to documents that straddle to more than one genre and to documents that belong to no genre.

a kind of structured noise. In this scenario, Vidulin et al.'s [41] accuracy results are high (94%),<sup>10</sup> while their F-measure average on 20 genres is moderate (50%).

*One problem with structured noise is that it requires a major annotation effort because all the classes that the supervised classifier should consider as negative examples must be clearly defined and labelled.* Additionally, the underlying hypothesis is quite compelling because it presupposes that all documents fall into well-defined genres. As it will be stressed in Section 5.4, this is not always the case with documents on the web. Many web documents might simply not belong to any genre or embody several genres. For this reason, I tried to explore whether *unstructured noise*, which is pervasive in real-world conditions, can be handled in some way.

SANTINIS (see next section) incorporates the unstructured noise of the SPIRIT sample. The initial observations formulated in 2005 were that it is difficult to annotate by genre the whole web or a large or representative slice of it (due the annotation problems described in the Introduction), so the challenge is to devise a classification model robust to the *unknown*. The classification model presented here was provocatively designed with this idea in mind and in reaction to the fully supervised Machine Learning (ML) approach, which I employed extensively in other experiments.

The composition of the SPIRIT sample has remained completely *unknown* up to very recently. But I carried out a preliminary annotation in summer 2008 to have some insights of this *unknown*. My annotation reveals that the SPIRIT sample contains a large proportion of web pages that could not be labelled either because I did not know if they belong to any genre at all, or because genre labels did not come to my mind. The difficulty of formulating or uttering a genre name is the main drawback of a genre labelling activity carried out outside any real need, context or task.

The SPIRIT samples contain *also* some genres that are in the model's palette. This would be confusing for a standard fully supervised ML classifier, which can only handle structured noise.

### 5.2.2 Description of the Corpora Used for Cross-Testing

The four genre collections briefly described below have been built by different people, for different purposes, having different priorities in mind. In the break down that follows, I will point out the main differing characteristics, with respect to *composition and size, collection method and annotation, main purpose, assumptions or hypotheses, and noise*.

---

<sup>10</sup> As the authors point out “By splitting the multi-labeled ML problem into 20 binary sub-problems, we got 20 unbalanced data sets with high numbers of negative and low number of positive examples. Sub-classifiers that would recognize only negative examples would still be highly accurate” [41].

### 5.2.2.1 SANTINIS

*Composition and size.* Santini’s web corpus (henceforth SANTINIS) includes the BBC web genre corpus and the 7-webgenre collection, which is, together with KI-04 (see below) a de facto standard in AGI. The four BBC web genres (20 web pages each) are: EDITORIALS, DIY MINI-GUIDES, SHORT BIOGRAPHIES and FEATURE ARTICLES. The seven novel web genres (200 web pages each) are: BLOGS, ESHOPS, FAQs, FRONT PAGES, LISTINGS, PERSONAL HOME PAGES and SEARCH PAGES. Language: English. BBC and novel genres represent the known part of the web, i.e. about 60% of the sample. The SPIRIT sample contains 1,000 random English web pages extracted from the SPIRIT collection [15]. The SPIRIT sample amounts to about 40%. It is chronologically older than the rest of the SANTINIS (it was crawled in 2001) and represents the unknown and unclassified part of the web. The selection of the genres and the proportions of the different parts are purely arbitrary. This corpus was created in 2005. See Table 5.15 in the Appendix.

*Collection method and annotation.* The annotated part of this collection has not been manually labelled. The collection has been collected and annotated applying the principles of “objective sources” and “consistent genre granularity” [26]. The concept of being “objective” does not refer to any undeniable self-evident reality (if such a thing exists). It refers to social or public behaviour and naming habits. Basically, the principle of “objective sources” exploits the socio-cultural aspect of the concept of genre. In simple words, it relies on the membership of web pages in genre-specific archives or portals and uses their membership in these containers as evidence of an automatic membership in a specific genre, no matters who and how many decided that a certain web pages is an appropriate member for a specific archive. In order to avoid biases, it is safe to download web pages from several independent genre-specific archives or portals. Also the title of the documents can be used as public acknowledgment of a certain genre. The “objective sources” that I used to build this collection were then selected on the basis of the *genre names* included in the palette that I wished to explore (see Section 5.3.1). For example, the PERSONAL HOME PAGE genre class was selected from URL or archives containing the string “personal home page”.<sup>11</sup> Arguably, a genre collection annotated by objective sources tends to be more representative for intra-genre variation and closer to real-world conditions than a collection annotated relying on the genre stereotypicality that two or a few more people have in mind. Additionally, annotating a collection using objective sources is faster. The genre labels derived with the principle of “objective sources” do not exclude the co-existence of other genres in a web page.

The principle of “consistent genre granularity” relies on the intuition that an AGI classifier performs better when level of generality of genre classes is consistent. This collection has been built with classes at basic level (cf. the prototype theory summarized in Lee [19]), because this is the level, according to the prototype theory, where genre classes are better acknowledged and discriminated (cf. [19]). The

---

<sup>11</sup> The list of objective sources is listed in Santini [29, Appendix A].

class LISTING, however, is a super-genre included in the collection for experimental purposes (see [26]).

*Main purpose, assumptions or hypotheses.* The main purpose of SANTINIS composition is to provide a noisy environment (presumably similar to the real web) to assess the performance of an AGI classifier. As described in the previous paragraphs, SANTINIS is heterogeneous in many ways. Regardless the labels assigned and the selection method, the underlying assumption is that each web page might belong to zero, one or multiple genres. However, the performance is assessed on the available labels. These labels might be refined and augmented in the future.

*Noise.* In this collection, noise can be paraphrased as “DON’T KNOW”. Basically, the SPIRIT sample included in SANTINIS represented the *noise* that can be found on the web. Simply put, the SPIRIT sample is a random slice of the web whose content is unknown. Therefore, it contains not only genres that are different from those included in the model’s genre palette, but also genres that might be in the palette. Since we do not know the number and the distribution of genres on the web, this DON’T KNOW class is an attempt to bypass the constraint underlying ML-based models, where the documents must be necessarily pre-assigned to known and well-defined classes. In July 2008 I provided the SPIRIT sample with some genre labels. It is important to stress that the SPIRIT genre annotation is to be considered a starting point, not a validated annotation. In the future, annotation by other people (maybe through a social network) can be added and validated through agreement coefficients.

In my manual annotation of the SPIRIT sample, I used also “judgements”, namely *overlabelling* and *zerolabelling*. Overlabelling is counted as a NM (No Match) and indicates that the model’s genre palette did not contain the genre I had in mind. For example, SPRT\_002\_060\_117\_0058030 is an ERROR MESSAGE, but this genre is not present in the palette. In these cases, I used the label with NOTHING SUITABLE IN THE GENRE PALETTE. My expectation is that the model assigns no genre, i.e. I expect zerolabelling (as in the case of SPRT\_010\_049\_112\_0055944 and SPRT\_022\_009\_162\_0080850). Some other time, I could assign one of 15 genres, but the others that came to my mind were not in the model’s palette. In these cases I assigned the label NOTHING ELSE SUITABLE IN THE GENRE PALETTE with the expectation that the labels belonging to the palette could be matched. The SPIRIT sample contains also web pages for which I could not find a genre. In this case I used the label IDK (i.e. I DON’T KNOW). I considered the IDK pages as NC (Not Classified) because I could not say whether or to what extent the model was correct in its classification. The stand-off annotation of the SPIRIT sample is available online.<sup>12</sup>

Although not validated, this annotation gave me an idea of the composition of the SPIRIT sample. I assigned zero, one or more of the 15 genre classes of the model palette, up to four genre labels. The limit of four genre labels was the spontaneous

---

<sup>12</sup> The spreadsheet containing my standoff annotation is available at <<http://sites.google.com/site/marinasantiniacademic/site/>>: see *my\_manual\_genre\_labelling\_1000SPIRIT\_webpages\_NOVEMBER2008\_matching\_with\_the\_initial\_corpus.xls*.

boundary that I found to be comfortable when manually annotating web pages. This limit can be discussed and compared with other experiences in the future.<sup>13</sup>

### 5.2.2.2 KI-04

*Size.* The KI-04 includes 1,295 English web pages (HTML documents), but only 800 web pages (100 per genre) were used in the experiment described in Meyer zu Eissen and Stein [22]. KI-04 is a *de facto* standard in AGI, together with the 7-webgenre collection. Language: English. See Table 5.16 in the Appendix.

*Collection method and annotation.* The KI-04 corpus was collected using bookmarks from about five people. Some genres were extended to get a better balance. The corpus was sorted by three people, one of whom wrote a bachelor thesis (in German) on the corpus building process. One of the creators (S. Meyer zu Eissen) checked many of the pages, and most of the sorting complied with his understanding of the genre categories. The download date was January, 2004.

*Main purpose, assumptions or hypotheses.* The KI-04 corpus was built following a palette of eight genres suggested by a user study on genre usefulness [22]. Hence, the main purpose of this collection is to represent genres that are useful in retrieval tasks.

*Noise.* KI-04 does not contain any class representing noise.

### 5.2.2.3 HGC

*Size.* The Hierarchical Genre Collection (henceforth HGC) contains 32 genre classes, 40 files per class. Language: English. Collected in 2005/2006. 1,180 HGC web pages were used in the cross-testing experiments illustrated in this chapter. HGC is described in Stubbe and Ringstetter [36] and Stubbe et al. [37]. No detailed description of the genre classes is provided. See Table 5.17 in the Appendix.

*Collection method and annotation.* This collection was manually selected and annotated by Andrea Stubbe. She tried to gather a broad distribution of topics for each genre in order to avoid bias.

*Main purpose, assumptions or hypotheses.* HGC relies on the assumption that genre should exclusively represent the dimensions of the form and function of a text. The classification ought to be task oriented and hierarchical. It has to be logically consistent and complete. A certain text can be assigned to different classes, but

---

<sup>13</sup> It would be interesting to define the amount of the critical mass for genre annotation, i.e. to establish the point when the majority agrees on a number of labels for the same document. It seems that genre annotation based on the agreement of small number of people (2, 3, 4, or a few more) does not guarantee reliability. For instance Mikael Gunnarsson, made the following observations on the ARTICLE genres included in the KI-04 corpus, which is defined as “Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters” [22]. In this class, Gunnarsson found: a book announcement, a redirect page, a table of contents, bibliography, three documents authored in German, 2 commercial portrayals, 2 help pages, 2 discussion pages, 1 link list, and 1 personal homepage among the 127 articles (personal communication). Although intra-genre variation is, in my opinion, a positive characteristic, as well as a certain degree of noise, after Gunnarsson’s breakdown one might wonder about the criteria for representing a genre class.

this should not be the norm. HGC is based on the genre palette proposed by Dewe et al. [7], but it contains a finer grained hierarchy of genres, which, presumably, meet the demands of genre focused corpus construction.

*Noise.* HGC includes the NOTHING class, containing error messages, empty pages, and frame sets.

#### 5.2.2.4 MGC

*Size.* The Multi-Labelled Genre Collection (henceforth MGC) was built by Mitja Luštrek and Andrej Bratko and consists of 1,539 web pages classified into 20 genres. Each web page can belong to multiple genres. This collection is described in Vidulin et al. [41]. Language: English. See Table 5.18 in the Appendix.

*Collection method and annotation.* The corpus was manually labelled with genres by two independent annotators. Their labels disagreed on about a third of the web pages in the corpus, so these were reassessed by a third and sometimes even a fourth annotator. The web pages were collected from the Internet using three methods, i.e highly-ranked Google hits for popular keywords; gathered random web pages; finally, searching for web pages belonging to the genres underrepresented to that point to obtain a more balanced corpus.

*Main purpose, assumptions or hypotheses.* Genre categories were chosen with the intention to cover the whole Internet. The genre of a web page is intended to represent the communicational intention that shapes the page.

*Noise.* MGC does not contain any class representing noise.

### 5.3 The Web as Communication

The genre model presented here emphasizes the linguistic and pragmatic aspects of the web. Web pages are instantiations of communicative situations where language is used to interact in and with a context. The model relies and builds upon Biber's observation [2, p. 33] saying that linguistic features can be used to derive the communicative situation in which texts have been produced, thus identifying their communicative purposes. The genre palette and feature set (both described below) rely on this assumption.

#### 5.3.1 Genre Palette

Choices for the genre palette (i.e. the genres that the model can automatically identify) depend on the specific type of genre-enabled application. Presumably, a genre palette for a digital library will be different from a palette for web retrieval, for intranet searches, or for applications for corpus linguistics. Arguably, we cannot create a genre palette containing all the genres in use, since they amount to thousands (cf. [13]), which would also be very confusing for end users. Therefore, choices must be made. It is important to note that the palette used in this experiment is not

an “ideal” or an “all-purpose” palette. As emphasized above, each field of application will work out the best palette and the best nomenclature/taxonomy for its specific needs. The aim of the genre palette employed here is to investigate genres at different granularity or level of specificity, namely four rhetorical genres, four standardized genres coming from one domain (the BBC domain) and seven common web genres (see Table 5.1). This genre palette is static and non adaptive. This means that no suggestions are proposed to automatically incorporate new genre labels in the initial set.

In order to simplify the terminology, I will refer to these categories in the following way: “rhetorical genres”<sup>14</sup> indicate four rhetorical patterns, while “web genres” indicate both the BBC genres and the novel web genres. The idea is that all of them are genres, though with different characteristics. This view is very similar to the one proposed by Bruce ([5]; see also Chapter 15 by Bruce, this book), where social genres (my web genres) are built upon cognitive genres (my rhetorical genres). These two genre levels are tightly interrelated but they highlight different aspects in textual communication. While rhetorical/cognitive genres represent universal communicative purposes, social/web genres are historical entities with a life cycle. On the one hand, rhetorical/cognitive genres help harness the instability of the web or other noisy digital environments, because they are more stable than social genres. On the other hand, social/web genres, that come and go, and are very linked to technology, can be analysed and identified in terms of the communicative purpose they convey. It can also be said that rhetorical genres are more general and social genres more specific. It is worth highlighting that the computation of rhetorical genres as an

**Table 5.1** Genre palette

---

<b>Rhetorical Genres (A.K.A. Cognitive genres or text types)</b>
(1) Descriptive_narrative
(2) Explicatory_informational
(3) Argumentative_persuasive
(4) Instructional
<b>Traditional BBC web genres</b>
(5) BBC DIYs
(6) BBC editorials
(7) BBC short biographies
(8) BBC feature articles
<b>Novel web genres</b>
(9) Blogs
(10) Eshops
(11) FAQs
(12) Online newspaper front pages
(13) Listings
(14) Personal home pages
(15) Search pages

---

<sup>14</sup> Following Biber’s tradition [2], I had named them “text types” in my previous publications.

intermediate step is useful if we see genres as conventionalised and standardized cultural objects raising expectations about the purposes of communication. For example, what we expect from an BBC EDITORIAL is an “opinion” or a “comment” by the editor, which represents, broadly speaking, the view of the newspaper or magazine. Opinions are a form of ARGUMENTATION. ARGUMENTATION is a rhetorical genre expressed by a combination of linguistic features (the facets). If a document shows a high probability of being argumentative, i.e. if it has a high gradation of ARGUMENTATION, this document has a good chance of belonging to argumentative genres, such as EDITORIALS, SERMONS, PLEADINGS, and ACADEMIC PAPERS. It has less chances of being a STORY or a BIOGRAPHY, which are narrative genres.

### 5.3.2 Linguistically- and Functionally-Motivated Features

The genre model relies on features that I call *facets*. Broadly speaking, the word “facet” indicates an “aspect” of a situation, a concept, and so on. I used the word “facet” because each facet represents an “aspect” of communication. My facets are macro-features, i.e. they contain several micro-features. The advantage of these features is that they allow inference. While, shallow features are often unmeaningful for human understanding (e.g. character or byte n-grams), facets are higher order features can be easily understood and employed for reasoning.

For example, the first person facet includes first person pronouns, singular and plural. The first person facet indicates that the communication context is related to the text producer. A high frequency of first person facets in a text signals an impressionistic or subjective stance of the text producer. While in previous genre classification approaches, pronouns were used individually without any further interpretation, with the first person facet my aim is to interpret, or assess, whether first person pronouns indicate a particular stance in communication, and if this stance is linked to a genre. For instance, a high frequency of first person facet is often used in ARGUMENTATIVE genres, like COMMENTS and OPINIONS that can be found in newspapers and magazines.

I created 100 facets (listed in Table 5.19, in the Appendix). Facets can be refined and their number increased if they prove to be useful for AGI. The motivation, creation extraction, and drawbacks of facets are fully described in Santini [25, 29].

## 5.4 The Genre Model

The simple genre model that I describe and cross-test in this chapter has been built to fill a specific gap, namely the computability of the relation between genres and rhetorical patterns. In this respect, it complements more habitual approaches to AGI.

The model challenges two commonplaces in AGI, namely the predilection for shallow features and the use of fully-supervised ML approaches.

The use of shallow features is well justified because they are supposed to be potentially crosslingual and computationally inexpensive (cf. Chapter 7 by Sharoff; Chapter 8 by Stein et al., this book). Shallow features that have been used recently include n-characters n-grams [17], byte n-grams [21], harmonic descriptors (Chapter 6 by Kim and Ross, this book), and POS trigrams (Chapter 7 by Sharoff, this book).

These features with standard or adapted classifiers have high performance in small genre collections (see Table 5.5). However, the actual potential of these features in larger collections, for multi-lingual genre classification or in a more realistic environment is virtually unknown, although Chapter 7 by Sharoff (this book) and WEGA (Chapter 8 by Stein et al., this book; Stein and Meyer zu Eissen [35]) show some preliminary attempts towards multi-linguality. Additionally, the performance of the two genre-enabled existing IR applications relying on shallow features – namely, WEGA (Chapter 8 by Stein et al., this book and X-Site [11]) – still requires substantial enhancements.

Another common stance in AGI is the preference for fully supervised ML algorithms. Unfortunately, there are several disturbing factors that hinder the plain application of fully supervised ML to AGI. As ML models are build from examples, the penury of genre annotated material and the approximation of the genre population are major stumbling blocks. Current genre collections are tiny, ranging from the 200 web pages used in Chapter 7 by Sharoff (this book) to the 1,539 web pages used in Vidulin et al. [41], to the 3,685 pdf files employed in Chapter 6 by Kim and Ross (this book). Above all, these collections are very subjective, since they have all been built for individual or specific needs, collected with differing selection methods, following different conception of genres, including different genre palettes. Additionally, the performance of fully supervised methods relies heavily on the ideal combination of the following elements: a predictable population, an ideal genre palette, and large quantities of manually annotated stereotypical web pages. Consequently, it is difficult to draw any significant conclusions about the effectiveness of AGI since current findings are based on the tiny sizes of existing genre collections. Although there are ideas on how to create larger corpora of consensual genre-annotated material,<sup>15</sup> these ideas have not been implemented yet. In sum, being AGI in such a preliminary stage of research, there is no reason for not exploring other approaches that are alternative and complementary to shallow features,<sup>16</sup> manual annotation and ML.

The genre model, described and cross-tested in the following sections, explores the possibility of encoding genre knowledge in the classification model, rather than deriving it from stereotypical examples. This model has no ambition of becoming an industrial or commercial prototype, at least not in the implementation presented here. This is simply an explorative model that investigates the power of language

---

<sup>15</sup> For example, Rosso suggested that genre tags could be added (with a special genre-enabled tool) within social networks (personal communication).

<sup>16</sup> Cf. also the interesting experiments with “heavy” visual features carried out by Levering et al. [20] in order to detect subgenres.

in detecting genre classes and the relationship between rhetorical patterns and web genres.

The model relies upon linguistically rich features and layout information, and follows a multilabel scheme. It has been devised thinking of an open digital environment, like the web, where the level of noise is high and the population is difficult to approximate. Its task is to apply either no genre – when a document is highly individualized – or one genre – when the document belongs to a single genre – or multiple genres – when a document contains several genres or is hybrid. The unit of analysis is an English individual web page, including boilerplates and navigational text.

Empirical observations have led me to include the attributes of genre hybridism and individualisation in the characterization of the genre of web pages [28]. These two attributes account for classification hurdles, and help pinpoint the range of flexibility that an automatic genre classification system should have. I suggested that genre hybridism accounts for multi-genre classification, whereas individualisation accounts for zero-genre classification. Such a broad range of flexibility is not permitted by standard discrete single-label supervised classification algorithms. The efficacy of ML multi-label classifier, like LIBSVM, has been little explored in AGI (one experience is described in Vidulin et al. [42]).

The model goes beyond the single-label assignment and does not require any annotation of web pages by genre. This model has two main characteristics: (i) it makes a clear-cut distinction between rhetorical genres and web genres, and (ii) it is based on inference rather than supervised learning.

The first original trait relies on the separation of the concepts of rhetorical genres and web genres, where rhetorical genres represent a middle layer between functionally interpreted features – the *facets* (Section 5.2) – and web genres. This intermediate level gives flexibility to an automatic genre classification system because rhetorical genres are linguistic devices that represent the purpose of communication, and are more universal than genres since they span across all cultures and all times. Web genres, on the other hand, are (like all other social genres, see Bruce [5]) cultural artefacts, linked to a historical context, and in constant evolution. By using rhetorical genres, an analysis will remain possible on all media (printing, the web, mobile phones, etc.) even if genres evolve and texts cannot be safely ascribed to any existing genre (zero-genre assignment), or if texts show several genres at the same time (multi-genre assignment), since rhetorical genres help relate genres to one another across old and new media.

The second original trait relies on inference rather than supervised ML. More precisely, rhetorical genres are inferred using a modified form of Bayes' theorem – the odds-likelihood or subjective Bayesian method – and web genres are derived using a few inferential *if-then* rules. For this reason, I refer to this model as the *inferential model*. Inference is possible thanks to high-level, linguistically rich and functionally motivated features (Section 5.2). While shallow features are opaque to human understanding, deeper linguistic features allow functional interpretation of texts. Biber [2] showed that functionally interpretable, linguistic features correlated through factor analysis return textual dimensions (Biber's text types) that can tell us

something about the communicative situation in which a text has been produced.<sup>17</sup> The ultimate goal is to investigate to what extent complex linguistic features can allow genre classification with noisy and heterogeneous corpora, more similar to a real-world scenario. For this reason, the inferential model will be cross-tested with a number of different genre collections.

### 5.4.1 Methodology

The model implements a simplified form of Bayes' theorem called odds-likelihood or subjective Bayesian method, suggested by Duda and co-workers [9, 10] to handle uncertainty in PROSPECTOR, a rule-based system for classifying mineral exploration prospects. The main reason for choosing the odds-likelihood form of Bayes' theorem is that the model is very simple, but allows more complex reasoning through the use of weights. Like the standard Bayesian version, the odds-likelihood method is based on probabilities. In the flow, probabilities are converted into odds. Odds and probabilities contain exactly the same information and are interconvertible. But odds are not limited to the range 0–1, like probabilities. In other words, odds is a positive integer (without any limitation) that tells us how much more likely one hypothesis is than the other. Odds are usually used for games of chance, where the probabilities are expressed in the form of integer-to-integer (e.g. “six-to-one”) where the first figure represents the number of ways of failing to achieve the outcome and the second figure is the number of ways of achieving a favourable outcome. The main difference between the regular Bayes models and the subjective one is that in the latter attributes are NOT considered to be equally important, but are, instead, weighted according to their probability value. Therefore, in the odds-likelihood version of Bayes' theorem much of the effort is devoted to weighing the contributions of different pieces of evidence in establishing the match with a hypothesis. These weights are confidence measures: Logical Sufficiency (LS) and Logical Necessity (LN). LS is used when the evidence is known to exist (larger value means greater sufficiency), while LN is used when evidence is known NOT to exist (a smaller value means greater necessity). One important point to make is that the facets, i.e. the pieces of evidence on which the model relies upon, are not just either present or absent: their presence or absence is considered to be uncertain. Therefore, LS and LN can be viewed as the limits of the interval in which lies the value indicating the degree to which a facet influences the prior probability of H (the hypothesis). In this implementation of the model, LS was set to 1.25 and LN was set to 0.8 on the basis of previous experience and empirical adjustments. While in this implementation of the model, LS and LN have a single value for all the facets (i.e. LS is always 1.25 and LN always 0.8), it is also possible to compute a

---

<sup>17</sup> “The notion of function is closely associated with the notion of situation. A primary motivation for analysis of the components of situation is the desire to link the functions of particular linguistic features to variation in the communicative situation” [2, p. 33].

weight for each facet, because some facets can be more indicative than others. But in order to do so, it would be necessary to have a corpus of documents already classified by rhetorical genres, which is not a trivial endeavour, as pointed out earlier.

### 5.4.2 *Flow and Hypotheses*

The inferential model is based on the following steps:

1. Extraction, count and normalization of linguistic facets.
2. Conversion of normalized counts into z-scores, which represent the deviation from the “norm” coming out from the web corpus.
3. Conversion of z-scores into probabilities, which means that facet frequencies are seen in terms of probabilities distribution.
4. Calculation of prior odds from prior probabilities of a rhetorical genre. The prior probability for each of the four rhetorical genres was set to 0.25 (all rhetorical genres were given an equal chance to appear in a web page). Prior odds are calculated with the formula:

$$\text{prior\_Odds}(H) = \text{prior\_Prob}(H) / 1 - \text{prior\_Prob}(H)$$

5. Calculation of weighted facets, or multipliers (M). If a facet, i.e. a piece of evidence (E), has a probability  $\geq 0.5$ , LS is applied, otherwise LN is applied. Multipliers are calculated with the following formulae:

$$\text{if Prob}(E) \geq 0.5 \text{ then} \\ M(E) = 1 + (LS - 1) (\text{Prob}(E) - 0.5) / 0.25$$

$$\text{if Prob}(E) < 0.5 \text{ then} \\ M(E) = 1 - (1 - LN) (0.5 - \text{Prob}(E)) / 0.25$$

6. Multiplication of weighted probabilities together, according to the co-occurrence decided by the analyst on the basis of previous studies.
7. Posterior odds for the rhetorical genre is then calculated by multiplying prior odds (Step 5) with co-occurrence of weighted facets (Step 7).
8. Finally, posterior odds is re-converted into a probability value with the following formula:

$$\text{Prob}(H) = \text{Odds}(H) / 1 + \text{Odds}(H)$$

At the end of this flow, the model returns the inferred rhetorical genre associated to a score. Scores are interpreted in terms of degree or gradation. For example, a web page with a score of 0.9 of being argumentative shows a very high degree, or gradation, of ARGUMENTATION. As explained later, the different gradations are independent from each other. In other words, the different scores accounting for the four rhetorical genres in a web page do not sum up to 1.0, but they simply indicate the gradation, and not the proportion, of a certain rhetorical genre. Scores

are then ranked in descending order (the highest score gets the first position). After the ranking, two hypotheses are tested:

1. The first hypothesis says that the combination of a number of rhetorical genres is sufficient to derive four BBC web genres – EDITORIALS, DIY MINI-GUIDES, SHORT BIOGRAPHIES, and FEATURE ARTICLES –, more traditional in their textuality. This hypothesis is tested with rules that combine only rhetorical genres, without any additional features. An example of these rules to derive BBC genres is shown in Box 5.1. These rules are used to assess the classification accuracy of four BBC web genres immersed in increasingly larger corpora (see Table 5.3).
2. The second hypothesis says that the combination of two predominant rhetorical genre, i.e. the top-ranked rhetorical genres, plus a combination of additional traits is sufficient to derive seven web genres – BLOGS, ESHOPS, FAQs, FRONT PAGES, LISTINGS, PERSONAL HOME PAGES, and SEARCH PAGES –, more influenced by the functionalities allowed by the web. This hypothesis is tested with rules take combines rhetorical genres plus additional features. An example of these rules to derive webgenres is shown in Boxes 5.2 and 5.3. These rules are used to assess the classification accuracy of seven novel web genres immersed in increasingly larger corpora (see Table 5.4).

#### Box 5.1 Rules for BBC DIY Mini-Guides

```
if (text_type_1=/instructional_1|argumentative_persuasive_1/)
if (text_type_2=/argumentative_persuasive_2|instructional_2/)
if (text_type_3/expository_informational_3|descriptive_narrative_4/)
```

#### Box 5.2 Positive Rules for Blogs

```
if (text_type_1=descr_narrat_1|argum_pers_1)
    then add 1 to goodBlogCandidate
if (text_type_2=descr_narrat_2|argum_pers_2)
    then add 1 to goodBlogCandidate
if (page_length=LONG)
    then add 1 to goodBlogCandidate
if (blog_words >= 0.5 probabilities)
    then add 1 to goodBlogCandidate
if goodBlogCandidate >=3
    then good BLOG candidate.
```

#### Box 5.3 Negative Rules for Blogs

```
if (frontpage_words > blog_words)
    then subtract 1 to goodBlogCandidate
```

There is no special reason for combining only two predominant rhetorical genres instead of three or more. The basic assumption is that web pages are mixed. With BBC web genres, I stress the intra-genre linguistic mixture, while with seven novel web genres I include additional traits, like layout and functionality, in the genre profiling. Obviously, web pages may contain many other rhetorical genres, not only the ones included in this implementation of the model. This palette is a starting point and can be enlarged and adjusted in future.

Simple *if-then* rules combine inferred rhetorical genres with additional traits for determining web genres in web pages. The main reason for not applying odds-likelihood here is that the combination of rhetorical genres with other features is more tentative; inference rules allow us to better understand how a conclusion is reached. Naturally, any future enhancement of this model would include the development and debugging of more sophisticated rules and the setting of thresholds.

The number of positive rules, i.e. those that confirm the presence of positive attributes for the genre under assessment, is very limited:

- 3 rules for each of the BBC genres;
- 4 rules for BLOGS;
- 7 rules for ESHOPS;
- 5 rules for FAQs;
- 8 rules for FRONT PAGES;
- 5 rules for LISTINGS;
- 4 rules for PHP;
- 9 rules for SEARCH PAGE.

The number of negative rules, i.e. those that disconfirm the presence of positive attributes for the genre under assessment is very low for BLOGS, and slightly higher for other web genres.

## 5.5 Results

As emphasized earlier, the main purpose of this chapter is to explore how to assess the robustness of genre models in noisy scenarios representing the web. These scenarios are represented by a number of genre collections, which will be used in combination with each other and in isolation. In this section, results are described. First, the performance on the single labels of the four BBC genres and the seven novel genres is shown (Sections 5.5.1, 5.5.2 and 5.5.3). Then the performance achieved on the 7-webgenre collection is compared with the results obtained by other AGI models and differing feature sets (Section 5.5.4). Finally, an attempt to assess multi-labelling is illustrated (Section 5.5.5).

For single label experiments, the evaluation measure employed to compare results is accuracy, i.e. the percentage of correct guesses on the BBC and the 7-webgenre collections.

file_name	descripti	expositor	argume	instructional																
BBC_DIY_guide_0C	0.17	0.22	0.33	0.75	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.16	0.25	0.25	0.37	GOOD	DIY	i	BAD	editori	BAD	bio	GOOD	fea	BAD	blog	i	GOOD	est	BAD	fe
BBC_DIY_guide_0C	0.16	0.25	0.25	0.37	GOOD	DIY	i	BAD	editori	BAD	bio	GOOD	fea	BAD	blog	i	GOOD	est	BAD	fe
BBC_DIY_guide_0C	0.27	0.19	0.19	0.36	GOOD	DIY	i	BAD	editori	GOOD	b	GOOD	fea	BAD	blog	i	BAD	esho	GOOD	fe
BBC_DIY_guide_0C	0.22	0.23	0.33	0.62	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.17	0.15	0.21	0.58	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.09	0.15	0.25	0.24	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.24	0.27	0.42	0.38	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.22	0.29	0.23	0.52	GOOD	DIY	i	BAD	editori	BAD	bio	GOOD	fea	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.17	0.29	0.30	0.65	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.27	0.14	0.32	0.70	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	GOOD	fe
BBC_DIY_guide_0C	0.11	0.13	0.23	0.34	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.14	0.19	0.29	0.35	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.13	0.15	0.29	0.32	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	GOOD	fe
BBC_DIY_guide_0C	0.14	0.19	0.19	0.40	GOOD	DIY	i	BAD	editori	BAD	bio	GOOD	fea	BAD	blog	i	BAD	esho	GOOD	fe
BBC_DIY_guide_0C	0.10	0.17	0.18	0.27	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	GOOD	fe
BBC_DIY_guide_0C	0.10	0.11	0.11	0.37	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.32	0.18	0.22	0.57	BAD	DIY	ca	BAD	editori	BAD	bio	GOOD	fea	BAD	blog	i	BAD	esho	BAD	fe
BBC_DIY_guide_0C	0.15	0.13	0.21	0.40	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	GOOD	fe
BBC_DIY_guide_0C	0.19	0.21	0.25	0.56	GOOD	DIY	i	BAD	editori	BAD	bio	BAD	featu	BAD	blog	i	BAD	esho	BAD	fe

Fig. 5.1 Excerpt from the output of the inferential model

For the multilabelling, matching accuracy and overlap coefficients are computed on the SPIRIT sample.

It is worth noting that the model has not been adjusted, adapted or optimised when applied to the different genre collections. Classification settings and feature set have been kept constant. An excerpt of the classification output is shown in Fig. 5.1. All the spreadsheets containing the inferential model’s classification are available online.<sup>18</sup>

### 5.5.1 Cross-Testing Performance on Single Labels: BBC and 7-Webgenre Collections

In this subsection, the model *scalability* in noisy environments is tried out by progressively increasing the size of the web corpus in three steps. In the first step, the model is applied to 2,480 web pages; in the second step on 3,685 web pages; in the final step on 6,404 web pages. The classification performance is compared on the single labels of the BBC and 7-webgenre collections.

#### 5.5.1.1 SANTINIS (2,480 Web Pages)

The inferential model achieves an accuracy of about 86% on SANTINIS (see Table 5.2, forth column) on the single labels of the 7-webgenre collection. The

<sup>18</sup> See all the excel files whose names start with “GIMS” at <http://sites.google.com/site/marinasantiniacademic/site/>.

accuracy of 86% returned by the model is a good achievement for a first implementation, especially if we consider that the standard Naïve Bayes classifier returns an accuracy of about 67% (see Table 5.2, third column). Although SVM achieves an accuracy of about 89% (see Table 5.2, second column), i.e. about +3% more than my genre model, it is worth stressing that both Naive Bayes and SVM standard classifiers were run on 1,400 web pages, i.e. they were built only on the 7-web-collection (Section 5.3). The inferential model, on the other hand, is run on the noisy SANTINS, made of 2,480 web pages (Section 5.3). This means that the model is robust to a certain level of noise, represented the SPIRIT sample and the BBC collection.

Although theoretically unsound, for explanatory purposes I built an UNKNOWN class. This means that I built an SVM classifier with SANTINIS 2,480 web pages and 100 facets. Web pages belonging to the BBC collection and the SPIRIT sample, i.e. 1,080 web pages, were labelled as DONTKNOW.

The stratified 10-fold-cross-validated accuracy (seed 1) on the SVM model built with eight classes is about 76%, i.e. about -13% of the accuracy achieved with a corpus of 1,400 web pages. In this respect, the inferential model is much more *corpus independent*, returning an accuracy of about 86% on 2,480 web pages, i.e. about +10% more than the model built with SVM on eight classes. The DONTKNOW class (Fig. 5.2, column “h”) causes many misclassifications. This confirms the observation made by Shepherd et al. [34] that the introduction of noise – in this case, unclassified web pages – significantly decreases the accuracy of the a fully supervised classifier.

In order to test its robustness to scalability, an increase of corpus size was simulated by adding first the KI-04, and finally both HGC and MGC.

**Table 5.2** Comparing accuracies: SVM, Naïve Bayes and the inferential genre model

Web genres	SVM (1,400 web pages (%))	Naïve Bayes (1,400 web pages (%))	Inferential model (2,480 web pages)
Blogs	96	92	91% (18 bad blog; 182 good blog)
Eshops	88	76	83% (4 bad eshop; 166 good eshop)
FAQs	94.5	67	88.5% (23 bad FAQ; 177 good FAQ)
Front pages	100	98	97% (6 bad frontpage; 194 good frontpage)
Listings	80	29	75.5% (49 bad listing; 151 good listing)
Pers. home pages	79	27	77% 46 (bad PHP; 154 good PHP)
Search pages	85	82	88% (24 bad spage; 176 good spage)
Total	About 89%	About 67%	About 86%

Correctly Classified Instances	1885	76.0081 %							
Incorrectly Classified Instances	595	23.9919 %							
Total Number of Instances	2480								
=== Confusion Matrix ===									
	a	b	c	d	e	f	g	h	<-- classified as
161	0	0	1	1	6	0	31		a = BLOG
0	123	3	0	6	2	7	59		b = ESHOP
0	1	152	0	1	1	0	45		c = FAQs
0	0	0	186	0	0	0	14		d = FRONTPAGE
3	5	2	1	78	1	8	102		e = LISTING
8	2	1	0	1	130	0	58		f = PHP
0	6	0	2	8	0	153	31		g = SPAGE
25	25	33	16	33	28	18	902		h = DONTKNOW

Fig. 5.2 SVM confusion matrix: misclassifications caused by the DONTKNOW class

**5.5.1.2 SANTINIS+KI-04 (3,685 Web Pages) & SANTINIS+KI-04+HGC+MGC (6,400 Web Pages)**

The accuracies achieved with the inferential models on BBC web genres and on the 7 novel genres are shown in Tables 5.3 and 5.4.

Differences in accuracy are statistically (chi-square<sup>19</sup>) not significant for the BBC web genres (Table 5.3), but significant for the 7 novel web genres (Table 5.4).

On the first enlarged corpus (i.e. 3,685 web pages), the accuracy achieved on a single genre is about 81%. These results are encouraging, since a size increase of 35% causes only 5% decrease in accuracy. The same is true with the second enlarged corpus (6,404 web pages), where the accuracy achieved on a single genre is about 76%. This means that a size increase of 60% causes only 10% decrease in accuracy. It is interesting to note that *the size increase is noisy*, i.e. the corpus has been enlarged with additional genre collections containing diverse genres, not just incrementing the size of existing, well-defined genre classes. As mentioned above,

**Table 5.3** Accuracies on the four BBC web genres

BBC Web Genres	Accuracies on SANTINIS (2,480 web pages)	Accuracies SANTINI + KI-04 (3,685 web pages)	Accuracies on SANTINIS + KI-04+HGC+MGC (6,404 web pages)
BBC DIY	95% (1 bad DIY; 19 good DIY)	85% (3 bad DIY; 17 good DIY)	85% (3 bad DIY; 17 good DIY)
BBC editorial	75% (5 bad editorial; 15 good editorial)	75% (5 bad editorial; 15 good editorial)	70% (6 bad editorial; 14 good editorial)
BBC bio	85% (3 bad bio; 17 good bio)	75% (5 bad bio; 15 good bio)	85% (3 bad bio; 17 good bio)
BBC features	60% (8 bad feature; 12 good feature)	50% 10 bad feature; 10 good feature	60% (8 bad feature; 12 good feature)
Total	About 79%	About 71%	About 75%

<sup>19</sup> Chi-square calculator: <[http://www.physics.csbsju.edu/cgi-bin/stats/contingency\\_form.sh?nrow=2&ncolumn=2](http://www.physics.csbsju.edu/cgi-bin/stats/contingency_form.sh?nrow=2&ncolumn=2)>. (April 2009)

**Table 5.4** Accuracies on the seven novel web genres

7 novel web genres	Accuracies on SANTINIS (2,480 web pages)	Accuracies on SANTINI + KI-04 (3,685 web pages)	Accuracies on SANTINIS+KI-04+HGC+MGC (6,404 web pages)
Blogs	91% (18 bad blog; 182 good blog)	72% (56 bad blog; 144 good blog)	93% (14 bad blog; 186 good blog)
Eshops	83% (34 bad eshop; 166 good eshop)	78.5% (43 bad eshop; 157 good eshop)	66% (68 bad eshop; 132 good eshop)
FAQs	88.5% (23 bad FAQ; 177 good FAQ)	84% (32 bad FAQ; 168 good FAQ)	88.5% (23 bad FAQ; 177 good FAQ)
Front pages	97% (6 bad frontpage; 194 good frontpage)	96.5% (7 bad frontpage; 193 good frontpage)	61% (78 bad frontpage; 122 good frontpage)
Listings	75.5% (49 bad listing; 151 good listing)	74% (52 bad listing; 148 good listing)	75% (50 bad listing; 150 good listing)
Personal home pages	77% 46 bad PHP; 154 Good PHP	77.5% (45 bad PHP; 155 good PHP)	81% (38 bad PHP; 162 good PHP)
Search pages	88% (24 bad spage; 176 good spage)	85.5% (29 bad spage; 171 good spage)	69% (62 bad spage; 138 good spage)
Total	About 86%	About 81%	About 76%

such an environment presumably represents the unpredictable population of the web. Apparently, the model is robust enough to withstand, to some extent, the chaos of the web.

### 5.5.2 Performances of Other Single-Label Models on the 7-Webgenre Collection

It is interesting to compare the performances that other researches have obtained on the 7-webgenre collection, which has been extensively used in other genre experiments. As mentioned in Section 5.3, this collection (together with KI-04) has become a de facto standard in AGI research. Table 5.5 shows some results.

When the 7-webgenre collection is used in isolation (rows 1–7), classification results increase with the number of features. Accuracies are all very high: from 88.8% achieved with 100 features (row 3) to 96.5 with >3,000 features (row 6). Since long vectors always raise the suspicion of overfitting, Kanaris and Stamatatos [17] have cross-checked their features by classifying the 7-webgenre collection using features extracted from the KI-04 corpus with a very good accuracy, i.e.

Table 5.5 Performances on the 7-webgenre collection with and without noise

Row	Experiment	Features	Classifier	# of web pages	Genre collections	Accuracy on the 7-webgenre collection (%)
1	Santini [26]	118 (function words, POSs, punctuation, genre-specific words, HTML tags)	SVM	1, 400	7-webgenre collection	90.6
2	Santini [26]	140 POS trigrams	SVM	1, 400	7-webgenre collection	89.4
3	Santini [26]	100 facets (see Section 5.3.2)	SVM	1, 400	7-webgenre collection	88.8
4	Waltinger and Mehler [43]	Ranked profiles of their n-gram frequencies (# of features N/A)	Category profiling (9 sub models associated to one categ.)	1, 400	7-webgenre collection	93
5	Mason et al. [21]	Most frequent 1,000 7-g per web page	Genre comparison method (based on a centroid feature set for each genre)	1, 400	7-webgenre collection	94.6
6	Kanaris and Stamataos [16]	>3,000 character n-gram + structural information	SVM	1, 400	7-webgenre collection	96.5
7	Kanaris and Stamataos [17]	7,203 character n-gram and structural features	SVM	1, 400	7-webgenre collection	96.6
8	Chapter 6 by Kim and Ross (this volume)	7,431 harmonic descriptors	SVM	3, 452	7-webgenre collection + 24 KRYIS-01 genres (structured noise)	96

Table 5.5 (continued)

Row	Experiment	Features	Classifier	# of web pages	Genre collections	Accuracy on the 7-webgenre collection (%)
9	Santini et al. [33]	100 facets	Inferential model	2,480	7-webgenre collection + structured noise (BBC genres) and unstructured noise (SPIRIT sample)	86
10	Santini [29]	100 facets	Inferential model	3,685	7-webgenre collection + structured noise (BBC genres + KI-04) and unstructured noise (SPIRIT sample)	81
11	Santini (this chapter)	100 facets	Inferential model	6,404	7-webgenre collection + structured noise (BBC genres + KI-04 + HGC + MGC) and unstructured noise (SPIRIT sample)	76

95.2%. The best performance achieved by Kanaris and Stamatatos is 96.6% using 7,203 features (row 7). Waltinger and Mehler propose a graph-based method capable of classifying the 7-webgenre collection with competitive accuracy<sup>20</sup> (row 4). However, the experiments listed in row 1–7 are carried out on a very small collection (i.e. 1,400 web pages), balanced (200 web pages per genre), and without any noise. It is very unlikely that the web can be represented in this way. A step forward towards a more diversified representation of genres of web documents is done in Chapter 6 by Kim and Ross (this book). They reach an accuracy of 96% on the 7-webgenre collection while classifying it together with other 24 genres from the KRY5-01 pdf collection (the overall accuracy on 31 genres is about 70%). However, the high number of features (7,421) and the small size of the corpus (3,685 documents) are suspicious elements for assessing the model’s generality and corpus-independence. Additionally, all the 3,685 web documents are supposed to fall into well-defined genre classes. This means that only structured noise is represented in this experiment.

The inferential model uses only 100 features and shows an accuracy of 86% when the 7-webgenre collection is classified in a corpus of 2,480 web pages including structured and unstructured noise. Its accuracy decreases to 81% when the corpus is extended up to 3,685 web pages, and to 76% in a very noisy corpus of 6,404 web pages (see also Section 5.5.1). I propose these three accuracies as baselines for future experiments with structured and unstructured noise and an increasing corpus size.

### 5.5.3 Cross-Testing Performance on Single Labels: Mapped Web Genres

In this subsection, the inferential model’s *exportability* is tried out by mapping some of the web genres in the genre collections (i.e. KI-04, HGC and MGC) to my palette. The model performance on these mapped genres is shown in the tables below.

The performance on KI-04 mapped genres (Table 5.6) is stable when the corpus is increased, and in line with the accuracy achieved by Meyer zu Eissen and Stein [22]. Differences in accuracy are not significant.

Although Stubbe et al.’s [37] results are not directly comparable, given the different evaluation measures employed by the authors, Table 5.7 tells us that the performance on HGC mapped genres is not depreciable on 6404 web pages. The most penalized genre is certainly FEATURES.

Conversely, the performance on MGC is much lower than the results reported in Vidulin et al. [41]. Only PERSONAL BLOGS, which in MGC corresponds to web pages that have been labelled both as PERSONAL and as BLOG (double label), perform satisfactorily on 6404 web pages (Table 5.8).

---

<sup>20</sup> The same method can be used for language identification and subject-based text classification.

**Table 5.6** Cross-testing on KI-04 mapped genres

Three of KI-04's genres	Accuracies from Meyer zu Eissen and Stein [22] on 800 web pages with discriminant analysis	Accuracies on SANTINI + KI-04 (3,685 web pages)	Accuracies on SANTINIS+KI-04+HGC+MGC (6,404 web pages)
KI-04 linklists ( <i>mapped to my listings</i> )	67.6% (out of 100 web pages)	63.9% (out of 205 web pages)	62.4% (out of 205 web pages)
KI-04 portrayal-priv ( <i>mapped to my personal home page</i> )	67.7% (out of 100 web pages)	83.3% (out of 26 web pages)	82.5% (out of 126 web pages)
KI-04 shops ( <i>mapped to my eshop</i> )	66.9% (out of 100 web pages)	71.8% (out of 167 web pages)	66.4% (out of 167 web pages)
Total	67.4%	73%	70.4%

**Table 5.7** Cross-testing on HGC mapped genres

Three of HGC mapped genres	Stubbe et al. [37] on 1,280 web pages		Accuracies on SANTINIS+KI-04+HGC+MGC (6,404 web pages)
	Precision (%)	Recall (%)	
Features ( <i>mapped to my BBC feature articles</i> )	53.8	35	12.5% (5 correct guesses out of 40 HGC features)
Blogs ( <i>mapped to my personal blogs</i> )	92.9	65	76.2% (32 correct guesses out of 42 HGC blogs)
FAQs ( <i>mapped to my FAQs</i> )	86.7	65	63.4% (26 correct guesses out of 41 HGC FAQs)
Total	77.8	55	50.7%

**Table 5.8** Cross-testing on MGC mapped genres

Four MGC mapped genres	Vidulin et al. [41] on 1539 web pages			Accuracies on SANTINIS+KI-04+HGC+MGC (6,404 web pages)
	Accuracy (%)	Precision	Recall	
Personal blogs ( <i>mapped to my personal blogs</i> )	(All kinds of blogs) 96	71	56	79.4% (27 correct guesses out of 34 MGC personal blogs)
All kinds of shopping ( <i>mapped to my eshop</i> )	97	89	38	37.9% (25 correct guesses out of 66 MGC shopping pages)
All kinds of faqs ( <i>mapped to my FAQs</i> )	99	94	77	47.1% (33 correct guesses out of 70 FAQs)
All kinds of index pages ( <i>mapped to my listings</i> )	85	53	42	32.6% (74 correct guesses out of 227 index pages)
Total	94.25	76.75	71	49.25%

### 5.5.4 Cross-Testing Performance on Single Labels: HGC and MGC in Isolation

In this section, the exportability of the inferential model is tried out on HGC and MGC in isolation. Performance on the mapped genres can be compared when the model is applied to these collections in isolation and to the increased corpus.

#### 5.5.4.1 HGC

When the model is run on the HGC in isolation, the performance on the three mapped genres is much higher than when it is run on 6,404 web pages (Table 5.9). Interestingly, the performance on the FEATURE genre is perfect when the model is applied to HGC alone.

**Table 5.9** Accuracies on HGC

Three of HGC mapped genres	Accuracies on the mapped genres of HGC in isolation (1,180 web pages)	Accuracies on SANTINIS+ KI-04+HGC+MGC (6,404 web pages)
Features ( <i>mapped to my BBC feature articles</i> )	100% (40 correct guesses out of 40 HGC features)	12.5% (5 correct guesses out of 40 HGC features)
Blogs ( <i>mapped to my personal blogs</i> )	73.8% (31 correct guesses out of 42 HGC blogs)	76.2% (32 correct guesses out of 42 HGC blogs)
FAQs ( <i>mapped to my FAQs</i> )	87.8% (36 correct guesses out of 41 HGC FAQs)	63.4% (26 correct guesses out of 41 HGC FAQs)
Total	87.2%	50.7%

#### 5.5.4.2 MGC

The performance on MGC is quite idiosyncratic. When the model is run on MGC in isolation, the performance on the four mapped genres is slightly lower than when it is run on 6,404 web pages (differences on accuracy are significant). Apparently, the model performs better when MGC is immersed in a larger corpus (Table 5.10).

### 5.5.5 The SPIRIT Sample: An Attempt to Assess Multilabelling

#### 5.5.5.1 Matching Accuracy (SANTINIS)

In this section I will describe the manual matching that I carried out in order to compare and analyse the similarities and discrepancies between my genre annotation and the model classification of the SPIRIT sample on SANTINIS (i.e. 2,480 web pages, see Section 5.2.2). This comparison has been very time consuming and I performed it in order gain a deeper insight into the model's classification behaviour. For a first

**Table 5.10** Accuracies on MGC

	Accuracies on the mapped genres of MGC in isolation (1,539 web pages)	Accuracies on SANTINIS+ KI-04+HGC+ MGC (6,404 web pages)
Four MGC mapped genres		
Personal blogs ( <i>mapped to my personal blogs</i> )	79.4% (27 correct guesses out of 34 MGC personal blogs)	79.4% (27 correct guesses out of 34 MGC personal blogs)
All kinds of shopping ( <i>mapped to my eshop</i> )	18.2% (12 correct guesses out of 66 MGC shopping pages)	37.9% (25 correct guesses out of 66 MGC shopping pages)
All kinds of faqs ( <i>mapped to my FAQs</i> )	47.1% (33 correct guesses out of 70 FAQs)	47.1% (33 correct guesses out of 70 FAQs)
All kind of index pages ( <i>mapped to my listings</i> )	23.7% (54 correct guesses out of 227 index pages)	32.6% (74 correct guesses out of 227 index pages)
Total	42.1%	49.25%

assessment of this comparison, I used the approach similar to that utilized by Freund et al. [12]. Although the principle is the same, my coding is slightly different and follows the criteria listed in Table 5.11.

Figure 5.3 shows a excerpt of my manual annotation of the SPIRIT sample, while Fig. 5.4 illustrates the output of the inferential model on the SPIRIT sample. The first file (SPRT\_002\_060\_117\_0058000) was classified as a FRONTPAGE and a LISTING genre by me. The model correctly guessed FRONTPAGE, but not LISTING. According to the model, this page could also be a good candidate SHORT BIOGRAPHY and EDITORIAL genres. In this case, only a FM (Fair Match) is scored since there is only one match between my labels and the model's labels.<sup>21</sup> Matching results are shown in Table 5.12.

**Table 5.11** Matching criteria

PM = Perfect Match	All my genre labels match all the predicted genre labels
EM = Excellent Match	4 of my genre labels match the predicted genre labels (e.g. when the model assigns 5 or more genre labels to the same web page)
VG = Very Good Match	At least 3 of my genre labels match the predicted genre labels
G = Good Match	At least 1 of my genre labels match the predicted genre labels
FM = Fair Match	At least 1 of my genre labels match the predicted genre labels
NM = No Match	No matches between my annotation and the model classification
NC = Not Classified	Either the genres are not in the model's palette, or i do not know how to classify the page

<sup>21</sup> The spreadsheet containing the matches is available at <<http://sites.google.com/site/marinasantiniacademic/site/>>: see *my\_manual\_genre\_labelling\_1000SPIRIT\_webpages.xls*.

SPRIT WEB PAGES	my_genre_label_1	my_genre_label_2	my_genre_label_3	my_genre_label_4
SPRT_002_060_117_0058000	frontpage	listing		
SPRT_002_060_117_0058001	listing			
SPRT_002_060_117_0058002	instructional	expository_informational	argumentative_persuasive	listing
SPRT_002_060_117_0058003	argumentative_persuasive	descriptive_narrative		
SPRT_002_060_117_0058004	listing			
SPRT_002_060_117_0058005	listing	descriptive_narrative		
SPRT_002_060_117_0058007	spage	listing		
SPRT_002_060_117_0058008	descriptive_narrative	expository_informational		
SPRT_002_060_117_0058009	listing			
SPRT_002_060_117_0058010	descriptive_narrative	listing	php	
SPRT_002_060_117_0058011	listing			
SPRT_002_060_117_0058012	expository_informational	listing		
SPRT_002_060_117_0058015	descriptive_narrative	listing	argumentative_persuasive	
SPRT_002_060_117_0058016	listing	expository_informational		
SPRT_002_060_117_0058017	argumentative_persuasive	expository_informational		
SPRT_002_060_117_0058019	spage	listing	descriptive_narrative	argumentative_persuasive
SPRT_002_060_117_0058020	listing	expository_informational		
SPRT_002_060_117_0058023	listing	spage		
SPRT_002_060_117_0058024	expository_informational	listing		
SPRT_002_060_117_0058027	nothing suitable in the genre palette			
SPRT_002_060_117_0058028	listing	nothing else suitable in the genre palette		
SPRT_002_060_117_0058030	nothing suitable in the genre palette			

Fig. 5.3 SPRIT manual annotation

SPRT_002_060_117_0058000	0.20	0.11	0.10	0.11	BAD DIY c GOOD edi GOOD bio BAD featu BAD blog i BAD est
SPRT_002_060_117_0058001	0.06	0.11	0.05	0.04	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i BAD est
SPRT_002_060_117_0058002	0.41	0.51	0.59	0.72	GOOD DIy BAD edito. BAD bio ci. BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058003	0.08	0.16	0.10	0.06	BAD DIY c GOOD edi GOOD bio BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058004	0.12	0.22	0.24	0.15	GOOD DIy GOOD edi BAD bio ci. BAD featu BAD blog i. GOOD e
SPRT_002_060_117_0058005	0.09	0.10	0.07	0.08	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. GOOD e
SPRT_002_060_117_0058007	0.09	0.15	0.08	0.08	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058008	0.56	0.68	0.74	0.74	GOOD DIy BAD edito. BAD bio ci. BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058009	0.06	0.09	0.05	0.06	BAD DIY c BAD edito. GOOD bio GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058010	0.19	0.21	0.25	0.29	GOOD DIy BAD edito. BAD bio ci. BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058011	0.06	0.10	0.05	0.03	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. GOOD e
SPRT_002_060_117_0058012	0.11	0.20	0.13	0.18	GOOD DIy BAD edito. BAD bio ci. GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058015	0.50	0.57	0.56	0.41	BAD DIY c GOOD edi GOOD bio BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058016	0.33	0.42	0.34	0.31	BAD DIY c GOOD edi GOOD bio BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058017	0.17	0.30	0.41	0.18	GOOD DIy GOOD edi BAD bio ci. BAD featu BAD blog i. BAD est
SPRT_002_060_117_0058019	0.37	0.43	0.29	0.65	BAD DIY c BAD edito. GOOD bio GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058020	0.18	0.23	0.11	0.11	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. GOOD e
SPRT_002_060_117_0058023	0.08	0.11	0.05	0.04	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058024	0.07	0.13	0.06	0.07	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058027	0.15	0.36	0.18	0.21	GOOD DIy BAD edito. BAD bio ci. GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058028	0.08	0.10	0.05	0.05	BAD DIY c GOOD edi GOOD bio GOOD fea BAD blog i. BAD est
SPRT_002_060_117_0058030	0.07	0.06	0.08	0.07	BAD DIY c GOOD edi BAD bio ci. BAD featu GOOD bio BAD est
SPRT_002_060_117_0058033	0.12	0.12	0.06	0.10	BAD DIY c GOOD edi GOOD bio BAD featu BAD blog i. BAD est

Fig. 5.4 SPRIT genre classification by the inferential model

**Table 5.12** Results of the matching

PM = Perfect Match	4	0.4%	(Including 2 zerolabelling)
EM = Excellent Match	3	0.3%	
VG = Very Good Match	33	3.3%	
G = Good Match	92	9.2%	
FM = Fair Match	279	27.9%	
NM = No Match	415	41.5%	(Including 118 overlabelling)
NC = Not Classified	174	17.4%	
Total number of web pages	1,000		

By summing up PM, EM, VG, G, and FM, we get a percentage of 41.1%. If we exclude from this preliminary assessment NC web pages (i.e. 17.4%), we basically have a parity between misclassifications (NM) and matching accuracy (PM+EM+VG+G+FM). These results are in line with Freund et al. [12], WEGA [31], where similar assessment criteria were applied.

### 5.5.5.2 Overlap Coefficients

In this section, I list the overlap coefficients that result from the comparison between my manual annotation and the actual results returned by the model on the SPIRIT sample, immersed in three corpora of different size.

The overlap coefficients are functions that measure the agreement in the attribute sets of two objects. There are many different overlap coefficients. They measure the similarity between the items in two vectors. Here two common coefficients are used, the Dice and the Jaccard coefficients. Both measures range from 0.0 (no overlap) to 1.0 (perfect overlap).

In Table 5.13, the overlap coefficients are measured on 1,000 web pages of the SPIRIT sample including the 380 web pages labelled as IDK (i.e. I DO NOT KNOW), NOTHING SUITABLE IN THIS PALETTE and NOTHING ELSE SUITABLE IN THIS PALETTE.

In Table 5.14, the overlap coefficients are measured on 620 web pages of the SPIRIT sample excluding the 380 labelled as IDK (i.e. I DO NOT KNOW), NOTHING SUITABLE IN THIS PALETTE and NOTHING ELSE SUITABLE IN THIS PALETTE.

In both cases, the overlap coefficients are quite low. This can be explained by the fact that my manual annotation is limited to a small number of labels (max four), while the inferential model tends to overextends some genres, for instance EDITORIAL.

**Table 5.13** Overlap coefficients on 1,000 SPIRIT web pages

Web corpora	Jaccard	Dice
1,000 spirit within 2,400 web pages (SANTINIS)	0.12	0.18
1,000 spirit within 3,685 web pages (SANTINIS+KI-04)	0.12	0.18
1,000 spirit within 6,400 web pages (SANTINIS+KI-04+HGC+MGC)	0.10	0.15

**Table 5.14** Overlap coefficient 620 SPIRIT web pages

Web corpora	Jaccard	Dice
620 spirit web pages within 2,400 web pages (SANTINIS)	0.17	0.26
620 spirit within 3,685 web pages (SANTINIS+KI-04)	0.18	0.26
620 spirit within 6,400 web pages (SANTINIS+KI-04+HGC+MGC)	0.15	0.22

## 5.6 Discussion

The inferential model appears to be robust when cross-tested on three noisy genre collections of increasing size. Unfortunately, these findings are not directly comparable with other results, because this experimental setting has never tried out earlier. Hopefully, the results presented here can be used as baselines for future experiments.

Performance on the four BBC web genres (only 80 web pages all in all) are stable (see Table 5.3). On the seven novel web genres, the initial performance on SANTINIS (i.e. 86%), only decreases of 5% (i.e. 81%) when the initial corpus is increased of 35% (i.e. SANTINIS+KI-04). Even more encouragingly, an increase of 60% of the initial corpus (i.e. SANTINIS+KI-04+HGC+MGC) only causes 10% decrease (i.e. 76%) with respect to the initial performance (i.e. 86%). (see Table 5.4).

Table 5.5 appears to be very informative. It shows that accuracy increases with the number of features, but then it is hard to assess if this high dimensionality ensures also a certain degree of generability of the resulting genre models. More experimentation with structured and unstructured noise is certainly wished for. For the time being, it seems encouraging that a small amount of corpus-independent features (i.e. the 100 facets) achieves an accuracy of 76% on the 7-webgenre collection in a very noisy corpus of 6,404 web pages. It would be interesting to investigate in future whether a slight increase in the number of facets (e.g. up to 200) would also enhance accuracy.

The inferential model's performance is also appreciable on the mapped genres of KI-04 (see Table 5.6). Its performance equals that reported in Meyer zu Eissen and Stein [22] and keeps steady also when the corpus is drastically increased in size. Performance is very good when it is applied to HGC mapped genres in isolation (1,180 web pages), while an unexpected drop of accuracy occurs on the FEATURE genre, when HGC mapped genres are cross-tested on the enlarged corpus (6,404 web pages) (see Tables 5.7 and 5.9). Leaving the error analysis of the feature genre drop to future work, it appears that the model can be safely and effectively exported. This is indeed the main advantage of a corpus independent classification model. However, the performance is not too brilliant on the MGC mapped genres (see Tables 5.8 and 5.10). While the accuracy on MGC PERSONAL BLOGS is remarkably stable regardless the size of the underlying corpora (1,539 web pages vs. 6,404 web pages), the other MGC mapped genres show a more controversial composition. For instance, the MGC SHOPPING class mapped to my model's ESHOPS includes online stores, classified ads, price comparators and pricelists. The model performs

disappointingly on this SHOPPING class because it was not designed to cover classified ads and price comparators. In short, the low performance on MGC SHOPPING, FAQs and INDEX genres is due, I conjecture, to the composition of these classes. This collection seems to be quite hard to handle (cf. also [17]) possibly because of the characterization of genre classes underlying its construction. More in-depth error analysis and additional comparative experiments will shed more light on this behaviour.

Both the multi-labelling assessment described here and other experiments (cf. [42]) show that multi-labelled genre classification has a long way to go. The performance on multi-labelled classification is problematic for a number of reasons. On the one hand, overlap coefficients (see Tables 5.13 and 5.14) show that that my manual annotation needs to be discussed, agreed upon and validated. On the other hand, the model classification skills need to be refined and optimised, e.g. with the introduction of a threshold that skims off less probable genre labels. However, matching accuracy (Table 5.12) conforms to the assessment of the WEGA genre add-on [31]. These can be considered the current baselines for genre multilabel classification.

## 5.7 Conclusion and Future Work

The genre model presented in this chapter has been conceived to be as corpus-independent as possible. Hence, it is not derived from a single, supposedly representative corpus. The model is grounded on the findings of previous linguistic and textual genre analyses. It relies on NLP tools that extract linguistic knowledge from texts. The design is based on the intuition that encoding genre knowledge in the model rather than deriving through ML might ensure more resilience in noisy environments, like the web. Being corpus independent, the model does not depend on genre-annotated examples and can be applied in a situation where not all the web pages are labelled by the genres included in the model palette. Corpus independence is important when dealing with genre classes, because, as pointed out earlier, annotating a web page by genre is one of the major burdens of AGI research. Essentially, the inferential model tries to make the best of theoretical and empirical findings documented by genre analysts by encoding them in hard-coded rules, rather than relying on the learning from small genre collections assembled with subjective criteria.

The 7-webgenre collection and the other collections included in the experiments have been used to evaluate results and not for learning genre classes.

The model employs only 100 facets. It would be interesting to investigate in the future if the increase of the number of facets would positively affect classification results.

The inferential model is also topic-independent and relies on the correlation between genres and the rhetorical patterns that express the communicative purposes

in a text. It implements a zero-to-multi classification scheme where a web page can be assigned to one, more, or none of the genres of the palette.

The genre palette is purely experimental. This palette is just one of the many plausibly useful genre palette for genre-enabled applications.

The model has been cross-tested with four different genre corpora. Performance has been monitored through scalability (i.e. size increase), exportability to genre collections (i.e. HGC and MGC in isolation), and multi-labelling. However, multi-labelling genre classification is such in a premature stage that any further experience will be beneficial.

All in all, the model has been cross-tested with 6,404 genre-annotated web pages. I conjecture that this final composite corpus of 6,404 web pages well represents a noisy environment like the web. In this difficult scenario, the model shows some robustness and stability, but results need to be enhanced. The results shown here must be taken as baseline that will hopefully be overperformed in future experiments.

Results on MGC show that diverse definitions of the concept of genre have a strong bearing on the characterization of genre classes. When the conception of genre is not so distant, as in the case of HGC, results are more encouraging. This means that the diverse definitions of the concept of genre might have a strong bearing on the characterization of genre classes, thus affecting the generability of genre models as a whole.

One urgent need is the creation of genre reference corpora for evaluation purposes (a proposal can be found in Santini and Sharoff [32]). The construction of these corpora would entail profitable discussion and the formation of more consent around the concept of genre for AGI. A preliminary genre palette for reference corpora has already been proposed in Rehm et al. [23]. This palette is a good starting point for future debate.

More generally, a view of genre population on the web could be provided by the application of webometric techniques. Chapter 12 by Lennart Björneborn (this book) shows some interesting relations between the genres of academic websites. Other approaches are also possible. For instance, a genre-oriented replication of the experiences described in Thelwall [38–40] could undoubtedly provide new insights and a better understanding of the dynamics underlying genre use on the web.

Semi-supervised ML techniques and the exploitation of the tagging habits encouraged by social networks are certainly paths to be explored to assemble large quantities of genre-annotated material.

## Appendix

The appendix contains tables describing the genre corpora used in the experiments explained in Chapter 5.

**SANTINIS (2,480 Web Pages). Cf. Also Santini ([29], Appendix B)****Table 5.15** SANTINIS composition

SANTINIS		no. of web pages	Description
Noise	A.K.A. the spirit sample	1,000	A randomly selected sample of english web pages from the SPIRIT collection
Blogs			These are personal blogs where the author (a blogger) expresses whatever s/he thinks, fears or experiences using entries posted in reverse chronological order
Eshops		200	Eshops are interactive documents, with their own purpose (i.e. selling products), rhetorical function (persuade potential buyers), textual conventions (i.e. use of exhortations and a special typographical organization), and expectations (i.e. prices, pictures, short descriptions, offers, etc.). An eshop is often organized as a list of products with prices
FAQs		200	Questions and answers can be organized in different ways. For example, a FAQs can be a single document with a regular pattern of question + answer sequence; or each page can contain a single question and an answer; or all the questions are listed in one page hyperlinked to other pages (answers are provided in one different page per question); and similar
Front pages		200	A front page is the first page of a newspaper. In the paper world it was mainly a component of a newspaper bearing the initial part of the most important articles. On the web, a front page has a more autonomous status than in the paper world

**Table 5.15** (continued)

SANTINIS	no. of web pages	Description
Listings	200	Namely hotlists, sitemaps, tables of contents, and checklists. broadly speaking, a list is a synthetic way of delivering information or giving instructions. the visual organization into bullets or numbers provides a visual support that replaces other linguistic devices, such as connectives. Lists can be seen as a textual solution where juxtaposition prevails over subordinating constructions
Personal home pages	200	A personal home page is a page published and maintained by an individual. Personal home pages have been defined as “narrative of self-evaluation” and analysed also in terms of the construction of identity. However, Döring [8] finds that only 42% of the personal home pages listed in the university directories corresponded to the image of the typical self-presentation page
Search pages	200	Search pages belong to the fourth generation sites, those that focus on the architecture with dynamic content
BBC editorials	20	Argumentative statements of views that are considered to be representative of a newspaper as a whole
BBC DIY mini-guides	20	Include some general information about the project, duration time, etc.; a list of tools needed to operate; a short headline introducing the topic of the following paragraph; finally a sequence of instructions

**Table 5.15** (continued)

SANTINIS	no. of web pages	Description
BBC short biographies	20	Characterized by narration. Several recurrent linguistic features can be identified in a biography, for instance the past tense is used to report events together with temporal markers and location markers
BBC features	20	Articles about a specific subject or theme
Total	2,480	

***KI-04 (1,205 Web Pages). Cf. Also Meyer zu Eissen and Stein [22]***

**Table 5.16** KI-04 composition

KI-04	no. of web pages	Description
Articles	127	Documents with long passages of text, such as research articles, reviews, technical reports, or book chapters
Download pages	151	Pages on which freeware, shareware, demo versions of programs etc. Can be downloaded
Link collections	205	Documents which consist of link lists for the main part
Portrayal (priv.)	126	Private self-portrayals, i.e. typical private homepages with informal content
Discussions	127	All pages that provide forums, mailing lists or discussion boards
Helps	139	All pages that provide assistance, e. g. Q&A or FAQ pages
Portrayal (non-priv)	163	Web appearances of companies, universities, and other public institutions. That is, home or entry or portal pages, descriptions of organization and mission, annual reports, brochures, contact information, etc
Shops	167	All kinds of pages whose main purpose is product information or sale
Total	1,205	

**HGC (Used 1,180 for Crosstesting). Cf. Also Stubbe et al. [37]****Table 5.17** HGC composition

HGC		# of web pages	
<i>A Journalism</i>			320
1	Commentary	40	
2	Review	40	
3	Portrait	40	
4	Marginal note	40	
5	Interview	40	
6	News	40	
7	Feature	40	
8	Reportage	40	
<i>B Literature</i>			120
9	Poem	40	
10	Prose	40	
11	Drama	40	
<i>C Information</i>			360
12	Science	40	
13	Explanation	40	
14	Receipt	40	
15	FAQ	40	
16	Lexicon	40	
17	Bilingual	40	
18	Presentation	40	
19	Statistics	40	
20	Code	40	
<i>D Documentation</i>			120
21	Law	40	
22	Official	40	
23	Protocol	40	
<i>E Dictionary</i>			160
24	Person	40	
25	Catalogue	40	
26	Resources	40	
27	Timeline	40	
<i>E Communication</i>			160
28	Mail, talk	40	
29	Forum, guestbook	40	
30	Blog	40	
31	Form	40	
<i>F Nothing</i>			40
32	Nothing	40	
Total		1, 280	1, 280

***MGC (1,539 Web Pages). Cf. Also Vidulin et al. [41]*****Table 5.18** MGC composition

MGC	Genre Assignments (no. of web pages)	Description of the communicative purpose
Blog	77	Presents updates on what is going on with an entity
Children's	105	Presents content in a simple and colorful way specifically suited for children
Commercial/ promotional	121	Web pages are intended to invoke the visitor's interest in goods or services, typically for commercial gain
Community	82	Type web page involves the visitor in the creation of the page and enables interaction with other visitors
Content delivery	138	Delivers content that is not a part of the page. Entertainment web pages entertain the visitor
Entertainment	76	Presents jokes, puzzles, horoscopes, games
Error message	79	Tells the visitor to go away
FAQ	70	Are intended to help a user to solve common problems by answering frequently asked questions
Gateway	77	Transfers the visitor to another page. Index transfers the visitor to a selection of multiple other pages
Index	227	Presents lots of links to other web pages
Informative	225	Conveys objective information of permanent interest suitable for general population
Journalistic	186	Conveys mostly objective information on current events
Official	55	Conveys information with legal or otherwise official consequences
Personal	113	Conveys subjective, personal information in an informal way
Poetry	72	Presents poems and lyrics with intention to evoke emotions
Pornographic/adult	68	Web pages have intention to sexually arouse the visitor
Prose	67	Fiction presents story about real or fictional event in artistic form with intention to evoke imagination and emotions
Scientific	76	Conveys objective information suitable for experts
Shopping	66	Web pages sell goods or services online
User input	84	Solicits the visitor's input

**100 Facets****Table 5.19** 100 facets

---

1. Predicators	42. Discourseal connectives	76. That omission
2. Nominals	43. Temporal connectives	77. Comparative clause
3. First person	44. While	78. Relative clause
4. Second person	45. Whereas	79. Phenomenon registering
5. Third person	46. When	80. Con. action recording
6. Inanimate pronoun	47. Since	81. Action recording
7. Present tense	48. If	82. Phenomenon identifying
8. Past tense	49. As	83. Phenomenon linking
9. Imperative	50. To verb	84. Quality attributing
10. Active voice	51. Concession clause	85. Phenomenon identifying mod
11. Passive voice	(initial)	86. Act. demanding com.
12. Time markers	52. Concession clause	87. Layout (HTML)
13. Location markers	(final)	88. Typography (HTML)
14. Instrument	53. Concession clause	89. Functionality (HTML)
15. Manner	(special)	90. Navigability [general] (HTML)
16. Negative particles	54. Contrast clause	91. Navigability [external] (HTML)
17. Probability verbs	55. Exception clause	92. Navigability [internal] (HTML)
18. Necessity verbs	56. Reason clause (initial)	93. Web page length [in words]
19. Existential there	57. Reason clause (final)	94. Blog words
20. Expressiveness	58. Space clause (initial)	95. Eshop words
21. Colon	59. Space clause (final)	96. FAQs words
22. Question mark	60. Time clause (initial)	97. Front page words
23. Quotes	61. Time clause (final)	98. Listing words
24. Activity verbs	62. Time clause	99. Personal home page words
25. Communication verbs	(instructional)	100. Search page words
26. Mental verbs	63. Time clause (split)	
27. Causative verbs	64. Conditional clause	
28. Occurrence verbs	(initial)	
29. Existence verbs	65. Conditional clause	
30. Aspectual verbs	(final)	
31. Enumerative connectives	66. Conditional clause	
32. Equative connectives	(special)	
33. Reinforcing connectives	67. Result clause	
34. Summative connectives	68. Similarity clause	
35. Appositive connectives	69. Complex np	
36. Resultative connectives	70. Verb+that clause	
37. Inferential connectives	71. Adjective+that clause	
38. Reformulatory connectives	72. Wh clause	
39. Replacive connectives	73. Adjective+to clause	
40. Antithetic connectives	74. Verb+ing clause	
41. Concessive connectives	75. Purpose clause	

---

## References

1. Berninger V., Y. Kim, and R. Ross. 2008. Building a document genre corpus: A profile of the KRYS I corpus. Corpus profiling for information retrieval and natural language processing. *Workshop Held in Conjunction with IiiX 2008*, 18th Oct 2008. London.
2. Biber, D. 1988. *Variations across speech and writing*. Cambridge, UK: Cambridge University Press.
3. Biber, D. and Kurjian, J. (2007). Towards a taxonomy of web registers and text types: a multi-dimensional analysis. In *Corpus linguistics and the web*, eds., M. Hundt, N. Nesselhauf, and C. Biewer, 109–131. Rodopi – Amsterdam – New York.
4. Blood, R. 2000. Weblogs: A history and perspective. Rebecca's pocket. [http://www.rebeccablood.net/essays/weblog\\_history.html](http://www.rebeccablood.net/essays/weblog_history.html). Accessed 7 Sep 2000.
5. Bruce, I. 2008. Academic writing and genre. A systematic analysis. London-New York: Continuum International Publishing Group Ltd.
6. Dewdney, N., C. Vaness-Dikema, and R. Macmillan. 2001. The form is the substance: Classification of genres in text. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics*. Toulouse.
7. Dewe, J., J. Karlgren, and I. Bretan. 1998. Assembling a balanced corpus from the internet. In *Proceedings of the 11th Nordic Conference of Computational Linguistics*. Copenhagen.
8. Döring, N. 2002. Personal home pages on the web: A review of research. *Journal of Computer-Mediated Communication (JCMC)* 7(3).
9. Duda, R., J. Gasching, and P. Hart. 1979. Model design in the prospector consultant system for mineral exploration. In *Expert systems in the micro-electronic age*, ed. D. Michie, 153–167. Edinburgh: Edinburgh University Press. Reprinted in 1984.
10. Duda, R., P. Hart, and N. Nilsson. 1981. Subjective methods for rule-based inference system. In *Readings in artificial intelligence*, eds. B. Weber and N. Nilsson, 192–199. Palo Alto, CA: Tioga Publishing Company.
11. Freund, L. 2008. Exploiting task-document relations in support of information retrieval in the workplace. Doctoral dissertation, Faculty of Information Studies, University of Toronto, Toronto. [http://faculty.arts.ubc.ca/lfreund/Publications/Freund\\_Luanne\\_S\\_200811\\_PhD\\_thesis.pdf](http://faculty.arts.ubc.ca/lfreund/Publications/Freund_Luanne_S_200811_PhD_thesis.pdf)
12. Freund, L., C.L.A. Clarke, and E.G. Toms. 2006. Genre classification for IR in the workplace. In *Proceedings of Information Interaction in Context (IiiX 2006)* Copenhagen, Denmark.
13. Görlach, M. 2004. *Text types and the history of English*. Berlin-New York: Mouton de Gruyter.
14. Heyd, T. 2008. *Email Hoaxes. Form, function, genre ecology*. Amsterdam; Philadelphia, PA: J. Benjamins Publishing Company.
15. Joho, H., and M. Sanderson. 2004. The SPIRIT collection: An overview of a large web collection. *SIGIR Forum*, 38(2), December 2004.
16. Kanaris, I. and E. Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence*. Washington, DC.
17. Kanaris, I., and E. Stamatatos. 2009. Learning to recognize webpage genres. *Information Processing and Management* 45(5):499–512.
18. Karlgren, J., and D. Cutting. 1994. Recognizing text genre with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*. Kyoto.
19. Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC Jungle. *Language Learning & Technology* 5(3):37–72.
20. Levering, R., M. Cutler, and L. Yu. 2008. Using visual features for fine-grained genre classification of web pages. In *Proceedings of the 41st Hawaii International Conference on System Sciences*. Big Island, Hawaii.

21. Mason, J., M. Shepherd, and J. Duffy. 2009. An n-gram based approach to automatically identifying web page genre. In *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences*. Big Island, Hawaii.
22. Meyer zu Eissen, S., and B. Stein. 2004. Genre classification of web pages: User study and feasibility analysis. In *Advances in artificial intelligence*, eds. S. Biundo, T. Frühwirth, and G. Palm, 256–269. Berlin: Springer.
23. Rehm, G., M. Santini, M. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. 2008. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of LREC 2008*, May 28–30. Marrakech, Morocco.
24. Rosso, M. 2008. User-based identification of Web genres. *Journal of the American Society for Information Science and Technology* 59(7):1053–1072.
25. Santini, M. 2005. Building on syntactic annotation: Labelling subordinate clauses. In *Proceedings of the Workshop on Exploring Syntactically Annotated Corpora (held in conjunction with Corpus Linguistics 2005 Conference)*. Birmingham.
26. Santini, M. 2006. Common criteria for genre classification: Annotation and granularity. In *Proceedings of the Workshop on Text-based Information Retrieval (TIR-06) (held in conjunction with ECAI 2006)*. Riva del Garda.
27. Santini, M. 2007a. Automatic genre identification: Towards a flexible classification scheme. *BCS IRSG Symposium: Future Directions in Information Access 2007 (FDIA 2007a) (held in conjunction with the European Summer School on IR (ESSIR 2007))*, Tuesday, 28th and Wednesday, 29th of Aug. Glasgow.
28. Santini, M. 2007b. Characterizing genres of web pages: Genre hybridism and individualization. In *Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40)*. Hawaii.
29. Santini, M. 2007c. Automatic identification of genre in web pages. PhD thesis, University of Brighton, Brighton.
30. Santini, M. 2008. Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing and Management* 44(2):702–737.
31. Santini, M., and M. Rosso. 2008. Testing a genre-enabled application: A preliminary assessment. In *Proceedings of Future Direction in Information Access (FDIA-2008)*. BCS, London.
32. Santini, M., and S. Sharoff. 2009. Web genre benchmark under construction. *Journal for Language Technology and Computational Linguistics (JLCL)* 24(1):129–145.
33. Santini, M., R. Power, and R. Evans. 2006. Implementing a characterization of genre for automatic genre identification of web pages. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL/COLING 2006)*. Main Conference Poster Paper. Sydney.
34. Shepherd, M., C. Watters, and A. Kennedy. 2004. Cybergenre: Automatic identification of home pages on the web. *Journal of Web Engineering* 3(3–4):236–251.
35. Stein, B., and S. Meyer zu Eissen. 2008. Retrieval Models for Genre Classification. *Scandinavian Journal of Information Systems (SJIS)* 20(1):91–117.
36. Stubbe, A., and C. Ringlstetter. 2007. Recognizing Genres. In *Abstract Proceedings of the Colloquium "Towards a Reference Corpus of Web Genres" (held in conjunction with Corpus Linguistics 2007)*, 27 Jul 2007, eds. M. Santini and S. Sharoff. Birmingham.
37. Stubbe, A., C. Ringlstetter, and K. Schulz. 2007. Genre to classify noise – noise to classify genre. In *Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*, 8 Jan 2007. Hyderabad, India. *International Journal on Document Analysis and Recognition (IJ DAR)*, Dec 2007.
38. Thelwall, M. 2008a. Text in social network web sites: A word frequency analysis of Live Spaces. *First Monday* 13(2).
39. Thelwall, M. 2008b. Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology* 59(11):1702–1710.

40. Thelwall, M. 2008c. Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology* 59(1):38–50.
41. Vidulin, V., M. Luštrek, and M. Gams. 2007. Using genres to improve search engines. In *Proceedings of Towards Genre-enable Search Engines: The Impact of Natural Language Processing Workshop*, Sept 2007. Borovets, Bulgaria.
42. Vidulin, V., M. Luštrek, and M. Gams. 2009. Multi-label approaches to web genre identification. *Journal for Language Technology and Computational Linguistics (JLCL)* 24(1):97–114.
43. Waltinger, U., and A. Mehler. 2009. The feature difference coefficient: Classification by means of feature distributions. In *Proceedings of the Conference on Text Mining Services (TMS 2009)*, 159–168. Leipzig, Germany.
44. Xu, J., Y. Cao, H. Li, N. Craswell, and Y. Huang. 2007. Searching documents based on relevance and type. In *Proceeding of ECIR 2007*. Rome, Italy.
45. Yeung, P., S. Büttcher, C. Clarke, and M. Kolla. 2007a. A Bayesian approach for learning document type relevance. *ECIR 2007*. Rome.
46. Yeung, P., C. Clarke, and S. Büttcher. 2007b. Improving retrieval accuracy by weighting document types with clickthrough data. *SIGIR'07*. Amsterdam, The Netherlands.
47. Yeung, P., L. Freund, and C. Clarke. 2007c. X-Site: A workplace search tool for software engineers. *System demo presented at the 30th International ACM SIGIR Conference*. Amsterdam.

# Chapter 6

## Formulating Representative Features with Respect to Genre Classification

Yunhyong Kim and Seamus Ross

### 6.1 Introduction

Document classification is one of the most fundamental steps in enabling the search, selection, and ranking of digital material according to its relevance in answering a predefined search. As such it is a valuable means of knowledge discovery and an essential part of the effective and efficient management of digital documents in a repository, library, or archive. Document classification has previously been dominated by the classification of documents according to topic. Recently, however, there has been a growing interest in the classification of documents with respect to factors other than topic (e.g. classification into forms of dissemination such as scientific papers, emails, blogs, and news reports). This type of classification has been labelled in many different ways, including the phrase *genre classification*. The vast number of different contexts in which genres have emerged across classification attempts illustrate that genre is a high-level, context-dependent concept (cf. literature review [24]). Genre has been referred to as aspects of the text described by level of information or degree of elaboration, persuasion and abstraction (cf. [5]), as well as, to common document forms such as FAQ, Job Description, Editorial or Reportage (e.g. [9, 14, 16]). In some cases, genre has been used to describe the classification of a document according to whether or not it is a narrative and the target level of audience (e.g. [16]), and whether it is fact or opinion, and, in the case

---

Y. Kim (✉)

Humanities Advanced Technology and Information Institute (HATII),  
University of Glasgow, Glasgow, UK; School of Computing, Robert Gordon University,  
Aberdeen, UK  
e-mail: ykim1@rgu.ac.uk

The work presented in this chapter was supported by DELOS: Network of Excellence on Digital Libraries (<http://www.delos.info>), funded by European Commission's IST FP6 [G038-507618], and the Digital Curation Centre (<http://www.dcc.ac.uk>), funded by the Joint Information Systems Committee (JISC, <http://www.jisc.ac.uk>) and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC, <http://www.epsrc.ac.uk>)[GR/T07374/01].

of opinion, whether it is positive or negative (e.g. [10]). On occasion it has been used to describe membership to selected journals and brochures (e.g. [1]), and, to denote similar feature cluster groups (e.g. [2, 21]).

Despite the elusive nature of genre, it is undoubtedly true that being able to bind together tools trained to retrieve information within selected genre domains would be invaluable to automating the ingest, management and preservation of material in digital repositories (cf. [23]). This is especially true where metadata describing the technical characteristics, function, source and content of digital material play a core role in the efficient and effective management and re-use of the same. The manual collection of metadata is labour-intensive, costly and susceptible to variation in quality and precision across different actors; automating the process of semantic metadata extraction is, therefore, essential. Past efforts (e.g. [3, 7, 11, 12, 15, 25]) to extract metadata automatically from digital documents have relied heavily on the structure that characterises the genre class to which the document under consideration belongs. The reliance of these methods on document structure emphasises the benefits of constructing a tool that enables automated genre classification. An effective automated genre classifier would function as an overarching tool for integrating genre-specific tools and, in any case, provide a first-level classification of documents into those of a similar structure, which would facilitate the extraction of further information.

The interest in forms of documents classification other than that of topic is also growing in the area of information retrieval and reflects the limitations of relevance measurements defined on the basis of topical similarity. Topic alone does not provide insight into whether or not a retrieved document is relevant to your purpose; a document with the same topic may be created with different objectives resulting in different levels of usefulness as a source of information (e.g. compare an advertisement about a camera to a product review of the same camera). These objectives of document creation seems to be at the centre of what characterises document genre. On the other hand, these objectives define the functional requirements imposed on the document (e.g. to narrate, to argue against, to argue for, to present research results, to record) and the structures found within the document are designed to meet these functional requirements. In this chapter we do not claim a deep understanding of the nature of genre, but merely are driven by the observation that the structural classification of documents is a fundamental component in understanding a document with respect to its purpose and function.

Classical models of document classification largely depend on term frequency weighting and counting instances of specified linguistic constructs. The former does not reflect much conceptual structure and the latter results in a highly language dependent model that incorporates some local conceptual structure but largely disregards the global structure of the document and its components. In this chapter we examine the role of word distribution pattern in classifying documents. More specifically,

- we describe an approach to document representation that incorporates more document structure by considering how strings are distributed throughout the document (Section 6.2.2), and,
- give evidence that this approach is better than the bag-of-words approach by comparing it against the rainbow classifier developed by McCallum (see [20] and Section 6.6.2).

It is not the purpose of this chapter to advocate the structural classification of documents as a definition for genre classification, but to show by experimental evidence that our model may be more appropriate in dealing with high level concepts (such as genre). We are not disputing the fact that genre is a social construct (cf. the Chapter 2 by Karlgren, this volume) and that it is the social context that defines genre. We wish to merely state that, just as the phenotype of a group of genetically distinct organisms (e.g. whales and fish) may lead to the extinction or survival of the entire group (e.g. if the water should become contaminated), the structure of a document is likely mirror the social objectives related to the document creation and provide a key to gauging the usefulness of a document and extracting further information. In particular, we report evidence that some of the previously established genre schemas and collections are better distinguishable by our distribution model than previously reported results.

The importance of structure has also been discussed elsewhere (e.g. the Chapter 1, by Lindemann and Littig, this volume) but, while others have introduce structure as the measurements of *structural entities* within the document distinct from topical terms or content, we will be discussing structure as an organisation of terms (regardless of topicality) throughout the document (Section 6.2.2) akin to *burstiness of terms* discussed in [6] and again in [8].

The combined representation of content and structure that we are attempting to establish in this chapter is also intended to raise questions about a prevailing notion in earlier analyses that genre classification is a task orthogonal to topic classification (e.g. the Chapter 8 by Stein et al., this volume). While this may be true on a conceptual level, there is reason to believe that this may not be a statistically sound approach. For example, the topic of *algebraic variety*, a well-known subject area in higher mathematics, would not be expected to appear as frequently in the genre class Reportage as it would in the genre class Research Article. In fact, preliminary results from a recent experiment, classifying documents belonging to 10 genre classes into twenty newsgroup topic classes, shows that, while there are genre classes whose documents are randomly distributed across the 20 topics (e.g. Poem), there are also genres 95% of whose documents are classified into only four newsgroup topics (e.g. Minutes). Given these examples where genre is interactively intertwined with topic, it would seem beneficial to build a general classification model that encompasses both tasks. With this in mind, we would like to introduce genre classification, not as a classification task distinct from topic classification, but as a point in a continuum of classifications, emphasising both genre classification and topic classification as a special case of a general abstract classification model.

## 6.2 Defining Genre Classification

### 6.2.1 Document Representation in Conventional Text Classification

The conventional method of text classification can be contracted to a formula for the weight of a term  $T$  within a document expressed by:

$$TF \times IDF \times N \quad (6.1)$$

where  $TF$  denotes the frequency of the term in the document,  $IDF$  denotes the inverse of the number of documents in the collection containing the term, and  $N$  denotes a normalisation factor dependent on the length of the document. The calculation method of each of these terms differs according to the research or application in question. This model is based on the notion that:

- if a term appears frequently in a document, it is likely to be a characterising feature of the document;
- if a term appears across several documents, then it is not likely to be a strong feature in distinguishing any one of those documents from the others; and
- if the same term appears in equal numbers within a short document and a long one, then it is likely to be a stronger feature of the short document.

While it may be considered a gross simplification to represent all the various classification methods by this one description, it still seems true that the basic principles that drive various text classification methods are closely related to this model. In a subject classification task, the term may surface as words or  $N$ -grams ( $N$  consecutive words or characters), while in other classification tasks term may manifest itself also as functional groups of words (e.g. verb) or combinations of such words and phrases and groups. Nevertheless, the mechanism driving the classification is largely dependent on counting patterns, and weighing the number against the pattern count throughout the collection being examined. The location of patterns, the relationship between instances of the patterns, and the interplay between different types of patterns are largely by-passed and only represented implicitly through the pattern of the expression being counted.

### 6.2.2 Harmonic Descriptor Representation (HDR) of Documents

A document can be described as a sequence of symbols. Symbols should not be confused with the alphabet of a natural language, although they may take the form of alpha-numeric characters in some instances. In the present terminology, each symbol may form any group of these characters or a much larger set of characters (e.g. white space, %, + and ?) and could also refer to the functional category of a group of characters (e.g. the part-of-speech).

Because of its static appearance, a document is often misunderstood to be time independent, but the interpretation of each symbol is possible only as a consequence of its temporal relationship to other symbols. In this light, document classification can be considered to be a subtask of signal processing. Viewed in this way, an accurate measure of term frequency is expressed by how many times a symbol occurs with respect to time. The term weight calculated in Section 6.2.1 presents no awareness of the role of temporal progression in the semantic analysis of the document. That is, if the word “clock” were to appear in two documents 10 times, then the weight of this word would be equal with respect to both documents: the fact that the word appears only in the first half of the document with respect to one of the documents in contrast to being evenly distributed throughout the document (which may be the case with respect to the other document) would be disregarded. A proper consideration of the time dimension would suggest “clock” in the first document as a signal having twice the frequency of that of the second document, but lasting only half the length of time. Time should not be taken to be the length of the text. Although the two are closely related, the length of the text is not equivalent to the tempo of the piece of writing, beginning with an introduction and ending with a conclusion. To understand the notion of time, we will compare a document to a string of a musical instrument. An occurrence of a symbol within the document partitions the document into two parts. If the two partitions are equal in length, then the phase division is akin to a harmonic with twice the frequency of the fundamental of the string (the document with zero occurrence of the symbol). If the division is not equal, then the frequency can not be considered to be uniform throughout the document.

In the case of topic detection, a loose application of time (e.g. taking the frequency to be uniform throughout the document) may be sufficient to capture salient vocabulary, but in other types of classification, where the main interest lies in the physical or conceptual structure of the object, the lack of temporal and relational placement of symbols contributes to a considerable loss of information. To fill this gap, we propose incorporating the symbol’s range and period as an effective means of characterising the symbols in the document. We will refer to this characterisation as the Harmonic Descriptor Representation (HDR) of the document (inspired by the musical analogy given above). We define range as the interval between the initial and ultimate occurrence of the symbol, and period as the time duration between two consecutive occurrences of the symbol. When the symbol occurs at regular intervals, the resulting signal in the document is akin to a harmonic of the document as a wave. Brookstein et al. [6] observed that content-bearing words would clump together and therefore result in non-harmonic behaviour. In contrast to the content-bearing words that they discuss, our research focuses on words that may be indicative of style and structure. We observe that document structure is captured by words displaying both harmonic and non-harmonic behaviour; harmonic words define the physical structure of the document, while non-harmonic words define conceptual landmarks or structure. In our description, we attempt to capture the degree of non-harmonic behaviour using three quantities derived from the range and period of each symbol:

1. The time duration before the first occurrence within the document of the symbol ( $FP$ ), measured by the number of characters (including white space) before the symbol, divided by the number of characters in the entire document.
2. The time duration after the last occurrence of the symbol to the end of the document ( $LP$ ), measured by the number of characters after the last symbol divided by the number of characters in the entire document.
3. The average period ratio ( $AP$ ), defined as 1 if the maximum number of characters between two occurrences is zero, and, otherwise, as  $T/(N \times MP)$ , where:
  - $N$  is the total number of occurrences of the symbol plus one;
  - $MP$  is the maximum number of characters found between two consecutive occurrences of the symbol; and,
  - $T$  is the total number of characters in the document minus  $N$ .

The average period ratio is an average ratio of the distance between two occurrence over the maximum distance. It is intended to measure how regular the occurrences are, regardless of how far apart the actual occurrences are, as. The more harmonic the behaviour of a symbol, the closer  $AP$  will be to 1. The other two measures  $FP$  and  $LP$ , on the other, hand are intended to measure when the term is first introduced and how focused the occurrences are against the entire document. In Fig. 6.1, we display an example of six documents (D1–D6) of different lengths, portrayed as light-coloured strips where the top of the strip is the beginning of the document. Occurrences of symbols in the documents ( $s_1$ – $s_7$ ) have been represented as horizontal lines across the strips. The period between two consecutive occurrences have been indicated to be  $x$ . This example will be used in Figs. 6.2, 6.3, and 6.4 to demonstrate how  $FP$ ,  $LP$ , and  $AP$  change under different conditions.

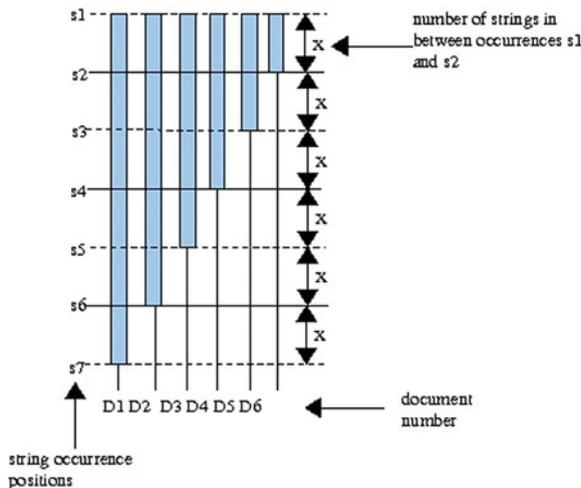


Fig. 6.1 Example of symbol occurrence in six documents of different lengths

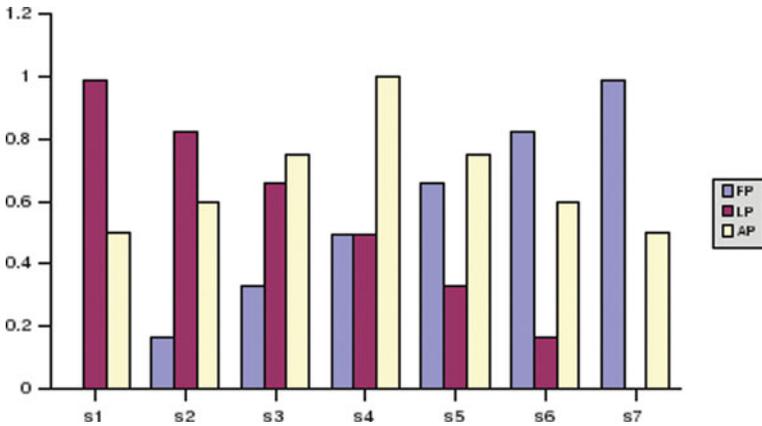


Fig. 6.2 *FP*, *LP*, and *AP* with respect to the position (*X*-axis) of a single occurrence of a symbol in *D1*

We present in Fig. 6.2, a graph illustrating how *FP*, *LP* and *AP* change as the position of a symbol occurring once in *D1* (see Fig. 6.1) changes from *s1*–*s7*. In Fig. 6.3, we show how *FP*, *LP* and *AP* for a symbol occurring twice in *D1* change with respect to the period between the two instances, as the second occurrence of the symbol moves away from the first occurrence. Finally, the graph in Fig. 6.4 presents how *FP*, *LP* and *AP*, for a symbol occurring once halfway between *s1* and *s2*, change as the document length varies.

Given a document, each word or symbol in the document is associated to their *FP*, *LP* and *AP* values. By taking all the words in a collection or by using a pre-compiled list of indicative words (say, in either case, the resulting word list is

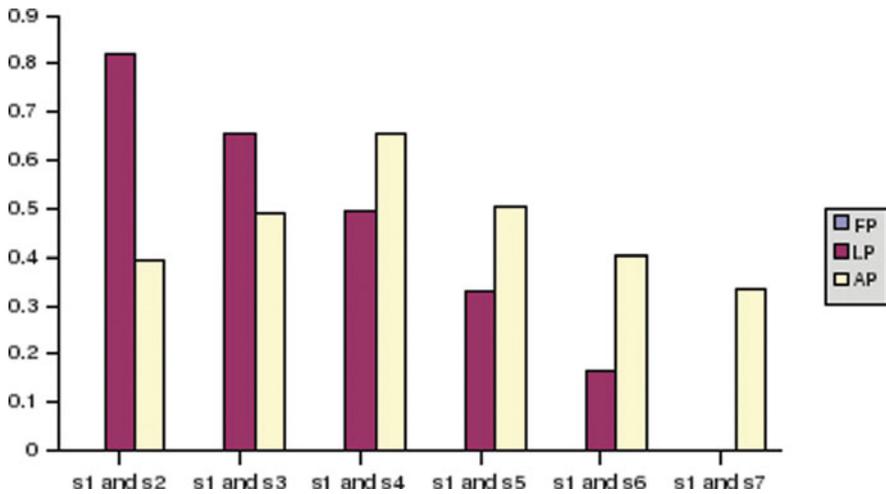
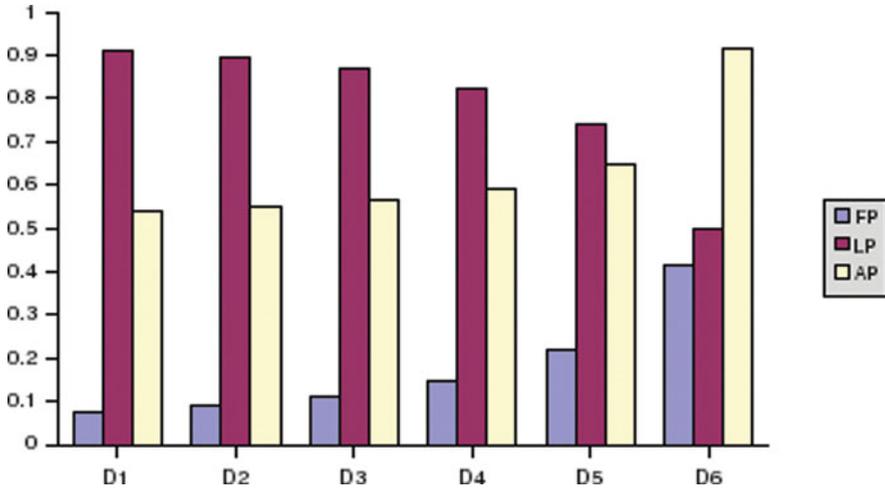


Fig. 6.3 *FP*, *LP*, and *AP* for a symbol occurring twice in *D1* as the period between the two instances become larger



**Fig. 6.4** *FP*, *LP*, and *AP* for a symbol occurring once in the same position relative to the beginning of different length documents

of size  $N$ ), each document can be represented as a vector of dimension  $3N$ , where each term in the vector is the *FP*, *LP*, or *AP* value of each word. In our model we pre-compiled a list of words from a sample dataset (which is discarded from the test dataset after the words are collected) by aggregating a list of words that appear in 75% of all the documents in at least one genre class in the sample dataset.

The relevance of term distribution has been mentioned by others including Manning et al. (see [19]), and, more recently, by De Roeck et al. (e.g. [8]) who carried out a study of profiling datasets to determine the degree of homogeneity or heterogeneity in the distribution of frequent terms. However, there have only been few explicit implementations of the measurement for the purpose of automated classification, and most of these previous analyses have been based on a count of words in selected chunks of the texts. Term *dispersion* measured using Juilland's  $D$  coefficient (formula to be found in [22]) also depends on examining selected texts within a larger collection, for variations of standard deviation in word frequency. The model presented here, on the other hand, compares relative distances between term instances, viewing the entire document as a time dependent whole, and does not involve arbitrary choices of text chunk sizes.

### 6.2.3 Defining Genre

While the definition of genre may not be easily pinned down, there is a shallow agreement that genre is a concept that can be used to categorise documents by structure and function. In fact, the structural properties (e.g. the existence of a title page, chapter, section, the number of columns, use of diagrams, and font variations) evolve in ways that are designed to optimise the document's capability to fulfil its functional intention(s) (e.g. to describe, to inform and to argue, to advertise) within

its target environment (e.g. the user community, publisher and creator), much the same as the structure of an organism evolves to optimise its survival function in the natural environment (cf. Kim and Ross [18]). As a consequence, genre reflects one or more of the following:

- the intention of the creator (e.g. to inform, to argue, to instruct);
- the interpretation of the user community (e.g. as a collection of facts, an expression of opinion, a piece of research);
- the prescription of a process (e.g. article for journal publication, job description for recruitment, minutes of a meeting); and
- the type of data structure (e.g. table, graph, chart, list).

The model described in Section 6.2.1, while effective in distinguishing some intentional and interpretive aspects of genre, seems insufficient to capture distinguishing features in the case of prescriptive, conceptual or physical structure. Such structure can be characterised even by low frequency terms of the class (e.g. single occurrence of “minutes” in the title of meeting minutes, or paragraph headings in a curriculum vitae), and the distributional pattern of words throughout the document (variation of density) is often bound to its class (e.g. the even distribution of wh-words in a FAQ sheet). The last observation is a generalisation of the observation by Brookstein et al. [6], who noted the clumping properties of content-bearing words and their role in text classification. In contrast to the content-bearing words that they discuss, we are interested also in words indicative of style and structure. These words can exhibit both clumping and uniform distributional properties. We present evidence that documents of each genre class display distinctive distributional characteristics and these can be more effectively captured using the HDR of documents introduced in Section 6.2.2.

A genre schema of seventy classes (KRYIS I corpus) was introduced in Kim and Ross [17, 18]. The schema was constructed and populated to represent the diverse range of intentional and structural aspects of genre listed above. At the time of building the corpus, we were focusing on document genres and, therefore, did not include webpage genres. In the experiments described in this chapter, we have compensated for the deficiency by further augmenting the schema with 7 webpage genres identified within the 7-webgenre collection introduced by Santini [24]. The inclusion of the 7-webgenre collection also enables us to compare our method to other results that have been achieved on the same dataset.

### 6.3 Classifiers

In Section 6.6, we will compare support vector machine (SVM) classification using the harmonic descriptor representation of documents modelled using Weka machine learning software [27] against the SVM classification performed using the Bow Toolkit rainbow text classifier developed by MacCallum [20], and the classification attempts of Santini [24], to show that the performance is consistently better when using the new description. The reason we have selected SVM as the classification method is that it showed the best results for rainbow when compared with

Rocchio/TFIDF and Naive Bayes. Also it has been evidenced to be effective in other text classification tasks as demonstrated by Yang et al. [26]. The rainbow text classifier, included in the BOW toolkit [20], indexes the alpha-numeric content of the text as a bag-of-words for an analysis of significant term frequencies, while Santini's method employs a combination of linguistically motivated features. The three way comparison was motivated by a desire to make a comparison of term distribution models (e.g. HDR), term frequency models (e.g. BOW) and linguistically motivated models (e.g. [24]).

## 6.4 Dataset

The dataset in our experiment consists of 24 classes from KRYIS I and the seven classes from the 7-webgenre collection, altogether consisting of 3,452 documents in 31 genres (see Table 6.1). The test was initially confined to 31 genres, partly, due

**Table 6.1** Scope of genres

Creative	Book of Fiction(29) Poem(90)
Determined by user context	Email(90) Exam/Worksheet (90) Form (90) Handbook (90) Letter (91) Minutes (99) Resumé/CV (96) Sheet music (90) Speech transcript (91) Technical manual (90)
Determined by organisational prescription	Abstract (89) Academic monograph (99) Advertisement (90) Business report (100) Magazine article (90) Scientific article (90) Memo (90) Periodicals (67) Poster (90) Slides (90) Technical report (91) Thesis (100)
Webpage genres	Blog (190) Eshop (190) FAQ (190) Front page (190) List (190) Personal home page (190) Search page (190)

to some computing problems. Although clever distributed computing might have circumvented the problem observed, it was not uncommon for the support vector machine on Weka to crash due to lack of memory. This problem seemed to arise especially when many classes or number of features are introduced into the classification. Increasing the number of documents did not seem to affect the system as badly as long as the number of classes and features are moderate (e.g. experiments on a newsgroup data consisting of nearly 20,000 samples in 20 classes represented by less than 300 features did not seem to cause the same difficulty). The 24 classes from KRYIS I were selected to reflect a proportion of classes from each of the ten genre groups presented in Kim and Ross [18].

A comparison of automated classification methods on a dataset that has not been tested for human agreement can give misleading information as human agreement analysis conveys to us how clean the dataset is and the nature of the genre class schema of the dataset. The experiments reported here were carried out on a collection consisting of the genres in Table 6.1 (numbers of documents in each genre, excluding those used to construct the word list in the previous section, are indicated in parentheses). The dataset for the twenty-four document genres were collected by:

1. assigning genres to collectors (in this case students) who retrieved from the Internet as many PDF files as they could find in English; and
2. having two classifiers (in this case secretaries) reclassify the PDF documents using the initial schema but without the knowledge of the initial label for each document.

None of the labellers were given a definition for the genres in the schema. This was partly to establish whether there was already a well understood genre vocabulary. The human performance was examined by taking the number of labels given by a single labeller in agreement with the other two labellers over the total number of documents on which the other two labellers agreed. The three numbers obtained in this way are 0.675, 0.73 and 0.829. Although the difference between the lowest and the highest recall is a noticeable 14%, this should be viewed with the knowledge that the highest recall is the result of student classification while the lowest recall is that of secretary classification. The human classification agreement on the KRYIS I corpus has been further analysed in the research presented in Berninger et al. [4]. User studies that have been presented here and Berninger et al. [4] are speculative and far from conclusive. The results that have been presented here have been provided mainly to give context to the dataset being used in the experiments. User studies with respect to the 7 webgenre collection is found in [24].

Other human labelling analyses of genre classification from the bottom up approach (i.e. giving the users the freedom to assign and define the genres) have been carried out in Chapter 3 by Rosso and Haas (this book). Note, however, that the numbers in their work are slightly different from the numbers that have been presented here, and, in [4]: while they examine overall agreement (e.g. number

of labels in agreement per document regardless of the labeller), [4] examine the agreement of selected labellers as well as overall agreement on a document.

## 6.5 Features

For the HDR SVM experiments reported in Section 6.6, we set aside a sample dataset consisting of ten random documents from each of the genres classes in the whole collection, and compiled all the symbols that appear in more than 75% of the documents in each genre. The symbols examined with respect to SVM HDR in the experiments reported here are simply white space delimited words in the document text,<sup>1</sup> inclusive of any HTML (Hyper Text Markup Language) tags. These tags are part of the vocabulary that indicates document structure and relations between entities in the HTML hybrid language, just as functional words (e.g. auxiliary verbs) might do in natural language. The compiled word list, in the current experiment, consisted of 2,477 words. Each of these words/symbols represent three features FP, LP, and AP (see Section 6.2.2) in our HDR of documents (i.e. each document is represented by a vector of dimension 7,431). The words/symbols compiled are expected to represent symbols that are prolific within at least one of the genre classes being examined (but not necessarily prolific within any one document). The list is expected to include stop words as well as HTML tags. As an illustration of the varying characteristics of vocabulary with respect to genre, we present (in Table 6.2) the number of selected word types (the range of types are indicated in the column labelled “WT”, in the table) found to be prolific (based on ten random documents from each genre) within the classes Poem, Letter and Thesis. The numbers were estimated manually by the author.

Most of the numbers in Table 6.2 are not very illuminating by itself in that the median lengths of documents belonging to Poem, Letter and Thesis are 1,718, 4,265, and 132,994, respectively (in bytes), that is, we expect the numbers to be increasing in that order for each type of word. However, we immediately notice an exception in this pattern with respect to subject pronouns, and, closer examination of the actual words show that at least one of the two subject pronouns found to be prolific in poems (i.e. “you” and “I”) is not found to be as prolific in letters (i.e. “it”) and theses (i.e. “I”, “we”, “they”, “it”). Further, the word “Dear” is only found to be prolific within letters.

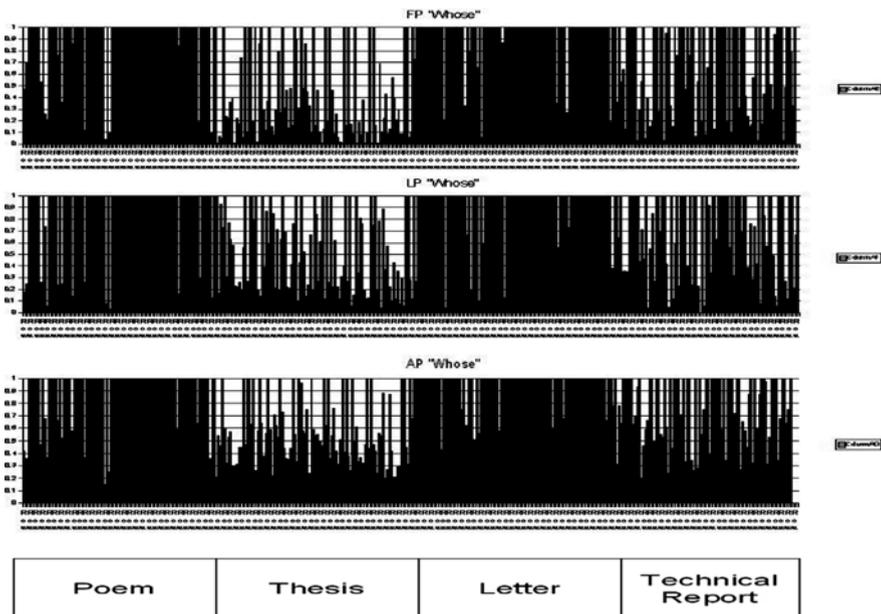
To illustrate how the FP, LP and AP of the HDR description varies across documents of the same genre we present a snapshot of these values with respect to the word “whose” across 90 poems, 100 theses, 91 letters and 91 technical reports in Fig. 6.5. The segments corresponding to the documents belonging each genre are indicated at the bottom of the figure. The figure shows that FP, LP, and AP are similar for documents belonging to the same genre but diverge as we move across documents belonging to different genres.

---

<sup>1</sup> Text was extracted from the PDF using the XPDF pdftotext tool (<http://www.foolabs.com/xpdf/>)

**Table 6.2** Number of words found in seven out of ten documents belonging to three genres (top row) with respect to word type (left column). Median length of documents in each genre are expressed in the parentheses next to the genre label as number of bytes

	Poem (1,718)	Letter (4,265)	Thesis (132,993)
Article	2	2	3
Wh-word	0	0	6
Modal	0	1	9
Have verb	0	1	3
Be verb	1	3	7
Verb	0	0	29
Noun	0	0	46
Subject pronoun	2	1	4
Object pronoun	0	0	1
Possessive pronoun	0	0	0
Possessive adjective	0	0	2
Adjective	1	1	43
Adverb	0	1	29
Quantifier	0	1	9
Demonstrative	1	2	6
Conjunction	1	3	9
Preposition	5	8	20
Punctuation	2	3	4
Other	1	1	12



**Fig. 6.5** Example of FP (*top*), LP (*middle*), and AP (*bottom*) values with respect to the word “whose” across documents belonging to four distinct genres (the documents corresponding to each of these genres are noted by segmentation indicated at the *bottom* of the figure)

## 6.6 Results

The performance will be evaluated using one or more of three conventional metrics: accuracy, precision and recall. To re-visit the definition for these terms, let  $N$  be the total number of documents in the test data,  $N_c$  the number of documents in the class  $C$ ,  $TP(C)$  the number of documents correctly predicted to be a member of class  $C$ , and  $FP(C)$  the number of documents incorrectly predicted as belonging to class  $C$ . Accuracy,  $A$ , is defined to be:

$$A = \frac{\sum TP(C)}{N}, \quad (6.2)$$

precision,  $P(C)$ , of class  $C$  is defined to be:

$$P(C) = \frac{TP(C)}{TP(C) + FP(C)}, \quad (6.3)$$

and recall,  $R(C)$ , of class  $C$  is defined to be:

$$R(C) = \frac{TP(C)}{N_c}. \quad (6.4)$$

In addition we also examine the average of  $P(C)$  and  $R(C)$  expressed as the  $F$ -measure  $F(C)$  defined as  $F(C) = 2 * (P(C) * R(C)) / (P(C) + R(C))$ . Although some debate surrounds the suitability of accuracy, precision and recall as a measurement of information retrieval tasks, for classification tasks they are still deemed to be a reasonable indicator of classifier performance.

It should also be mentioned here that all the results reported in this section are based on the average taken on ten-fold cross validation.

### 6.6.1 Overall Accuracy

The figures in Table 6.3 are the overall accuracies of the support vector machine rainbow classifier (SVM rainbow), the support vector HDR classifier (SVM HDR), and the average human agreement estimated by assuming that human agreement on the 7-webgenre collection is perfect. The classifier we are considering to be a baseline classifier in this comparison is the SVM rainbow classifier. The human agreement is included to indicate the cleanliness level of the dataset being used.

**Table 6.3** Overall accuracy across all 31 genre classes

Classifier	SVM rainbow	SVM HDR	Human avg
Overall accuracy	0.73	0.80	0.84 <sup>a</sup>

<sup>a</sup>Estimated assuming agreement is perfect on the 7-webgenre collection.

The numbers in Table 6.3 suggest that the performance level of the SVM rainbow classifier is already comparable to the average performance of three human labellers, and shows that the SVM HDR improves on the SVM rainbow classifier by 7%.

To test the limits on a cleaner dataset, we analysed the classification results with respect to the 7-webgenre collection. This is the overall accuracy of the classification when the recall of the documents belonging to the webpage genre classes is calculated upon the classification of the entire dataset into 31 classes. There is a slight increase of 0.002 when the webpage classes are classified on their own. The results are shown in Table 6.4: the numbers suggest that SVM HDR is a strong contender in webpage genre classification.

**Table 6.4** Overall accuracy of classifiers across webpage genres (Blog, Personal Home Page, FAQ, List, Search Page, EShop, Front Page)

Classifier	SVM rainbow	Santini's result	SVM HDR
Accuracy	0.92	0.89	0.96

### 6.6.2 Precision and Recall

The challenge in document classification is to improve the overall accuracy of the classification without compromising the performance with respect to any one class in the schema. In this section we will show that SVM HDR meets this challenge.

In Figs. 6.6 and 6.7, we present the recall and precision of SVM rainbow and SVM HDR with respect to each of our classes. The graphs show that SVM HDR outperforms SVM rainbow with respect to most of the classes in both recall and precision. The recall of SVM rainbow with respect to Academic Monograph, Book of Fiction, Front Page (of a website), Minutes, periodicals, Technical Manual and Thesis is Marginally higher than SVM HDR and the precision of SVM rainbow with respect to Abstract, Exam/Worksheet, Home Page, Poem, and Slides is somewhat higher than that of SVM HDR. However, with respect to the majority of the classes, SVM HDR outperforms SVM rainbow.

The graphs also demonstrates that SVM rainbow's performance varies widely across different genres, while the deviation of performance is much more confined in the case of SVM HDR. The recall (resp. precision) of SVM rainbow ranges from 0.08 to 1 (resp. 0.24–0.99), while recall (resp. precision) of SVM HDR ranges from 0.42 to 1 (resp. 0.38–0.99). The difference between precision and recall with respect to each class is also notable: the maximum absolute difference between precision and recall across the genre classes for SVM HDR is observed at approximately 0.24, while the same for SVM rainbow is observed at 0.46. The small deviation of performance across classes and the comparability of precision and recall with respect to each class seems to suggest that HDR is more successful in characterising the genre classes.

The graph in Fig. 6.8 presents the F-measures of SVM rainbow and SVM HDR with respect to each class. This graph shows that the F-measures of SVM HDR

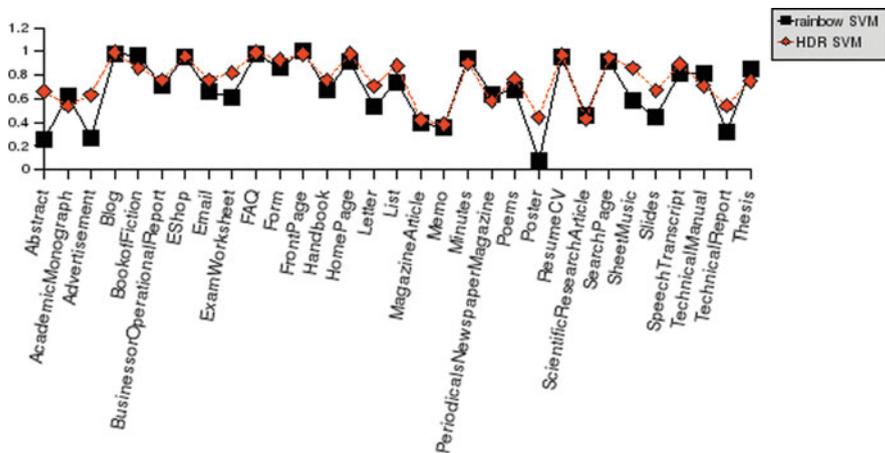


Fig. 6.6 Recall: a comparison, SVM rainbow and SVM HDR

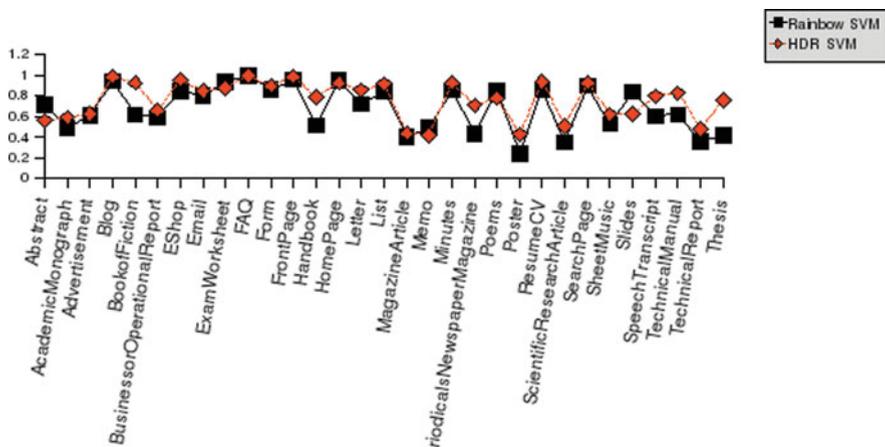


Fig. 6.7 Precision: a comparison, SVM rainbow and SVM HDR

are greater than those of SVM rainbow with respect to every class except the class Memo. With respect to Memo, the difference is 0.02 in favour of SVM rainbow. Latest experiments using HDR to analyse a newsgroup dataset of 19,597 documents in 20 topical classes (obtained from McCallum’s website<sup>1</sup>), show that the same SVM HDR model is also promising in topic classification, with an overall accuracy of over 95% (detailed report of this experiment available shortly). A list of 82 words was compiled from 400 documents (20 documents from each genre) set aside from the original 19,997 documents for this experiment. We have also calculated the F-measures of SVM HDR with respect to the classes in this dataset to find them all greater than the best results (overall accuracy 93.7%) of the rainbow classifier. The details of this experiment will be published shortly.

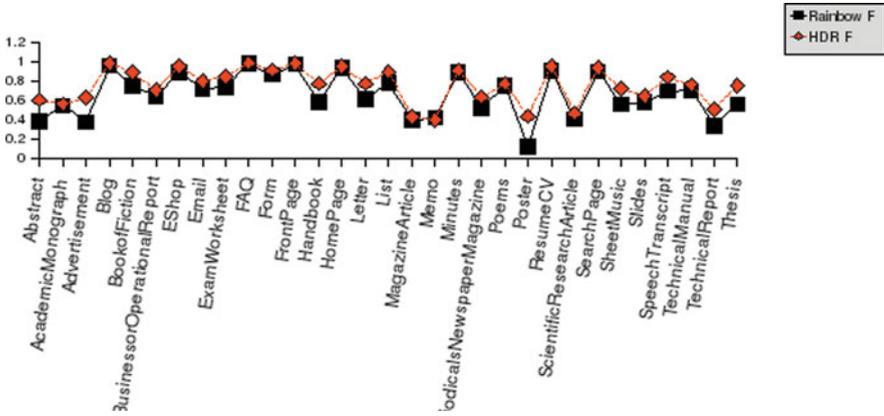


Fig. 6.8 F-measure: a comparison, SVM rainbow and SVM HDR

In the HDR of documents we have presented here, we have measured FP, LP and APR with respect to the length of the whole document. Just as performing discrete Fourier transform to obtain the harmonics of waves in signal processes involves sampling the signal, documents can also be examined at different resolutions by varying the range in which harmonic behaviour is examined (e.g. when examining the string “axbxcdefghijklmn”, and examining the occurrences of “x” throughout the string, it does not seem to exhibit harmonic behaviour but, if you select the first seven letters “axbxcxd”, it is perfectly harmonic). It is likely that shorter windows of examination will produce interesting comparisons.

### 6.7 Conclusions

The results of automated experiments described in this chapter provide evidence that the overall accuracy of the support vector machine rainbow text classifier is already comparable to that of an average human classifier in genre classification. Here we have shown that the SVM HDR, which uses the layout of words in the document, outperforms the SVM rainbow text classifier. This makes it a promising candidate for further study. In particular, a comparison of the SVM HDR classifier against classifiers other than SVM rainbow is required for fuller analysis. It would also be desirable to make direct comparisons of LP, FP and AP across genre classes.

The results with respect to the 7-webgenre collection suggest SVM HDR as a promising candidate for comparison to classifiers that rely on counts of terms or patterns. There have been reports of high accuracy levels of classification on the same dataset carried out by Kanaris and Stamatatos [13]. Although their numbers are similar to ours, it must be noted that the accuracy presented by them is from classifications of the set carried out in isolation while, the accuracy reported in this chapter is obtained from a classification of the seven webpage genres when accompanied by a classification of 24 additional document genres.

Previous text classification methods actively integrate mathematical methods in feature selection, statistical modelling and error analysis, but the concept we are trying to capture is still only described through examples in the domain. This leads to a semantic gap (especially with high-level concepts such as those represented by genre classes) not dissimilar to that encountered in image retrieval.

A more rigorous study of genre is required to reflect two considerations: first, we need to scope different communities for potentially useful genre classes that can support other applications and, second, we need to incorporate basic mathematical concepts into the actual description of the identified genres. Hence, future efforts in this field should not only study the implication of term distribution versus term frequency further by:

- examining the resolution mentioned at the end of Section 6.6.2;
- looking at, and comparing, other forms of symbols apart from words; and
- considering ways in which the two approaches might be integrated

but also include user studies of genres to identify the possible applications to guide genre classification work, and isolate base mathematical concepts that can be used to build the concepts gradually to describe higher-level concepts of genre.

## References

1. Bagdanov, A., and M. Worring. 2001. Fine-grained document genre classification using first order random graphs. In *Proceedings of the 2001 Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, 79–90. Seattle, WA. USA. <http://doi.ieeecomputersociety.org/10.1109/ICDAR.2001.953759>
2. Barbu, E., P. Heroux, S. Adam, and E. Turpin. 2005. Clustering document images using a bag of symbols representation. In *Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)*, 1216–1220. Seoul, Korea. <http://doi.ieeecomputersociety.org/10.1109/ICDAR.2001.953759>
3. Bekkerman, R., A. McCallum, and G. Huang. 2004. Automatic categorization of email into folders: Benchmark experiments on Enron and Sri corpora. Technical Report IR-418, Center for Intelligent Information Retrieval, UMASS. <http://www.cs.umass.edu/~homedirmccallum/papers/foldering-tr05.pdf>
4. Berninger, V.F., Y. Kim, and S. Ross. 2009. Building a document genre corpus: A profile of the KRY5 I corpus. In *Proceedings of Corpus Profiling Workshop with 'BCS-IRSG Workshop on Corpus Profiling'*. <http://www.bcs.org/server.php?show=conWebDoc.26115>
5. Biber, D. 1995. *Dimensions of register variation: a cross-linguistic comparison*. New York, NY: Cambridge University Press.
6. Bookstein, A., S.T. Klein, and T. Raita. 1998. Clumping properties of content-bearing words. *Journal of the American Society of Information Science* 49(2):102–114.
7. Dc-dot: UKOLN Dublin Core Metadata Editor (webpage last updated Aug 2000). <http://www.ukoln.ac.uk/metadata/dcdot/>
8. De Roeck, A., A. Sarkar, and P. Garthwaite. 2004. Frequent term distribution measures for dataset profiling. Technical Report 2004/2006, Faculty of Mathematics and Computing, Open University, Milton Keynes. <http://computing-reports.open.ac.uk/index.php/>
9. Dong, L., C. Watters, J. Duffy, and M. Shepherd. 2008. An examination of genre attributes for web page classification. In *Proceedings 41st Hawaiian International Conference*

- on System Sciences*. IEEE Computer Society Press, Waikoloa, Big Island, HI, USA. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=4438836>
10. Finn, A., and N. Kushmerick. 2006. Learning to classify documents according to genre. *Journal of American Society for Information Science and Technology* 57(11):1506–1518.
  11. Giuffrida, G., E. Shek, and J. Yang. 2000. Knowledge-based metadata extraction from PostScript Files. In *Proceedings 5th ACM International Conference on Digital Libraries*, 77–84. San Antonio, TX, USA. <http://portal.acm.org/citation.cfm?id=336597.336639>
  12. Han, H., L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E.A. Fox. 2003. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 77–84. Houston, TX, USA. <http://portal.acm.org/citation.cfm?id=827146>
  13. Kanaris, I., and E. Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings 19th IEEE International Conference on Tools with Artificial Intelligence*. Patras, GR. <http://portal.acm.org/citation.cfm?id=1337285>
  14. Karlgren, J., and D. Cutting. 1994. Recognizing text genres with simple metric using discriminant analysis. In *Proceedings 15th Conference on Computational Linguistics* 2:1071–1075. Kyoto, Japan.
  15. Ke, S.W., C. Bowerman, and M. Oakes. 2006. PERC: A personal email classifier. In *ECIR 2006* (London, UK), eds. M. Lalmas et al., LNCS 3936, 460–463, Heidelberg: Springer-Verlag, <http://www.springerlink.com/content/r27700t736786455/fulltext.pdf>
  16. Kessler, G., B. Nunberg, and H. Schuetze. 1997. Automatic detection of text genre. In *Proceedings 35th Annual Meeting ACL*, 32–38. Madrid, Spain.
  17. Kim, Y., and S. Ross. 2007a. Detecting family resemblance: Automated genre classification. *CODATA Data Science Journal* 6:S172–S183. ISSN: 1683–1470. [http://www.jstage.jst.go.jp/article/dsj/6/0/S172/\\_pdf](http://www.jstage.jst.go.jp/article/dsj/6/0/S172/_pdf)
  18. Kim, Y., and S. Ross. 2007b. Searching for ground truth: A stepping stone in automated genre classification. In *Digital libraries: R&D* (Tirrenia, Italy), eds. C. Thanos, F. Borri, and L. Candela, LNCS 4877, 248–261, Heidelberg: Springer-Verlag, <http://www.springerlink.com/content/lt760613m2731723>
  19. Manning, C., and H. Schutze. 1999. *Foundations of statistical language processing*. Cambridge, MA: MIT Press.
  20. McCallum, A. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/~verb1~11mccallum/bow>
  21. Rauber, A., and A. Müller-Kögler. 2001. Integrating automatic genre analysis into digital libraries. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. Roanoke, VA, USA. <http://doi.acm.org/10.1145/379437.379439>
  22. Rayson, P., A. Wilson, and G. Leech. 2002. Grammatical word class variation within the British National Corpus sampler. In *New frontiers of corpus research: Papers from the 21st International Conference on English Language Research on Computerized Corpora*, Sydney 2000, eds. P. Peters, P. Collins, and A. Smith, 295–306. Amsterdam: Rodopi.
  23. Ross, S., and M. Hedstrom. 2005. Preservation research and sustainable digital libraries. *International Journal of Digital Libraries* 5(4):317–325.
  24. Santini, M. 2007. Automatic identification of genre in web pages. PhD Thesis, University of Brighton, Brighton. [http://www.itri.brighton.ac.uk/~homedirMarina.Santini/MSantini\\\_PhD\\\_Thesis.zip](http://www.itri.brighton.ac.uk/~homedirMarina.Santini/MSantini\_PhD\_Thesis.zip)
  25. Thoma, G. 2001. Automating the production of bibliographic records. Technical report, Lister Hill National Center for Biomedical Communication, US National Library of Medicine. <http://archive.nlm.nih.gov/pubs/thoma/mars2001.php>
  26. Yang, Y., J. Zhang, and B. Kisiel. 2003. A scalability analysis of classifiers in text categorization. In *Proceedings 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 96–103. Toronto, ON, CA. <http://doi.acm.org/10.1145/860435.860455>
  27. Witten, H.I., and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann.



# Chapter 7

## In the Garden and in the Jungle

### Comparing Genres in the BNC and Internet

Serge Sharoff

#### 7.1 Introduction

The jungle metaphor is quite common in genre studies. The subtitle of David Lee's seminal paper on genre classification is "navigating a path through the BNC jungle" [16]. According to Adam Kilgarriff, the BNC is a jungle only when compared to smaller Brown-type corpora, while it looks more like an English garden when compared to the Web [15]. Intuitively this claim is plausible: if we consider the whole Web as a corpus, it probably contains a much greater variety of text types and genres than the 4,055 texts in the BNC classified into 70 genres. However, we still need to study this jungle.

Nowadays it is relatively easy to collect a large corpus from the Web, either using search engines [24] or web crawlers [13, 3], so it is easy to surpass the BNC in size. However, we know little about the domains and genres of texts in corpora collected in this way. Even if we collect domain-specific corpora [2] and can be sure that all texts in our corpus are about, e.g., epilepsy, we still do not know the amount of research papers, newspaper articles, webpages advising parents, tutorials for medical staff, etc. in it.

Traditional corpora have been annotated manually, which did not create a significant overhead: such corpora have been also compiled manually, so it was possible to annotate each text according to a reasonable number of parameters. Even then there can be problems with manual classification. Spoken texts in the BNC are not classified into their domains at all, even though many of them are devoted to a well-defined topic, like computing, medicine or politics. Similarly, a single large text taken from a newspaper and classified as "world affairs" in the BNC can contain home and foreign news, commentaries, gossips, etc. Many genres also remain underdescribed. Even though there are textbooks in the BNC (for instance, texts

---

S. Sharoff (✉)

Centre for Translation Studies, University of Leeds, LS2 9JT Leeds, UK  
e-mail: s.sharoff@leeds.ac.uk

EVW or GVS<sup>1</sup>), their presence is not registered in the classification scheme: they are classified as written academic texts (according to David Lee's genre classification) or as books for professional readers in the respective domains of natural sciences and arts (according to the original BNC database), but nothing in the scheme indicates that they are teaching materials. However, such complaints are only minor quibbles if we compare this situation to the sheer lack of information about even very basic characteristics of Web corpora, such as I-EN [24], SPIRIT [13] or deWaC [3].

The task of classifying Web corpora and comparing their composition to traditional corpora is difficult for several reasons. First, no established classification of genres exists, even for traditional written texts. Practically every study uses its own list of genres, e.g., compare the 15 classes in the Brown Corpus to the 70 genres in David Lee's classification of the BNC to the 120 genre labels in the Russian National Corpus (RNC). Second, the relationship between traditional genres and genres existing on the Web is not clear. Some web genres can be compared to traditional printed media, e.g., on-line newspapers, while others are markedly different from any known printed counterpart, e.g., chat rooms. Third, given the large number of pages in Web-derived corpora, e.g., more than 60,000 in I-EN [24], we need automatic methods that can identify genres reliably and be applicable to an arbitrary webpage. The fourth problem concerns the very design of the genre inventory. If the goal is to classify every text existing on the Web, the number of genres is too large to be listed in a flat list. Only within the genres of academic communication we can come across research articles (with different genre conventions applicable to the humanities, engineering or natural sciences), as well as popular articles, reviews, books, calls for participation, emails, mailing lists and forums, project proposals, progress reports, minutes of meetings, job descriptions, etc. A recent overview of traditional genre labels refers to a list of "more than 4,500" categories [21]. The fifth problem concerns "emerging" genres: new technologies can offer new avenues for communication, which readily produce new genres, e.g., blogs, personal homepages or spam. However, we can expect greater stability of underlying communicative intentions, which are realised in new forms using new technologies. For instance, if our list of webgenres includes a simple entry for blogs, this category cannot be compared to anything in the BNC (blogs did not exist at the time of its compilation), whereas the function of blogs is similar to that of diaries or opinion columns in newspapers, while it is different from them in the audience size, distribution mode and authorship.

One way of studying genres on the Web is to start with a genre (or a group of genres), such as blogs [17] or conference websites [19]. Then we can analyse linguistic features specific to this genre and learn how to identify a text as belonging or not belonging to it. Another way of studying web-genres is aimed at saying "sensible and useful things about any text" that exists on the Web.<sup>2</sup> Such studies can offer a

---

<sup>1</sup> Throughout this chapter I refer to BNC texts using their ids from the BNC Index, which is available from <http://clix.to/davidlee00>

<sup>2</sup> The quote refers to the purposes Michael Halliday intended for his "Introduction to Functional Grammar" [11].

very superficial description for many genres studied in the first approach, but if we do not use a compact text typology, this study risks ending with an infinite list of genre types to account for all possible webpages.

In this chapter I will follow the latter research path by outlining an approach to text classification that can be used to describe the majority of texts on the Web using a small number of categories (less than 10), so that we can broadly assess the composition of genres in a Web-derived corpus, compare it against any other collections of webpages and traditional corpora, as well as against corpora in other languages (Section 7.2). Then (in Section 7.3) I will present an experiment for detection of text categories in traditional reference corpora against English and Russian Internet corpora. Traditional corpora used in this study are the BNC and Russian National Corpus (RNC), which is comparable to the BNC in its size and composition [23]. Finally, in Section 7.4 I will discuss the similarities and differences between these Internet corpora and their manually collected counterparts.

The study concerns English and Russian corpora collected from the Web using random queries to search engines [24]. Below these corpora are referred to as “the Internet corpora” (or I-EN and I-RU more specifically). However, there is nothing in the methodology specific to this method of corpus collection, so the study should be applicable to any sufficiently large corpus of webpages (in the discussion below I refer to such corpora as “Web-derived corpora”). In the last section, I also report on a small experiment of applying the same methodology to classifying ukWac, another English corpus collected by crawling websites in the .uk domain [10].

## 7.2 Text Typology for the Web

Approaches to classifying texts into genres can be grouped into two main classes. The first class identifies genres of documents on the basis of what can be called “look’n’feel” properties, e.g., FAQ, forum or recipe, while the second class detects broad functional classes, e.g., description or argumentation, cf. the discussion in [16] or [5].

Look’n’feel approaches are based on traditional labels, so they reflect the practice of their use and it is relatively easy to annotate a significant amount of texts manually by human annotators without extensive training. For instance, if a page looks like a blog, applying this label is not difficult for anyone familiar with this genre. If we use a folksonomy-based genre typology in a search engine, again its users can recognise labels easily, for instance, to refine the results of their search. At the same time, this approach assumes an established genre inventory, which does not exist (Problem 1 identified above), it results in proliferation of categories (Problem 4), and it is not flexible enough to allow comparison of webcorpora to their traditional counterparts (Problem 5).

This is the reason for taking the functional approach to genre classification in this project. However, even if we narrow our search of a suitable genre classification scheme down to functional studies, which classify texts from the viewpoint of the function they fulfill in the society, we still find a large number of options. Marina

Santini mentions such classes as Descriptive-narrative, Explicatory-informational, Argumentative-persuasive and Instructional identified in traditional text typology studies along with the several variations of this inventory, e.g., separating descriptive and narrative texts [22, Chapter 2]. Without giving an explicit text typology, James Martin defines genres as the results of “staged, goal-oriented, purposeful activity in which speakers engage as members of our culture” [18, p. 25]. In [14], genres are also defined functionally, but using traditional labels taken from reflective practice, e.g., “editorial is a shortish prose argument expressing an opinion on some matter of immediate public concern”. In another study of genre detection [7], the classification is done into five functional styles: fiction, journalism, official, academic, everyday language, following a tradition that stems from Jakobson [12].

The functional approaches mentioned above are still not precise enough for the goal of unambiguous classification of the majority of webpages. The classification scheme that gave the initial impetus to research presented in this chapter was proposed by John Sinclair, first in the context of the EAGLES guidelines [9, 26]. Among other dimensions of text classification Sinclair referred to the following six “intended outcomes of text production”:

1. information – reference compendia (Sinclair adds the following comment “an unlikely outcome, because texts are very rarely created merely for this purpose”);
2. discussion – polemic, position statements, argument;
3. recommendation – reports, advice, legal and regulatory documents;
4. recreation – fiction and non-fiction (biography, autobiography, etc.)
5. religion – holy books, prayer books, Order of Service (this label does not refer to religion as a topic);
6. instruction – academic works, textbooks, practical books.

The typology is compact and applicable to webpages: only six top-level categories, each of which represents a variety of webpages, e.g., a page from Wikipedia is aimed at informing, a forum – at discussing, etc.

However, an attempt to apply these classes to the Web without any modification results in several problems. First, the boundary between look’n’feel and communicative intentions is fuzzy. What is the reason for classifying a text as “recommendation”? Is this because it recommends an action or because it is classified as a report? A proposal issued by a think-tank of a political party can have “report” in its title, but in terms of its function it is very similar to a position statement published in a newspaper. The title of a publication is not the only reason for classifying it functionally, but in [9] no basis is given for classifying intentions.

Second, a functional classification assumes a certain degree of correlation between the function of a text and the language used to express this function. The function is *not defined* by linguistic features of respective texts, as otherwise the definition of genres depends on accidental features we choose to represent the genre, whereas its function in the society should be immune to such superficial variation. For instance, if narrative texts are defined by the number of past tense verbs [6], then narrative texts do not exist in Chinese, in which verbs do not have tenses. There might be a correlation between Chinese narrative texts and the amount of aspectual

particles (e.g., *le*, *zhe*) or temporal adverbs (e.g., *zuotian*), but the dimension of narrativity has to be defined without relying on the features of an individual corpus. In other words, categories, such as narration, have to be specified taking into account the function a text has in the society, so that any comparison between corpora is made on the basis of categories more stable than linguistic features.<sup>3</sup> Nevertheless, it is reasonable to expect that texts contained in a single class of communicative aims (or “outcomes”) are more or less similar, e.g., narrative texts can be defined as texts reporting a sequence of events, and this correlates with certain linguistic features, which can be language- or even corpus-specific. On the other hand, if there is no similarity between regulatory documents and adverts (the latter are considered as a subclass of advice in Sinclair’s classification), there can be fewer reasons to keep them in the same class of “recommendations”. The same applies to joining academic works (such as the present chapter) and practical books (such as recipes) in the same category of “instructions”.

Third, decisions on document categorisation from their look’n’feel can be made by any reasonably confident user of those texts, while much more training is needed to recognise more abstract functional categories. For instance, it is reasonable for the purposes of genre analysis to distinguish between blog entries aimed at discussion, news dissemination or recreation (entries with poetry or fiction), but naive annotators (much less ordinary Web-users) cannot make such distinctions reliably. In an experiment on webpage cleaning [4], we attempted to annotate two sets of 60 webpages each in Chinese and English using a functional set of categories derived from Sinclair (the categories were advert, academic discussions, non-academic discussions, information, interview, instruction, fiction, news). Each page was annotated by two translation students who were familiar with classification of texts by their function and were given training to recognise the categories from this list. Nevertheless, the students failed to produce appropriate classification labels for some texts. Often both decisions made by the two annotators of the same text were different from the principles used in the typology suggested to them. For instance, a diary-like blog entry (<http://blogs.bootsnall.com/michelle/archives/006670.shtml>) was classified by one student as “information”, by another one as “news”, while it should have been classified as “non-academic discussions” along with all other private blog entries, if the instructions given as the basis for document categorisation were followed. This experiment suggests a gap between genre theory and the actual practice of average users.

Finally, some texts can be inherently ambiguous with respect to categories from Sinclair’s list. For instance, academic works are typically aimed at discussing states of affairs and making position statements; the boundary between “recommendation” and “discussion” is also frequently fuzzy. The same argument applies to traditional

---

<sup>3</sup> This example assumes that the function of narration is actively used in the respective societies for approximately the same purposes, but for modern corpora this can be taken for granted.

rhetorical categories as well: the classes of descriptive, explicatory and argumentative texts often overlap.

These considerations have led to the following adaptation of the original Sinclair's typology:

1. *discussion* – all texts expressing positions and discussing a state of affairs
2. *information* – catalogues, glossaries, other lists (mostly containing incomplete sentences)
3. *instruction* – how-tos, FAQs, tutorials
4. *propaganda* – adverts, political pamphlets
5. *recreation* – fiction and popular lore
6. *regulations* – laws, small print, rules
7. *reporting* – newswires and informative broadcasts, police reports

The present study is based on this typology, but I would refrain from saying that this is the final version. The category of *discussions* might need splitting, as it comprises academic works and popular science, discussion forums and cases for support of academic projects, columns in newspapers and personal diaries, and so on. The difference between them can be described using other parameters of corpus classification, such as the audience (professional or layman), publication medium (newspapers, forums, blogs), authorship (e.g., single or corporate). A multidimensional classification of this sort is more complex than a flat list of microgenres. However, the reason for this complexity is that many microgenres actually contain diverse text types. For instance, the category of blogs (frequently studied as a microgenre) does not define its functional content. Blogs are often studied from what is retrieved from a blogging website, like [blogspot.com](http://blogspot.com), which by itself only provides a tool that can help in publishing a chronologically ordered sequence of (short) texts. The genre is defined by the way this tool is used, e.g., to post newsitems, publish fiction, discuss academic topics, or maintain personal diaries (with the two latter examples considered to be prototypical blogs). At the same time, a text can be published in a variety of possible publication media. For instance, a recipe (“instruction”) can be published in a blog entry, forum, newspaper or book.

Since the typology is meant to allow corpus comparison within and between languages, it should be complete: any webpage has to be classified according to a fixed number of predefined categories. Otherwise, it is difficult to compare corpora classified using different schemes. The functional principles for designing a typology mean that it is robust with respect to new emerging genres, as long as new communicative intentions do not emerge with new genres.

In designing a genre typology one open question is whether the typology is specific to an individual corpus, language or culture. Do we expect to use another typology to work with a corpus collected using different tools? Does the typology of English webpages apply to German, Russian or Chinese ones? The version proposed above corresponds to the mildest case of a culture-specific typology. It assumes that we derive the values of categories empirically from text categories which are more frequent in on the Web (across languages we are working with), also taking into account the typology used in traditional reference corpora. “Mild” cultural

dependence of the proposed typology means that it is specific to the current generation of Web-derived corpora for languages with well-developed Internet culture. The typology listed above was developed from my attempts to classify English, German, Russian and Chinese webpages in my Internet corpora [24]. Most probably, it can be applied to describing the majority of modern webpages in, say, Arabic or Tagalog, while it may lack categories important for describing many texts written in the eighteenth century or in languages without an existing Internet culture like Brahui or Yukaghir, which might use the Web for purposes different from major languages.

Another open question concerns the ambiguity. One of the aims of the typology presented above is to reduce the ambiguity in comparison to the original Sinclair's classification, e.g., by splitting recommendation or adding a new category of reporting. However, the ambiguity is wide-spread in real texts. This also concerns their communicative aims, so we can consider the possibility of using multiple labels, but the results of comparing two corpora with multiple labels are more difficult to interpret numerically. Therefore, in the study below each document gets a single label.

### 7.3 An Experiment in Automatic Classification of the Web

Once we have a typology, the next task is to classify I-EN and I-RU automatically and to compare their composition against traditional corpora (BNC and RNC respectively). A by-product of this study is the validation of the typology by checking whether its categories can be detected reliably and what confusion arises. One problem in this analysis is that supervised machine learning needs a large number of training examples, which are difficult to obtain from unclassified Web-derived corpora. Also, a comparison of I-EN and I-RU to their traditional counterparts implies classification of traditional corpora according to the same set of categories, while each corpus is documented using its own classification schemes.

Some genre labels used in BNC and RNC can be mapped to the more general functional categories listed above. For instance, academic (*W\_ac\_.\**) and non-academic (*W\_nonac\_.\**) papers from the BNC can be treated as "discussions", fiction and popular biographies as "recreational" texts, "propaganda" in the BNC is represented by *W\_advert*. Not all genre labels can be mapped unambiguously, e.g., *W\_commerce* or *W\_email*. In addition to this, newspaper files in the BNC frequently consist of an entire issue and they contain a combination of genres, so they cannot be used for training purposes. Thus, the training corpus is a subset of the BNC.

This unambiguous mapping results in a "crisp" training corpus, which consists of texts definitely within the boundaries of the respective categories. For instance, we can populate the "instructions" category with texts marked as *W\_instructional* in the BNC, 15 texts in total, such as recipe books, software manuals or DIY magazines. A clearer separation between text types is beneficial for the accuracy of cross-validation using the training corpus, but this eliminates other members

of this category, which do not have unambiguous labels in the BNC, e.g., textbooks or academic tutorials. If we apply the model trained on a “crisp” corpus to the rest of the BNC, there is little chance that such texts will be recognised as “instructions”. On the other hand, including texts not explicitly labelled as such in the BNC, e.g., texts having “textbook” in their title or keywords, results in a “fuzzy” training corpus, which has a better coverage for each individual category, but contains more ambiguity, which might adversely affect the accuracy of the classifier.

The second problem with crisp corpora is that some BNC genre categories are easier to convert to corresponding communicative aims than others, so the training corpus can get significantly more discussions and recreational texts than other text types, e.g., 514 text can be classified as “recreation” vs. only 15 as “instruction”. The lack of balance can cause problems to machine learning algorithms, which pay attention to the probability of a category in the training corpus. In the end for instructions and reporting categories I produced two versions, one was “crisp”, including, respectively, only `W_instructional` and `W_newsscript` texts. The other one was “fuzzy”, also including texts containing the word `textbook` in the title or keywords and `W.*_reportage` in its genre definition or news in the keywords. At the same time, the number of more frequent categories in the “fuzzy” corpus was reduced by random selection. Also, neither of the two corpora contains the category of “information”, as such texts (e.g., dictionaries or catalogue descriptions) have not been included in the BNC at all.

These subsets from traditional corpora were used to train SVM classifiers using the default parameters of Weka’s implementation of SVM [28]. Then, the models trained on a portion of traditional corpora were applied to the whole set. The features used for training were based on the frequency of POS trigrams describing individual texts, and also on the frequency of punctuation marks, e.g., quotes, exclamation and question marks each contributed to a feature. Given that the number of possible POS trigrams is fairly large resulting in a very sparse feature set, the study used the most significant POS trigrams selected using the Information Gain method, resulting in 593 features for English and 577 features for Russian (the accuracy on a subset actually improves by a few percentage points in comparison to the full feature set and the resulting model is much faster).

In principle, web-related parameters can be additionally used to describe webpages, such as the properties of originating URLs (e.g., the presence of `cgi-bin` or `~`), HTML tags (the use of fonts, tables or Javascript), navigation (links to other pages or links within a page), cf. [1, 20]. However, some information (such as HTML tags) has been lost in the process of corpus creation, and, more importantly, the chosen combination of POS trigrams with punctuation marks is applicable to both traditional written texts and webpages.

Table 7.1 compares the result of training using a “crisp” corpus against a “fuzzy” corpus. The accuracy is defined in Weka as the number of correctly classified instances (true positives) in the test corpus divided by its total size (averaged after 10-fold cross-validation). As we can see the overall accuracy can be very high (up to 97% with the crisp corpus), but this goes at the expense of the accuracy of assigning

**Table 7.1** Comparing confusion matrices in training corpora

Crisp BNC corpus (accuracy: 97%)							Fuzzy BNC corpus (accuracy: 86%)						
a	b	c	d	e	f	← Classified as	a	b	c	d	e	f	← Classified as
194	1	6	6	1	0	a = Discussion	244	26	2	4	0	13	a = Discussion
0	14	1	0	0	0	b = Instruction	19	49	3	4	1	0	b = Instruction
5	1	47	1	0	0	c = Propaganda	10	3	46	1	0	0	c = Propaganda
5	0	0	507	1	0	d = Recreation	3	1	0	194	0	1	d = Recreation
0	1	0	0	76	0	e = Regulation	2	0	0	0	78	0	e = Regulation
2	0	0	0	0	20	f = Reporting	14	0	0	0	0	29	f = Reporting

TP rate	FP rate	Precision	Recall	F-measure	Class
0.869	0.102	0.843	0.869	0.856	Discussion
0.623	0.046	0.608	0.623	0.615	Instruction
0.783	0.010	0.870	0.783	0.825	Propaganda
0.975	0.016	0.956	0.975	0.965	Recreation
0.950	0.001	0.987	0.950	0.968	Regulation
0.605	0.016	0.703	0.605	0.650	Reporting
0.800	0.030	0.830	0.800	0.810	<i>Average</i>

Fuzzy BNC, detailed accuracy by class

a	b	c	d	e	f	← Classified as
721	2	89	66	32	55	a = Discussion
41	17	14	4	12	2	b = Instruction
176	8	394	3	33	13	c = Propaganda
51	2	2	890	0	4	d = Recreation
55	12	45	0	339	19	e = Regulation
101	3	23	18	23	183	f = Reporting

Russian fuzzy training corpus (accuracy: 74%)

categories to examples outside clear-cut categories, when the classifier is applied to the rest of the BNC.<sup>4</sup> For instance, text A60, an introduction to international marketing, classified as *W\_commerce* in the BNC, is classified as “regulation” using the crisp training corpus, while it gets reclassified as “instruction” using the “fuzzy” one. This text does include formally written sentences that make it look like a piece of regulation (*International marketing is treated as a generic term covering the distinctions made in describing marketing activities as “international” or “multi-national” or “global”*), but the text as a whole is a textbook from the Kingston Business School. As a result, the crisp classifier treats only 86 texts in the whole BNC as “instructions”, while the fuzzy one finds 829 texts in this category, including A06 (a guide to becoming an actor), A0M (a karate handbook), A17 (a dog care magazine), none of which is treated as an instructional text in the BNC classification. Out of a random sample of 20 BNC texts automatically classified as “instructions”, only three texts should not belong to this category: C8X (poetry), KBS (a recorded dialogue) and KM4 (a recording from a business meeting). The results reported below are based on fuzzy training corpora.

<sup>4</sup> A similar pattern is evident in the accuracy drop from about 90% in the “crisp” 7-webgenre corpus to 66% in a fuzzy KI-04 corpus in experiments described in [22].

For English the procedure achieved the accuracy of 86% with 10-fold cross-validation, while the accuracy for Russian is significantly lower (74%), which can possibly be explained by the free word order, as well as by the greater number of morphological categories. For instance, the tagset used for English contains just four categories for nouns (common vs. proper, singular vs. plural), while in Russian nouns are described in terms of their number, gender, case, animacy, generating 92 categories actually occurring in the training corpus. These factors make POS trigram statistics sparser, especially on the RNC texts, which are generally shorter than their BNC counterparts. At the same time, the greater granularity of POS categories can help in distinguishing between genres. For instance, imperatives are a good indicator of instructions and propaganda, but in the English tagset such uses are treated identically to other base forms (infinitives and present simple forms). The same problem occurs with modal verbs: even if their functions are different and some modals are characteristic for specific genres (e.g., *shall* vs. *must*), in POS trigrams they are represented by a single tag.

Finally, the jungle of the Web was treated as being similar to the English garden, i.e., the models trained on the BNC and RNC were applied to English and Russian texts from the Internet corpora. First, the BNC and RNC models were applied to randomly selected subsets of 250 webpages from, respectively, I-EN and I-RU. The accuracy dropped considerably (down to 52% for English, 63% for Russian), but this gave the basis for creating a manually corrected training set to classify the entire Internet corpus. The drop in accuracy can be attributed to three factors<sup>5</sup>:

- the balance of genres even in the fuzzy training corpus is quite different from what we have in the testing corpus: some classes are under-represented (reporting), others are over-represented (fiction) or not represented in traditional corpora at all (information).
- the Internet corpora are dirty in the sense that they contain some elements from original webpages not presented in the traditional corpora, such as navigation frames, ASCII art, standard headers. In spite the best efforts to remove this noise, the accuracy of automatic cleaning is below 75% [4].
- the language of the Internet is to some extent different from the language used in traditional corpora, e.g., not only British English is included in the annotated genre sample, FAQs are organised differently from tutorials listed in the BNC, the core of BNC texts stems from 1980s (the accuracy on the Russian sample was higher because the RNC is based on more recent texts, while I-RU is much more homogeneous in terms of the dialects it contains).

---

<sup>5</sup> The BNC has been retagged with TreeTagger, the same tool used for tagging I-EN, so there was no difference in the tagset and tagging between the two corpora (this could have caused variations in accuracy otherwise).

## 7.4 Analysis of Results

The results of the automatic assessment of the composition of traditional and Internet corpora are presented in Table 7.2. The composition of the entire BNC and RNC was assessed by applying classifiers trained on their fuzzy subsets to their full content (BNC/F and RNC/F columns). I-EN and I-RU were assessed by their manually classified subsets of 250 texts each (I-EN/S and I-RU/S columns), and by applying classifiers trained on these subsets to their full content (I-EN/F and I-RU/F). Finally, the composition of ukWac, another corpus of English collected by crawling websites in the .uk domain, was also assessed by the same method (ukWac/F). To avoid data sparsity for classifiers, only texts longer than 300 words were used (this covers almost all texts in the BNC and more than 80% of I-EN and I-RU, 63% of ukWac).

**Table 7.2** Automatic assessment of corpus composition

Categories	BNC/F (%)	I-EN/S (%)	I-EN/F (%)	ukWac/F (%)	RNC/F (%)	I-RU/S (%)	I-RU/F (%)
Discussion	37.42	37.20	52.49	38.21	62.99	44.00	55.12
Information	0.00	6.00	4.03	5.03	0.00	0.40	0.06
Instruction	26.66	23.20	20.51	18.77	0.99	12.40	6.88
Propaganda	5.45	12.00	11.24	15.66	11.69	4.80	0.17
Recreation	21.43	4.00	0.97	1.03	14.17	24.80	27.46
Regulation	3.05	6.40	2.21	3.03	4.93	0.40	0.07
Reporting	6.00	11.20	8.54	18.27	5.22	13.20	10.24

### 7.4.1 Qualitative Assessment of Texts in Each Category

#### 7.4.1.1 Discussion

This is the biggest category with a variety of subtypes. Automatic classifiers in general tended to overestimate the membership for this category, i.e., /F columns list more members than corresponding /S columns (especially for Russian). Texts classified in this way mostly include academic and newspaper articles (texts written for the professional audience vs. for the general public), as well as discussion forums and archived mailing lists.

#### 7.4.1.2 Information

This macrogenre was not well represented in traditional corpora, such as the BNC and RNC, since corpus compilers tend to select running texts rather than catalogues or dictionaries. The procedure for collecting I-EN and I-RU also favoured running texts against incomplete descriptions by constructing longer queries, cf. [24, Section 2.2]. However, this macrogenre is common on the web. Pages classified as information include lists of people, places, businesses, objects, news stories, etc.

A fair amount of such texts (amounting to 15) managed to get into the random sample for English, even though fewer texts of this sort were detected in the full content of I-EN. There was only one text of this type in the Russian sample, which was not enough for training reliable classifiers. On the other hand, this macrogenre is more common in ukWac (which was produced by crawling, not by querying search engines). Texts of this type are important not only because of their amount, but also because of their potential to mislead POS taggers or other NLP tools. They often contain incomplete sentences with the visual boundary between their chunks often lost in the process of creating a plain text corpus.

#### 7.4.1.3 Instruction

The majority of texts classified with this label belong to two types:

- structured lists, such as FAQs, recipes, steps for assembling, repairing or maintaining something;
- advice written in a more narrative style, such as a recommendations, tutorials, as well as some research papers, e.g., [http://www.privcom.gc.ca/media/nrc/opinion\\_021122\\_lf\\_e.asp](http://www.privcom.gc.ca/media/nrc/opinion_021122_lf_e.asp)

Such texts constitute about one quarter of either I-EN or ukWac, making it the second most frequent text type. However, it is found to be much less common in I-RU, though it is less common in the RNC as well. One possible reason for the apparent scarcity of such texts (they do constitute 12% of the sample from the Web) is the greater difficulty of detecting them in Russian. According to the Russian confusion matrix in Table 7.1, the majority of texts classified as “instruction” in the training set were classified as “discussion” by the automatic classifier. More research is needed to find features that can detect this class in Russian reliably.

#### 7.4.1.4 Propaganda

The amount of texts with propaganda of various sorts is in the range of 11% in I-EN to 16% in ukWac, while it is much less common in the BNC (5.5%). Pages classified as propaganda typically promote goods and services, e.g., <http://www.hawaii-relocation.com/>, which is not strictly speaking spam; this speaks against the reputation of spam as the main polluter of Web-derived data.

#### 7.4.1.5 Recreation

It is known from other studies [24] that texts written with the purpose of recreation, such as fiction, are rare on the English Web (because of copyright restrictions), while they are quite frequent for Russian. The present experiment confirms this to a certain extent. Nevertheless, such texts do exist in the two English Internet corpora. The most common microgenres are science fiction (often published under a Creative Commons license), collections of jokes (without explicit authorship), as well as all sorts of out-of-copyright fiction. The

automatic classifier is also quite generous in assigning this category to texts, e.g., [http://42.blogs.warnock.me.uk/2006/05/cycling\\_fame.html](http://42.blogs.warnock.me.uk/2006/05/cycling_fame.html), that describes an event and is written in a chatty style (descriptions of events are normally classified as “reporting” otherwise). Anyway, one can argue that it is reasonable to classify texts of this sort as aimed at recreational reading.

#### **7.4.1.6 Regulation**

Texts classified in this way correspond to various rules, laws or official agreements, e.g., <http://contracts.onecle.com/talk/walsh.nso.2000.08.07.shtml>. According to the confusion matrix in Table 7.1 their detection in English is easy for the SVM classifier, so the figure for English in Table 7.2 can be assumed to be reliable. As for the Russian corpus, there was only one text of this type in the manually annotated sample, hence the classifier cannot be trained reliably. As a result there are numerous texts in I-RU automatically classified as “discussion”, while they can be reasonably treated as regulatory documents, e.g., <http://www.dmpmos.ru/law.asp?id=30020>.

#### **7.4.1.7 Reporting**

This category looks pretty uncontroversial. The original idea was to apply it to any type of newswires or reports about an event. Hence, the original classifier was trained on news scripts and reportage texts from the BNC (given the absence of police reports there). However, its application to webpages has identified other texts that can be reasonably treated as “reporting”, such as CVs, timelines of historic events or factual travel guides.

### ***7.4.2 Assessing the Composition of ukWac***

In this study I did not have time to evaluate the accuracy of genre assessment in ukWac on the basis of a large sample (around 250 documents). However, an initial estimate on transferring the classifiers trained on an I-EN sample to a new corpus can be made. Table 7.3 lists genres automatically assigned to documents collected from one website devoted to a large international conference. The results of classification in all cases seem to be reasonable. For instance, the rules for taking part in a competition are treated as “instruction”, texts about exhibitors, sponsors and possibilities for advertising are treated as “propaganda”, while the conference programme has been classified as “reporting”.

However, several pages reasonably belonging to the same category are classified differently. Three issues of the newsletter are classified as “propaganda”, while the fourth one – as “discussion”. Out of the seven CVs of conference speakers (the last one combines CVs of several panelists), three are treated as “reporting”, while the other four – as “discussion”. There are inherent reasons for the differences in their automatic classification. The first three newsletters promoted the conference or its sponsors, while the last one mostly consisted of an informative interview. The CVs

**Table 7.3** Assessing genres in ukWac

<a href="http://06.economie.co.uk/comp/rules.htm">http://06.economie.co.uk/comp/rules.htm</a>	Instruction
<a href="http://06.economie.co.uk/exhibitors/index.htm">http://06.economie.co.uk/exhibitors/index.htm</a>	Propaganda
<a href="http://06.economie.co.uk/location.htm">http://06.economie.co.uk/location.htm</a>	Discussion
<a href="http://06.economie.co.uk/newsletters/april2006.htm">http://06.economie.co.uk/newsletters/april2006.htm</a>	Propaganda
<a href="http://06.economie.co.uk/newsletters/aug1506.htm">http://06.economie.co.uk/newsletters/aug1506.htm</a>	Propaganda
<a href="http://06.economie.co.uk/newsletters/aug2806.htm">http://06.economie.co.uk/newsletters/aug2806.htm</a>	Propaganda
<a href="http://06.economie.co.uk/newsletters/may2006.htm">http://06.economie.co.uk/newsletters/may2006.htm</a>	Discussion
<a href="http://06.economie.co.uk/prog.htm">http://06.economie.co.uk/prog.htm</a>	Reporting
<a href="http://06.economie.co.uk/quiz.htm">http://06.economie.co.uk/quiz.htm</a>	Instruction
<a href="http://06.economie.co.uk/speakers/amy_domini.htm">http://06.economie.co.uk/speakers/amy_domini.htm</a>	Discussion
<a href="http://06.economie.co.uk/speakers/brian_spence.htm">http://06.economie.co.uk/speakers/brian_spence.htm</a>	Reporting
<a href="http://06.economie.co.uk/speakers/colin_baines.htm">http://06.economie.co.uk/speakers/colin_baines.htm</a>	Discussion
<a href="http://06.economie.co.uk/speakers/deborah_doane.htm">http://06.economie.co.uk/speakers/deborah_doane.htm</a>	Discussion
<a href="http://06.economie.co.uk/speakers/john_renesch.htm">http://06.economie.co.uk/speakers/john_renesch.htm</a>	Discussion
<a href="http://06.economie.co.uk/speakers/noreena_hertz.htm">http://06.economie.co.uk/speakers/noreena_hertz.htm</a>	Reporting
<a href="http://06.economie.co.uk/speakers/openforum.htm">http://06.economie.co.uk/speakers/openforum.htm</a>	Reporting
<a href="http://06.economie.co.uk/spons/additional.htm">http://06.economie.co.uk/spons/additional.htm</a>	Propaganda
<a href="http://06.economie.co.uk/spons/bursary.htm">http://06.economie.co.uk/spons/bursary.htm</a>	Propaganda
<a href="http://06.economie.co.uk/spons/index.htm">http://06.economie.co.uk/spons/index.htm</a>	Propaganda
<a href="http://06.economie.co.uk/spons/major.htm">http://06.economie.co.uk/spons/major.htm</a>	Propaganda
<a href="http://06.economie.co.uk/spons/opportunities.htm">http://06.economie.co.uk/spons/opportunities.htm</a>	Propaganda

in question were written in two different styles. One style describes the history of appointments (*Mike Kelly is Head of KPMG UK's Corporate Social Responsibility function. In 2002, Mike led KPMG's review of Environmental Risk Management at Morgan Stanley. Prior to coming to KPMG he was . . .*), while the other one emphasises the viewpoint of a person (*Variously described as a "business visionary" and as "a beacon lighting the way to a new paradigm", John Renesch stimulates people to think differently about work, leadership and the future. He believes that . . .*). The difference between these styles is obvious, but the decision made in each case is in the eye of the annotator (or automatic classifier), as views of the first person are described in his CV, even if they are less prominent than his function, while biographical details are also present in the second CV. The same argument applies to the difference between discussion and propaganda in the newsletters: the interview is informative, but it still promotes the company of the individual giving the interview.

## 7.5 Conclusions and Future Research

This chapter reports the first study, which was aimed at uncovering the genre composition of the entire jungle of the Web. The typology useful for classifying the entirety of webpages is still fluid. The main point of this study is to show that it is possible to estimate the composition of a corpus collected from the Web, even if it is a large corpus like I-EN (160 million words) or ukWac (2 billion words).

In short the proposed procedure looks like this:

1. take a corpus with known composition (source corpus);
2. train a classifier on a subset;
3. apply it to a sample of a corpus with unknown composition (target corpus);
4. correct the sample and train a new classifier;
5. apply the new classifier to the rest of the corpus.

If the system of genres used to describe the source corpus is identical to the genres needed to assess the target corpus, the whole source corpus can be used in Step 2. In another experiment, I classified I-EN and ukWac using the entire set of 70 genres of the BNC and four main genre categories of the Brown corpus (press, fiction, nonfiction and misc), following the results reported in [25]. This gives us data for comparing genre composition of a variety of corpora or for selecting subsets to study them more closely. For instance, 18,715 webpages in ukWac have been classified as *personal\_letters* using the BNC-trained classifier, with the vast majority of them being diary entries coming from blogs. So this classifier provides a useful mechanism for finding and studying diary-like blogs. However, the value of such tests is limited, as the experiments with the BNC and RNC (Section 7.3) show that the process of retraining using a subset of the target corpus (Steps 3 and 4) is necessary to improve the accuracy of the classifier on data from the target corpus.

Even the results for the validated classifiers have to be taken *cum grano salis*. It is tempting to refer to the results in Table 7.2 as saying that the composition of the Web is as follows: instructions – one quarter, advertising and propaganda – 10–15%, lists and catalogues – 5%, regulations – about 3%, etc. However, there are obvious limitations on extrapolating this study. First, the results are based on I-EN and ukWac, Web-derived corpora collected in a particular way. Both corpora contain only HTML pages (PDF files or Word documents were not used); the procedure for their collection favoured finding examples of running text at the expense of “index” pages or other collections of links (even though the methods for rejecting such pages were specific to each corpus), duplicate webpages in both corpora were discarded. Other methods of corpus collection might favour other slices of the Web and get different results.

Second, my training corpora used in Step 4 consisted of 250 webpages. This led to a limited number of training examples for less frequent categories. For instance, the Russian training sample contained just one example of texts classified as “information” and “regulation”, respectively. This is indicative of the fact that these text types are not very frequent in the rest of I-RU, see the discussion of sampling statistics in [24], but single examples do not give sufficient information for classifying unseen texts of this type. Some other macrogenres have more training examples, but they are still represented by a small number of microgenres. For instance, out of 16 texts classified as “regulation” in the English sample, there was no text belonging to the microgenre of “contractual agreements”, e.g., *Either party shall be entitled on written notice to terminate . . .* Thus, texts of this type from the full corpus are less likely to be classified as regulations. This suggests the need to have a greater

variety of texts in the training corpus, even at the expense of random selection of the sample, cf. the discussion about a representative corpus of webgenres in Chapter 5 by Santini's, this book.

The features discriminating between genres in the experiments described above were based on POS trigrams and punctuation statistics. However, more research is needed into detection of reliable genre indicators, including lexical features (e.g., keywords,<sup>6</sup> frequency bands, n-grams, lexical density, etc), grammatical features other than POS trigrams (the latter are quite sparse in morphologically rich languages, such as Russian), text statistics (average document or sentence length, web-specific markup statistics or URL components, etc). More research is also needed into methods for more efficient population of the feature set with features corresponding to individual categories.

A more general remark concerns the merits of using macrogenres (such as used in this study) vs. microgenres. As mentioned above, the use of the seven macrogenre categories studied in this chapter results in a very coarse classification. If our task to study the microgenre of prototypical blogs, i.e., short personal notes published in a chronological order, the results reported in Section 7.3 are of little help, as this microgenre is contained within in a much bigger macrogenre of "discussions". In addition to this, macrogenre categories are usually abstract, so their reliable recognition requires training. Unlike "look-n-feel" categories, ordinary Internet users or people outside of the community of genre scholars can find it difficult to use them, e.g., for refining the results of web searches.

However, we need a common yardstick for describing the composition of corpora collected using different methods from different sources, so that we can compare the proportion of genres in the BNC and ukWac, or in ukWac and deWac. Table 7.2 demonstrates the possibility of achieving this using a compact genre typology. A list of 70 genres of the BNC or 78 webgenres suggested in [21] would be more difficult to apply as a yardstick because of various reasons:

- the ambiguity usually increases with the number of categories, e.g., Wikipedia entries are (unintentionally) mentioned as an example in the categories of "Encyclopedias" and "Feature stories" in [21];
- the accuracy of automatic classification usually drops if the classifier has to distinguish between a larger number of possible choices, e.g., the F-measure reported in [27] is about 50% for 20 genres vs. 80% in Table 7.2, while machine learning methods used in the two studies are very similar;
- it is difficult to analyse results described in terms of a large number of different parameters (even the seven categories in Table 7.2 present problems for interpretation; if Table 7.2 was expanded to 78 categories, it would be almost impossible to interpret).

---

<sup>6</sup> The use of keywords for genre detection has been studied, e.g., in [29] or [8].

**Acknowledgments** I'm grateful to Silvia Bernardini, Adam Kilgarriff, Katja Markert and Marina Santini for useful discussions. The usual disclaimers apply. The tools for genre classification described in this chapter and the results of classifications of the Internet corpora are available from <http://corpus.leeds.ac.uk/serge/webgenres/>

## References

1. Allen, P., J.A. Bateman, and J.L. Delin. 1999. Genre and layout in multimodal documents: Towards an empirical account. In *Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding, or Retrieval of Documents*, eds. R. Power and D. Scott, 27–34. Cape Cod, MA: American Association for Artificial Intelligence. URL <http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/downloads/allen-bateman-delin.PDF>
2. Baroni, M., and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC2004*. Lisbon.
3. Baroni, M., and A. Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Companion Volume to Proceedings of the European Association of Computational Linguistics*, 87–90. Trento.
4. Baroni, M., F. Chantree, A. Kilgarriff, and S. Sharoff. 2008. Cleaneval: A competition for cleaning web pages. In *Proceedings of the 6th Language Resources and Evaluation Conference, LREC 2008*. Marrakech. URL <http://corpus.leeds.ac.uk/serge/publications/lrec2008-cleaneval.pdf>
5. Biber, D. 1988. *Variations across speech and writing*. Cambridge, MA: Cambridge University Press.
6. Biber, D., and J. Kurjian. 2006. Towards a taxonomy of web registers and text types: A multidimensional analysis. In *Corpus linguistics and the web*, eds. M. Hundt, N. Nesselhauf, and C. Biewer, 109–131. Amsterdam: Rodopi.
7. Braslavski, P. 2004. Document style recognition using shallow statistical analysis. In *ESLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, 1–9. Nancy.
8. Crossley, S.A., and M. Lowerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12(4):453–478.
9. EAGLES. 1996. Preliminary recommendations on text typology. Technical Report EAG-TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document. URL <http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>
10. Ferraresi, A. 2007. Building a very large corpus of English obtained by web crawling: ukwac. Master's thesis, University of Bologna.
11. Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Edward Arnold.
12. Jakobson, R. 1960. Linguistics and poetics. In *Style in Language*, ed. T.A. Sebeok, 350–377. Cambridge, MA: MIT Press.
13. Joho, H., and M. Sanderson. 2004. The SPIRIT collection: An overview of a large web collection. *SIGIR Forum* 38(2):57–61. doi: <http://doi.acm.org/10.1145/1041394.1041395>
14. Kessler, B., Nunberg, G., and H. Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th ACL/8th EACL*, 32–38. Madrid.
15. Kilgarriff, A. 2001. The web as corpus. In *proceeding of corpus linguistics 2001*. Lancaster. URL <http://www.itri.bton.ac.uk/techreports/ITRI-01-14.abs.html>
16. Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3):37–72. URL <http://lt.msu.edu/vol5num3/pdf/lee.pdf>
17. Macdonald, C., and I. Ounis. 2006. The TREC blogs06 collection: Creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow. URL <http://ir.dcs.gla.ac.uk/terrier/publications/macdonald06creating.pdf>

18. Martin, J.R. 1984. Language, register and genre. In *Children Writing: Reader (ECT language studies: Children writing)*, ed. F. Christie, 21–30. Geelong, VIC: Deakin University Press.
19. Mehler, A., and R. Gleim. 2006. The net for the graphs – towards webgenre representation for corpus linguistic studies. In *WaCky! Working papers on the Web as Corpus*, eds. M. Baroni and S. Bernardini. Bologna: Gedit.
20. Meyer zu Eissen, S., and B. Stein. 2004. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*. Ulm.
21. Rehm, G., M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. 2008. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the 6th Language Resources and Evaluation Conference, LREC 2008*. Marrakech.
22. Santini, M. 2007. Automatic identification of genre in web pages. PhD thesis, University of Brighton.
23. Sharoff, S. 2005. Methods and tools for development of the Russian reference corpus. In *Corpus linguistics around the world*, eds. D. Archer, A. Wilson, and P. Rayson, 167–180. Amsterdam: Rodopi.
24. Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*, eds. M. Baroni and S. Bernardini. Bologna: Gedit. <http://wackybook.sslmit.unibo.it>
25. Sharoff, S. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.
26. Sinclair, J. 2003. Corpora for lexicography. In *A practical guide to lexicography*, ed. P. van Sterkenberg, 167–178. Amsterdam: Benjamins.
27. Vidulin, V., M. Luštrek, and M. Gams. 2007. Using genres to improve search engines. In *Proceeding Towards Genre-Enabled Search Engines: The Impact of NLP*. RANLP, URL [http://dis.ijs.si/MitjaL/documents/Vidulin-Using\\_Genres\\_to\\_Improve\\_Search\\_Engines-RANLP-07-TGESE.pdf](http://dis.ijs.si/MitjaL/documents/Vidulin-Using_Genres_to_Improve_Search_Engines-RANLP-07-TGESE.pdf)
28. Witten, I.H., and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.
29. Xiao, Z., and A. McEnery. 2005. Three genres in modern American English. *Journal of English Linguistics* 33(1):62–82.

# Chapter 8

## Web Genre Analysis: Use Cases, Retrieval Models, and Implementation Issues

Benno Stein, Sven Meyer zu Eissen, and Nedim Lipka

### 8.1 Introduction

The genre of a Web document provides information related to the document's form, purpose, and intended audience. Documents of the same genre can address different topics and vice versa, and several researchers consider genre and topic as two orthogonal concepts. Though this claim does not hold without exceptions, genre information attracted much interest as positive or negative filter criterion for Web search results.

Though the undoubted potential of an automatic genre identification for Web pages, retrieval models for genre could not convince in the Web retrieval practice by now. The reasons for this are threefold. First, as was also observed by Santini [32], the proposed genre classifier technology is corpus-centered: their application within Web retrieval scenarios shows a significant degradation of the classification performance, rendering the technology largely useless for genre-enabled Web search. Second, the existing genre retrieval models are computationally too expensive to be applied in an ad-hoc manner. Third, there is no genre palette that fits for all users and all purposes. Ideally, a user should be able to adapt a genre classifier to his or her information need, e.g. by labeling documents as being of an interesting genre or not.

From the mentioned deficits the first one is the most severe: put in a nutshell, the existing Web genre retrieval models generalize insufficiently. Also the second deficit is crucial since it makes the important use case of a genre-enabled Web search unattractive for users who expect a result list from a search engine by the press of a button. We argue that the problems can be overcome, and this chapter will introduce elements of the necessary technological means.

---

B. Stein (✉)

Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany  
e-mail: benno.stein@uni-weimar.de

### 8.1.1 Contributions

Section 8.2 outlines use cases where knowledge about a Web document's genre is exploited to satisfy an information need in question. The scenarios show that genre analysis is not only amenable for standard Web search but represents a universal and powerful instrument for information extraction tasks.

The most important contributions of this chapter relate to the first two deficits mentioned at the outset: we propose concentration characteristics of genre-specific core vocabularies as both generalizable and efficiently computable features for genre retrieval models. In this connection Section 8.3 introduces methods for mining tailored core vocabularies as well as particular statistics as a means for sensible feature quantization. Section 8.4 then investigates the generalization capability of our genre retrieval model and presents new kinds of experiments and analysis methods.

Section 8.5 discusses two alternative realization approaches of a service for genre-enabled Web search. The presented approaches have been put into practice; an implementation in the form of a browser add-on can be downloaded from our Web site.<sup>1</sup>

## 8.2 Use Cases: Genre Analysis in the Retrieval Practice

Web genre analysis is of highly practical interest. In this section we underpin this statement and outline use cases where Web genre analysis forms an essential building block in the information processing chain. From an information retrieval viewpoint, a genre analysis is operationalized by means of a tailored retrieval model; see Section 8.3 for the respective definition and technical background.

The following use cases show the broad spectrum of genre applications, ranging from new kinds of retrieval services to auxiliary technology for information extraction. Section 8.2.1 illustrates topic-centered search technology which has been empowered by genre labeling. Section 8.2.2 shows the role of genre information in vertical search tasks. Section 8.2.3 reports on a feasibility study dealing with the identification of the governing classification principle in a document collection. Longterm objective is the development of smart document classification tools. In Section 8.2.4 genre information is used as a high-level feature for the tailored rendering of Web pages for visually handicapped people. The applications have been operationalized in our research group, and some have reached a mature development state.<sup>2</sup>

---

<sup>1</sup> <http://www.webis.de/research/projects/wega>

<sup>2</sup> While the use cases outlined here focus on the exploitation of genre in texts, the chapter of Paolillo et al. investigates genre emergence in Flash animations posted to Newgrounds.com.

### 8.2.1 Genre-Enabled Web Search

Search engines are the most influential and important applications for the World Wide Web. It stands to reason that an integration of genre-enabling technology may evolve into the most popular Web genre application. Such an integration can happen according to two different paradigms, namely filtering and Web search. Under the filtering paradigm, a user declares his or her information need in terms of a genre preference, and the retrieval process accounts for this constraint. Under the classical Web search paradigm using Google, Live Search, or Yahoo, Web genre information is introduced by assigning genre labels to the snippets in the search results (see Fig. 8.1 for an illustration). Both approaches have their pros and cons, pertaining to retrieval time and retrieval precision. Different Web genre palettes along with technology to identify the genre classes are compiled in Table 8.1.

### 8.2.2 Information Extraction Based on Genre Information

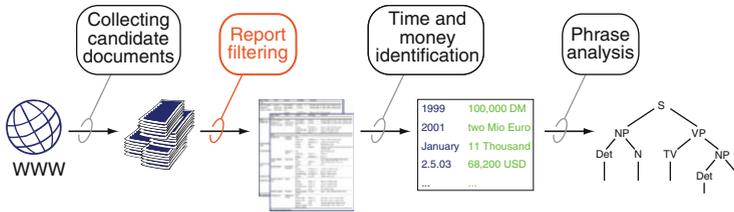
Web genre palettes provide a diversification of documents into text types that is oriented at search habits on the one hand and the emerged culture of Web presences on the other. In a technical sense, Web genre models can be understood as a collective term for retrieval models that quantify arbitrary structure- and presentation-related document features – while being topic-orthogonal at the same time. We have developed such retrieval models for high-level Web services that need a special text type as input. Examples:



Fig. 8.1 Genre labels are superimposed a few seconds after the result list is returned by the search engine. The snapshot shows the Firefox-add-on of WEGA, an acronym for “Web-based Genre Analysis”

**Table 8.1** Research in the field of automatic genre classification for Web-based corpora and digital libraries. An important use case is the development of a richer retrieval result representation in the search interface

Author analysis basis	Web genre palette $Q$	Document representation $d$
Bretan et al. [4] user study with 102 interviewees	Private, public/commercial, journalistic, report, other texts, interactive, discussion, link collection, FAQ, other listing	Simple part-of-speech features, emphatic and down-toning expressions, relative number of digits, average word length, number of images, proportion of links
Lee and Myaeng [19] 7,615 documents	FAQ, home page, reportage, editorial, research article, review, product specification	Genre-specific core vocabulary
Rehm [26] 200 documents	Hierarchy with three granularity levels for academic home pages	HTML meta data, presentation related tags, linguistic features
Meyer zu Eissen and Stein [21] user study with 286 interviewees, 800 documents	Article, discussion, shop, help, personal home page, non-personal home page, link collection, download	Word frequency class, part-of-speech, genre-specific core vocabulary, other close-classed word sets, text statistics, HTML tags
Kennedy and Shepherd [15] 321 documents	Personal, corporate, organizational	HTML tags, phone, email, presentational tags, CSS, URL, link, script, genre-specific core vocabulary
Boese and Howe [3] 342 documents	Abstract, call for papers, FAQ, sitemap, job description, resume, statistics, syllabus, technical paper	Readability scales, part-of-speech, text statistics, HTML tags, bow, HTML title tag, URL, number types, closed-world sets, punctuation
Lim et al. [20] 1,224 documents	Home page, public, commercial, bulletin, link collection, image collection, simple list, input, journalistic, research, official material, FAQ, discussion, product specification, informal	Part-of-speech, URL, HTML tags, token information, most frequent function words, most frequent punctuation marks, syntactic chunks
Freund et al. [13] 800 documents	Best practice, cookbook, demo, design pattern, discussion, documentation, engagement, FAQ, manual, presentation, problem, product page, technical, technote, tutorial, whitepaper	Bag of words
Santini [32] 1,400 documents	Blog, listing, eshop, home page, FAQ, search page, online newspaper front page	Most frequent English words, HTML tags, part-of-speech, punctuation symbols, genre-specific core vocabulary
Santini [32] 2,480 documents	[as before]	Text type analysis plus a combination of layout and functionality tags



**Fig. 8.2** A four stage approach to market forecast summarization. The second step, “Report filtering”, is achieved with a genre analysis

- *Market Forecast Summarization.* Market forecasting seeks to anticipate the future development of new technologies at an early stage. It is vital for most companies in order to develop reasonable business strategies and to make appropriate corporate investments. Market forecasting can be supported by automatically collecting, assessing, and summarizing information from the World Wide Web into a comprehensive presentation of the expected market volume. For this purpose we developed and implemented a four step approach [36]: collecting candidate documents, report filtering, time and money identification, and phrase analysis along with template filling (see Fig. 8.2).

The third as well as the fourth step are computationally very demanding, and the rationale of our approach is to reduce unnecessary NLP effort by a reliable identification of interesting business reports published on the Web. The heart of this strategy is a genre analysis in the report filtering step.

- *Retrieval of Scholar Material.* Specialized search engines and technology for vertical search are building blocks of future information extraction applications for the retrieval of scholar material. They shall be able to identify, synthesize, and present Web documents related to exercises, FAQs, introductory readings, definitions, or sample solutions – given a topic in question. The driving force is a reliable document type and genre analysis.
- *Focused Crawling for Plagiarism Analysis.* The discriminative power of a genre classifier can also be utilized at the crawling stage. Here, the challenges result from a retrieval model that must get by with few and small document snippets. An interesting application is plagiarism analysis, for which we are developing crawling technology that focuses on research articles, book chapters, and theses.

### 8.2.3 Organizing Collections in Both Topic and Genre Dimensions

The categorization of documents, bookmarks, or digital document identifiers in general can happen topic-centered, genre-centered, or in a combined fashion. Having identified the categorization paradigm one can support automatic classification, provide user guidance for insertion (*This is not the correct genre!*), give hints or special views for browsing and searching, and identify classes that are not properly organized. In [37] we have broken down this and similar analysis to the following question:

Given a categorization  $\mathcal{C}$  of documents (or bookmarks, links, document identifiers), can we provide a reliable assessment whether  $\mathcal{C}$  is governed by topic or by genre considerations?

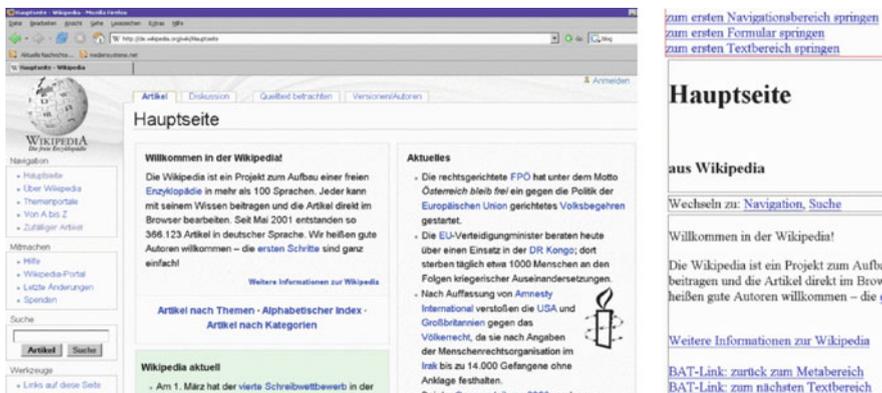
The question can be answered in the following five steps, where essentially a model fitting problem is solved. Let  $D$  be a set of documents, and let  $\mathcal{C}$  be a categorization of  $D$  that is either governed by topic or by genre considerations.

1. Construct for each  $d \in D$  two retrieval models, one under the genre retrieval model,  $R_G$ , and one under the topic retrieval model,  $R_T$ .
2. Construct two similarity graphs  $G_G$  and  $G_T$ . The edge weights in these graphs result from the similarity computations under  $R_G$  and  $R_T$  respectively.
3. Apply a clustering algorithm to the graphs  $G_G$  and  $G_T$ . The resulting clusterings are designated as  $\mathcal{C}_G$  and  $\mathcal{C}_T$ .
4. Compute the  $F$ -measure (or another external reference measure) to quantify the congruence between  $\mathcal{C}$  and  $\mathcal{C}_G$  as well as between  $\mathcal{C}$  and  $\mathcal{C}_T$ . The resulting values are designated as  $F_{\mathcal{C}_G}$  and  $F_{\mathcal{C}_T}$ .
5. If  $|F_{\mathcal{C}_G} - F_{\mathcal{C}_T}|$  is significant,  $\mathcal{C}$  is organized under genre considerations if  $F_{\mathcal{C}_G} > F_{\mathcal{C}_T}$ , and under topic considerations otherwise.

The analysis in [37] revealed that a definite answer to the above question can be given, if the impurity ratio, i.e., the ratio between topic classes and genre classes (or vice versa) is larger than 1:2.

### 8.2.4 Empower Web Page Abstraction with Genre Information

Web page abstraction is concerned with the preparation of Web pages in a consistent and clearly arranged form. Possible applications for such a technology are mobile Internet devices with small displays, but also the simplification of Web pages for visually handicapped people whose access to the World Wide Web is managed with



**Fig. 8.3** The browser add-on for Web page abstraction BAT, an acronym for “Blind Accessibility Tool”. The *left picture* shows the original Web page, the *right picture* the related abstraction, where navigational areas and main elements have been identified and reorganized in textual form

a braille reader. Since the use and the organization of a Web page's content elements is genre-dependent, the identification of the underlying genre class gives valuable hints for a fully-automated Web page abstraction. Figure 8.3 illustrates the use of the Blind Accessibility Tool add-on BAT that has been developed in our working group.<sup>3</sup>

The underlying retrieval model is based on the document object model, DOM: the genre-revealing features are computed heuristically from the DOM tree, exploiting node types, node neighborhoods, node depth information, and node content.

### 8.3 Construction of Genre Retrieval Models

It is necessary to distinguish between a real-world document  $d$  and its computer representation  $\mathbf{d}$ .  $d$  can stand for a paper, a book, or a Web site, while  $\mathbf{d}$  may be a vector of terms, concepts, or high-level features, but also a suffix tree or a signature file. Likewise,  $D$  denotes a collection of real-world documents, and  $\mathbf{D}$  denotes the set of the related computer representations.

Given an information need in question a retrieval model  $\mathcal{R}$  provides the rationale for constructing a particular representation  $\mathbf{d}$  from a real-world document  $d$ . Examples for retrieval models are the vector space model, the binary independence model, or the latent semantic indexing model [30, 28, 7]. Note that document representations and retrieval models are orthogonal concepts:  $\mathbf{d}$  defines the features computed from  $d$  while  $\mathcal{R}$  explains the retrieval performance of  $\mathbf{d}$  against the background of the retrieval task and linguistic theories.

A genre retrieval model is a retrieval model that addresses queries related to a palette of genre classes [38]; it is defined as follows.

**Definition 1** (Genre Retrieval Model) Let  $D$  be set of documents, and let  $Q, \mathcal{Q} = \{c_1, \dots, c_k\}$ , be a set of genre class labels, also called genre palette. A genre retrieval model  $\mathcal{R}$  for  $D$  and  $Q$  is a tuple  $\langle \mathbf{D}, \gamma_{\mathcal{R}} \rangle$ :

1.  $\mathbf{D}$  is the set of representations of the documents  $D$ .
2.  $\gamma_{\mathcal{R}}$  is a classifier and assigns one or more genre class labels to a document representation  $\mathbf{d} \in \mathbf{D}$ :

$$\gamma_{\mathcal{R}} : \mathbf{D} \rightarrow \mathcal{P}(\{c_1, \dots, c_k\})$$

The most important part of a genre retrieval model  $\mathcal{R}$  cannot be made explicit in the definition, namely, the theoretical basis and the rationale behind the mapping  $\alpha : D \rightarrow \mathbf{D}$  which computes the representation  $\mathbf{d}$  for a document  $d \in D$ .

The development of genre retrieval models is an active research field with several open questions, and only little is known concerning a user's information need and the adequacy of a retrieval model  $\mathcal{R}$ . Early work in automatic genre classification dates back to 1994, where Karlgren and Cutting presented a feasibility study

<sup>3</sup> <http://www.webis.de/research/projects/bat>

for a genre analysis based on the Brown corpus [14]. Later on followed several publications investigating different corpora, using more intricate or less complex retrieval models, stipulating other concepts of genre, or reporting on new applications [1, 8, 12, 16, 25, 34, 41].

Genres *on the Web* have been investigated since 1999 [4]. Table 8.1 compiles research that received attention: the table lists the basis of the analysis, the genre palette  $Q$ , and the document representation  $\mathbf{d}$ . The underlying use case is a genre-enabled Web search. The approaches from Crowston and Williams, Roussinov et al., Dimitrova et al., Chapter 3 by Rosso and Haas (this book) were not included since the authors provided suggestions rather than a technical specification about their genre retrieval models [6, 9, 29].

### 8.3.1 Problems of Genre Retrieval Models and Lessons Learned

In the following we concentrate on two problems:

1. the insufficient generalization capability of current genre retrieval models  $\mathcal{R}$ , and
2. the high computational effort of the mapping  $d \mapsto \mathbf{d}$ .

With respect to the third problem mentioned at the outset, the inadequacy of a unique, single-label genre palette, we propose no special solution but follow the argument of Santini [32]: Web page diversity and Web page evolution can be captured by a flexible genre classification palette, capable of performing a zero-, one-, or multi-label genre assignment. In this book, the problem of defining a suitable text typology for the Web is discussed by Sharov or by Rosso and Haas among others. karlgrens and Crowston et al. point out reasons why it is so difficult to develop a commonly accepted Web genre taxonomy.

#### *Insufficient Generalization Capability*

The authors of the approaches listed in Table 8.1 reported on classification results for the correct assignment of genre classes. The obtained (cross-validated) performances are surprisingly high, reaching from 75% with  $|Q| = 16$  genre labels [20] up to 90% with  $|Q| = 7$  genre labels [19]. These and similar results were achieved with rather small training corpora, containing between several hundred and a few thousand documents.

Let  $\gamma_{\mathcal{R}_1}$  be the genre classifier of a genre retrieval model  $\mathcal{R}_1$  trained with corpus  $D_1$ , and let  $\gamma_{\mathcal{R}_2}$  be the genre classifier of a genre retrieval model  $\mathcal{R}_2$  trained with corpus  $D_2$ . With respect to the common genre labels of two concrete retrieval models Santini investigated the generalization capability of classifier  $\gamma_{\mathcal{R}_1}$  to corpus  $D_2$  and, vice versa, of classifier  $\gamma_{\mathcal{R}_2}$  to corpus  $D_1$ .<sup>4</sup> It turned out that the retrieval precision

---

<sup>4</sup> Santini uses the term “exportability” in this connection. Actually, she measured the agreement between  $\gamma_{\mathcal{R}_1}$  and  $\gamma_{\mathcal{R}_2}$ , which is a particular facet of the generalization capability [39].

decreased by more than one order of magnitude, a truly disappointing result [32]. In this book Santini provides an extended analysis in this respect, by cross-testing a genre classifier’s performance on single labels.

The classification knowledge that is operationalized within a genre retrieval model  $\langle \mathbf{D}_1, \gamma_{\mathcal{R}_1} \rangle$  can be exported to a corpus  $D_2$  if the model captures the *intensional semantics* of the concept “genre”. The intensional semantics of a genre retrieval model can be understood as its capability to comply with the extensional semantics of genre in different worlds, say, as its capability to correctly classify documents from different corpora. If so, the model provides a high generalization capability. The generalization capability of a genre retrieval model depends on its bias, which in turn can be understood as the size of the hypotheses space wherein the learning algorithm is searching for the model.

The bias of a learning algorithm can be assessed with respect to two dimensions: the size and the exploration strategy of the hypothesis space. Different authors use different names for both types of biases, Table 8.2 gives an overview. The table also shows the impact of the bias strength: while a strong bias of Type I raises a classifier’s generalization capability, a strong bias of Type II raises the sensitivity of a learning algorithm with respect to the training data. The former is a highly appreciated property, whereas the latter is absolutely to be avoided.

There is also the question of the correctness of a biased learning algorithm [40]. Independent of its type, a strong bias decreases the probability of finding the correct hypothesis. In particular, a strong bias of Type I will inevitably compromise the correctness – simply due to the construction of a coarse hypothesis space, whereas a strong bias of Type II leaves – at least theoretically – the chance of choosing the correct hypothesis.

Most of the Web genre models listed in Table 8.1 have a very weak bias of Type I. If one analyzes the proportion between the number of training samples and the size of the underlying hypothesis space, running the risk of rote learning becomes obvious – despite sophisticated learning technology such as support vector machines. The generalization capability of a genre retrieval model can be

**Table 8.2** Two types of biases can be distinguished, pertaining to search space size and search space exploration. The table lists the synonyms that are used in the literature (upper row) and illustrates the impact of the bias strength towards the generalization capability of the learning algorithm and its training data sensibility (lower row)

Bias	Type I search space size		Type II search space exploration	
Synonyms	Exclusive bias	Rendell [27]	Preferential bias	Rendell [27]
	Representational bias	Quinlan [24]	Procedural bias	Quinlan [24]
	Restriction bias	Mitchell [22]	Search bias	Mitchell [22]
	<i>Strength</i> weak $\leftarrow$ $\rightarrow$ strong		<i>Strength</i> weak $\leftarrow$ $\rightarrow$ strong	
	<i>Generalization capability</i> low $\leftarrow$ $\rightarrow$ high		<i>Training data sensibility</i> low $\leftarrow$ $\rightarrow$ high	

measured by the “stability” and “efficiency” of its construction process with respect to training samples; Section 8.4.2 introduces the theoretical means and reports on respective experiments.

### *High Computational Effort*

Table 8.1 lists a wide range of feature types to compute the document representation  $\mathbf{d}$  for retrieval models:

- *Presentation-related Features*. Frequency counts for figures, tables, paragraphs, headlines, captions. HTML-specific analysis regarding colors, hyperlinks, URLs, or mail addresses.
- *Simple Text Statistics*. Frequency counts for clauses, paragraphs, delimiters, question marks, exclamation marks, and numerals.
- *Special Closed-Class Word Sets and Core Vocabularies*. Use of currency symbols, help phrases, shop phrases, calendar, or countries.
- *Word Frequency Class Analysis*. Use of special words, common words, or misspelled words.
- *Part-of-Speech Analysis*. Frequency counts for nouns, verbs, adjectives, adverbs, prepositions, or articles.
- *Syntactic Group Analysis*. Use of tenses, relative clauses, main clauses, adverbial phrases, or simplex noun phrases.

The effort to compute the mentioned features is between linear time in the text length, e.g. for simple frequency counts, and ranges up to cubic effort and higher for the parsing of syntactic groups. The usefulness and, even more importantly, the cost-benefit ratio of these features with respect to a reliable genre analysis is unclear. Hence the researchers who build genre retrieval models tend to include a feature instead of leaving it out. In this sense the model formation task is shifted to the learning algorithm, which identifies and weights the most discriminating features based on the training data. This strategy is acceptable if training data is plentiful and – with respect to the classification task – sensibly distributed. Both requirements are not fulfilled here: the construction of training corpora is expensive, as the small sample sizes in Table 8.1 show (see the first column). Moreover, considering the different user- and task-specific genre palettes and the impracticality to estimate the document type distribution on the World Wide Web, very little can be stated about the a-priori probabilities of document types.

The combination of rich feature models with small training corpora is crucial in two respects: it compromises generalization capability and makes the learning process sensible to the training data. A way out is the use of few features with a coarse domain.

### **8.3.2 New Elements for Genre Retrieval Models**

The potential of features related to genre-specific core vocabularies are underestimated. The reasons for this are twofold: (i) till now genre-specific core

vocabularies are compiled manually, following intuition. (ii) The evaluation of core vocabularies is limited to simple count statistics. This subsection outlines new elements for the construction of robust and lightweight genre retrieval models: an automatic extraction of core vocabularies and new features that quantify distribution information. Details can be found in [38].

For the set  $D$  of documents let  $\mathcal{C} = \{C_1, \dots, C_k\}$  be an exclusive genre categorization of  $D$ . I.e.,  $\bigcup_{C \in \mathcal{C}} C = D$  and  $\forall C_i, C_j, j \neq i \in \mathcal{C} : C_i \cap C_j = \emptyset$ . For a genre class  $C \in \mathcal{C}$ , let  $T_C$  denote the core vocabulary specific for  $C$ . Similar to Broder we argue that  $T_C$  is comprised of navigational, transactional, structural, and informative terms [5]. The combination, distribution, presence or absence of these terms encode a considerable part of the genre information.

- *Navigational Terms.* Appear in labels of hyperlinks and in anchor tags of Web pages. Examples: “Windows”, “Mac”, or “zip” in download sites, links to “references” in articles.
- *Transactional Terms.* Appear in sites that interact with databases, and manifest in hyperlink labels, forms, and button captions. Examples: “add to shopping cart”, “proceed to checkout” in online shops, buttons labeled “download” on download pages.
- *Structural Terms.* Appear in sites that maintain meta information like time and space. Examples include the meta information of posts in a discussion forum (“thread”, “replies”, “views”, parts of dates) and terms that appear in addresses on home pages (“address”, “street”).
- *Informative Terms.* Appear not in functional HTML elements but imply functionality though. Examples include “kb” or “version” on download sites, “price” or “new” on shopping sites, and “management”, “technology”, or “company” on commercial sites.

The terms in  $T_C$  are both predictive and frequent for  $C$ . Terms with such characteristics can be identified in  $\mathcal{C}$  with approaches from topic identification research, in particular Popescul’s method and the weighted centroid covering method [17, 18, 23, 35]. In order to mine genre-specific core vocabulary both methods must be adapted: they do not quantify whether a term is *representative* for  $C$ ; a deficit, which can be repaired without compromising the efficient  $O(m \log(m))$  runtime of the methods, where  $m$  designates the number of terms in the dictionary [38].

### *Concentration Measures*

In the simplest case, the relation between  $T_C$  and a document  $d$  can be quantified by computing the fraction of  $d$ ’s terms from  $T_C$ , or by determining the coverage of  $T_C$  by  $d$ ’s terms. However, if genre-specific vocabulary tends to appear concentrated in certain places on a Web page, this characteristic is not reflected by the mentioned features, and hence it cannot be learned by a classifier  $\gamma_{\mathcal{R}}$ . Examples for Web pages on which genre-specific core vocabulary appears concentrated: private home pages (e.g. address vocabulary), discussion forums (e.g. terms from mail headers),

and non-personal home pages (e.g. terms related to copyright and legal information). The following two statistics quantify two different vocabulary concentration aspects:

1. *Maximum Term Concentration*. Let  $d \in D$  be represented as a sequence of terms,  $d = \{w_1, \dots, w_m\}$ , and let  $W_i \subset d$  be a text window of length  $l$  in  $d$  starting with term  $i$ , say,  $W_i = \{w_i, \dots, w_{i+l-1}\}$ . A natural way to measure the concentration of terms from  $T_C$  in different places of  $d$  is to compute the following function for different  $W_i$ :

$$\kappa_{T_C}(W_i) = \frac{|W_i \cap T_C|}{l}, \quad \kappa_{T_C}(W_i) \in [0, 1]$$

The overall concentration is defined as the maximum term concentration:

$$\kappa_{T_C}^* = \max_{W_i \subset d} \kappa_{T_C}(W_i), \quad \kappa_{T_C}^* \in [0, 1]$$

2. *Gini Coefficient*. In contrast to the  $\kappa_{T_C}$  statistic, which quantifies the term concentration strength within a text window, the Gini coefficient can be used to quantify to which extent genre-specific core vocabulary is distributed unequally over a document. Again, let  $W_i$  be a text window of size  $l$  sliding over  $d$ . The number of genre-specific terms from  $T_C$  in  $W_i$  is  $v_i = |T_C \cap W_i|$ . Let  $A$  denote the area between the uniform distribution line and the Lorenz curve of the distribution of  $v_i$ , and let  $B$  denote the area between the uniform distribution line and the  $x$ -axis. The Gini coefficient is defined as the ratio  $g = A/B$ ,  $g \in [0, 1]$ . A value of  $g = 0$  indicates an equal distribution; the closer  $g$  is to 1 the more unequal  $v_i$  is distributed.

### Discussion

Concentration measures capture distribution information of different subsets of a document's terms. These subsets, called core vocabularies here, as well as their concentration analysis, form the basis for non-linear features that cannot be constructed by the state of the art machine learning technology. This is the reason why our research focuses on the idea of sensible genre retrieval models, instead of resorting to the standard bag of word model where the learning algorithms accomplishes a low-level feature (= term) selection. Note that in Chapter 6 by Kim and Ross (this book) also propose features that consider the word distribution in a document.

## 8.4 Evaluation

This section addresses evaluation-related issues of Web genre identification. We discuss approaches for improving the generalization capability and propose statistics to quantify this property for genre retrieval models. These statistics are used to evaluate different genre retrieval models with respect to two popular Web genre corpora.

### 8.4.1 Improving Generalization Capability

Improving a classifier's generalization capability means to restrict its representational bias. In practice, this goal is achieved by (i) reducing the number of features, (ii) reducing the number of values a feature can take, and (iii) replacing weak features by discriminative features.

The proposed concentration measures, i.e. maximum concentration and Gini coefficient of core vocabulary distributions, impose one feature per genre class, resulting in eight features for a document of a collection with eight genre classes. In comparison to a standard text classification approach with SVMs, the number of features introduced by these concentration measures is orders of magnitude smaller, addressing Point (i), and, as our experiments show, Point (iii).

Point (ii) can be tackled by discretizing continuous features. A standard approach is the substitution of categorical or nominal features for continuous features [11, 2]; see [10] for an overview of such methods. Although these methods might be powerful their evaluation for Web genre analysis is out of the scope of this chapter.

### 8.4.2 Measuring Generalization Capability

In the following, the concepts predictive accuracy, classifier agreement, and export accuracy are defined; the notation is adapted from [39]. Simply put, the concepts quantify the classification performance, the impact of classifier variation, and the impact of corpus variation.

**Definition 2** (Predictive Accuracy) Let  $D$  be a document set organized according to a genre palette  $Q$ . Moreover, let  $\alpha : D \rightarrow \mathbf{D}$  be a mapping that computes a document representation, and let  $\langle \mathbf{D}, \gamma_{\mathcal{R}} \rangle$  be a genre retrieval model for  $D$ . Then the predictive accuracy  $a_{\gamma_{\mathcal{R}}}$  of the classifier  $\gamma_{\mathcal{R}}$  is the probability that  $\gamma_{\mathcal{R}}$  will assign the correct genre class label to an unseen example  $(\mathbf{d}, c^*) \in \mathbf{D} \times Q$ :

$$a_{\gamma_{\mathcal{R}}} := P(\gamma_{\mathcal{R}}(\mathbf{d}) = c^*)$$

The *predictive* accuracy is estimated by classifying unseen examples from  $\mathbf{D} \times Q$ , and it may not be confused with the training set accuracy. It is possible that two classifiers that have the same predictive accuracy may disagree on predicting particular samples.

**Definition 3** (Classifier Agreement) Let  $D$  be a document set organized according to a genre palette  $Q$ . Moreover, let  $\alpha_1 : D \rightarrow \mathbf{D}_1$  and  $\alpha_2 : D \rightarrow \mathbf{D}_2$  be two mappings that compute two document representations, and let  $\langle \mathbf{D}_1, \gamma_{\mathcal{R}_1} \rangle$  and  $\langle \mathbf{D}_2, \gamma_{\mathcal{R}_2} \rangle$  be two genre retrieval models for  $D$ . Then the agreement of the classifiers  $\gamma_{\mathcal{R}_1}$  and  $\gamma_{\mathcal{R}_2}$  is defined as follows:

$$agree(\gamma_{\mathcal{R}_1}, \gamma_{\mathcal{R}_2}) := P(\gamma_{\mathcal{R}_1}(\mathbf{d}_1) = \gamma_{\mathcal{R}_2}(\mathbf{d}_2)),$$

where  $\mathbf{d}_1 \in \mathbf{D}_1$  and  $\mathbf{d}_2 \in \mathbf{D}_2$  are representations of the same document  $d \in D$ .

That is, the classifier agreement is the probability that two genre retrieval models make the same decision on the genre of a document. Consider that  $\alpha_1 = \alpha_2$  and hence  $\mathbf{D}_1 = \mathbf{D}_2$  can hold: the two genre retrieval models rely on the same document representation, but differ with respect to their machine learning settings. In particular,  $\gamma_{\mathcal{R}_1}$  and  $\gamma_{\mathcal{R}_2}$  can result from training on different samples while using the same classifier type. In this important analysis case the classifier agreement quantifies the *training data sensibility* of a genre retrieval model (see also Table 8.2, right column).

**Definition 4** (Export Accuracy) Let  $D_1 \subset D$  and  $D_2 \subset D$  be two document sets organized according to the genre palettes  $Q_1$  and  $Q_2$ ,  $Q_1 \cap Q_2 \neq \emptyset$ . Moreover, let  $\alpha : D \rightarrow \mathbf{D}$  be a mapping that computes the document representations  $\mathbf{D}_1 \subset \mathbf{D}$  and  $\mathbf{D}_2 \subset \mathbf{D}$ , and let  $\langle \mathbf{D}_1, \gamma_{\mathcal{R}_1} \rangle$  be a genre retrieval model for  $D_1$ . Then the export accuracy of the genre retrieval model  $\langle \mathbf{D}_1, \gamma_{\mathcal{R}_1} \rangle$  with respect to  $D_2$  is defined as follows:

$$e_{\gamma_{\mathcal{R}_1}, D_2} := P(\gamma_{\mathcal{R}_1}(\mathbf{d}_2) = c^*),$$

where  $\mathbf{d}_2 \in \mathbf{D}_2$  is the representation of a document  $d_2 \in (D_2 \setminus D_1)$  with genre class  $c^* \in (Q_1 \cap Q_2)$ .

That is, the export accuracy is the probability that the assigned genre of a document of an external corpus is correct. Note that the export accuracy is affected by the homogeneity of the training corpus. The export accuracy of a genre retrieval model  $\langle \mathbf{D}_1, \gamma_{\mathcal{R}_1} \rangle$  with respect to  $D_2$  quantifies whether the combination of  $D_1$ ,  $\alpha$ , and  $\gamma_{\mathcal{R}_1}$  captures the gist of the genre classes in  $Q_1 \cap Q_2$ . Only if the document set  $D_1$  is representative, if the mapping  $\alpha$  is sensible, and if  $\gamma_{\mathcal{R}_1}$  generalizes sufficiently, the classifier  $\gamma_{\mathcal{R}_1}$  will perform acceptably for the documents in  $D_2$ . Typically,  $D_2$  is compiled by different users, and the claimed conditions are not fulfilled. Hence we observe  $e_{\gamma_{\mathcal{R}_1}, D_2} < a_{\gamma_{\mathcal{R}_1}}$  in most cases.

### 8.4.3 Experiments

We now discuss the generalization capability of genre retrieval models regarding the measures introduced in the Definitions 2, 3, and 4. Our empirical analysis illustrates the theoretical observation from above: the stronger the representational bias of a retrieval model the higher is its generalization capability.

The analysis is based on the Web genre corpora “KI-04” with 8 Web genre classes [21], denoted as  $A$ , and the “7-Webgenre collection” [31], denoted as  $B$ .<sup>5</sup> These corpora are sketched in Table 8.1, row 4 and row 8. The questions to be answered refer to the generalization capabilities of different genre retrieval models. In particular, the following retrieval models are examined, which differ in the computed representation of a document:

---

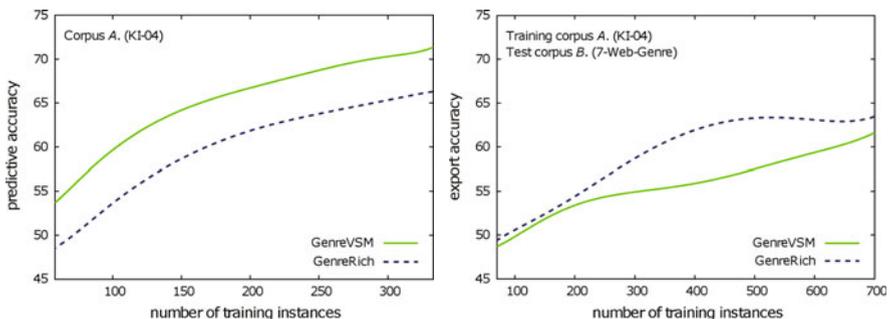
<sup>5</sup> KI-04 can be downloaded from <http://www.webis.de/research/corpora>. In the experiments the extended version of this corpus (1,200 Web pages) was used.

1. GenreVSM. The vector space model using  $tf \cdot idf$  term weighting scheme, comprising about 3,500 features.
2. GenreVoc. A genre retrieval model based on the core vocabulary analysis as introduced in Section 8.3, comprising a total of 26 features.
3. GenreBasic. A basic genre retrieval model based only of HTML features, link features, and character features, comprising a total of 54 features.
4. GenreRich. A rich genre retrieval model based on the features of GenreBasic along with part-of-speech features and vocabulary concentration features, comprising a total of 98 features.
5. GenreRichNoVoc. The GenreRich retrieval model without the vocabulary concentration features, comprising a total of 72 features.

Each experiment was repeated and averaged using 10 random draws of the respective number of training documents; the applied machine learning technology was a support vector machine.

The presumably most important property of a Web genre retrieval model is a high export accuracy. Consider in this connection Fig. 8.4: the left plot shows the predictive accuracy of the retrieval models GenreVSM and GenreRich – trained on and applied to documents of corpus *A* containing 1,200 documents. The right plot shows the export accuracy of these classifiers with respect to corpus *B* containing 600 documents, with  $Q_A \cap Q_B = \{\text{shop, personal home page, link list}\}$ . In both plots the *x*-axis shows the sample size of the training set taken from corpus *A*; the *y*-axis shows corresponding test set accuracies on corpus *A* (left plot) and the test set accuracies on corpus *B* (right plot), called the export accuracy.

Observe that the GenreVSM model achieves a significantly higher predictive accuracy than the GenreRich model (see Fig. 8.4, left plot); with respect to the sample size both show the same consistency characteristic. We explain the high predictive accuracy of GenreVSM with its higher training data sensibility, which is beneficial in homogeneous corpora. Even under a successful cross validation test the predictive accuracy and the export accuracy will considerably diverge.



**Fig. 8.4** Predictive accuracy (*left*) and export accuracy (*right*) of the retrieval models GenreVSM and GenreRich, depending on the size of the training set, which is always drawn from corpus *A* (KI-04). The predictive accuracy is estimated on a test set of corpus *A*, while the export accuracy is estimated on a test set of corpus *B* (7-Webgenre collection)

A corpus may be homogeneous because of the following reasons:

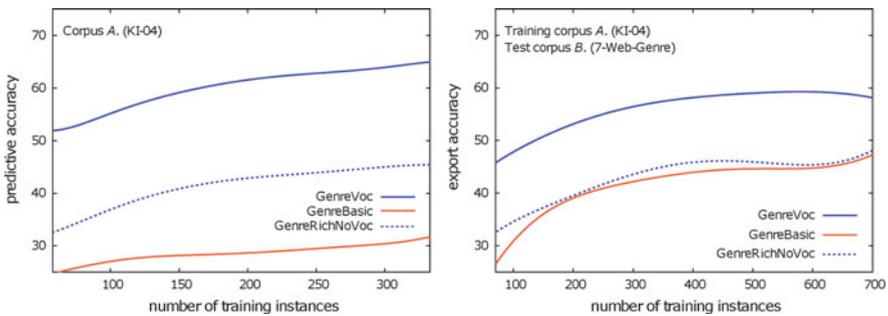
1. The corpus is compiled by a small group of editors who share a similar understanding of genre.
2. The editors introduce subconsciously an implicit correlation between topic and genre.
3. The editors collect their favored documents only.
4. The editors rely on a single search engine whose ranking algorithm is biased towards a certain document type.

Corpus homogeneity is unveiled when analyzing the export accuracy, which drops significantly (by 21%) for the GenreVSM model (see Fig. 8.4, right plot). For the GenreRich model the export accuracy drops only by 8%. The robustness of the GenreRich model is a consequence of its small number of features, which is more than an order of magnitude smaller compared to the GenreVSM model.

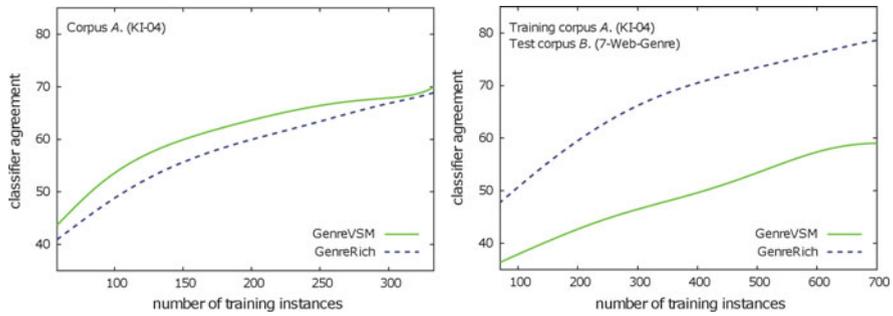
The plots shown in Fig. 8.5 quantify also the drop in export accuracy (left plot → right plot), but analyze different retrieval model variants whose feature sets are subset of the GenreRich model:

- The GenreVoc model shows a small drop in the export accuracy, which is rooted in the fact that the core vocabulary has a small – but acceptable – corpus dependency.
- For the GenreRichNoVoc model the export accuracy remains pretty constant. The reasons for this stability are the small hypothesis space and a small corpus dependency of the features.
- For the GenreBasic model the export accuracy is significantly higher than the predictive accuracy. We explain this behavior with the high discriminative power of the HTML features and link features with respect to the genre classes shop, personal home page, and link list.

Figure 8.6 shows results from an agreement analysis for classifiers of the GenreVSM model and the GenreRich model. The  $x$ -axis denotes the the size of the training set, which is always drawn from corpus A (KI-04). As expected, both plots show the monotonous characteristic of the classifiers subject to the training set size.



**Fig. 8.5** Predictive accuracy (*left*) and export accuracy (*right*) of the retrieval models GenreVoc, GenreRichNoVoc, and GenreBasic, using the same settings as in the experiments shown in Fig. 8.4



**Fig. 8.6** Classifier agreement of the retrieval models GenreVSM and GenreRich, depending on the size of the training set, which is always drawn from corpus *A*. In the *left plot* the agreement is analyzed on corpus *A*, while in the *right plot* the agreement is analyzed on corpus *B*

Observe in the left plot in Fig. 8.6 that the agreement of both classifiers is quite similar, although the representational bias of the GenreVSM model is weaker than the bias of the GenreRich model. Again, this behavior can be explained by the homogeneity of the corpus. However, the situation is different if the classifier agreement is analyzed on a test corpus different from the training corpus (see the right plot in Fig. 8.6): the agreement of classifiers under the GenreRich model is much better than the agreement of classifiers under the GenreVSM model. That is, classifiers under the GenreVSM model are corpus-specific (overfitted) whereas classifiers under the GenreRich model are not, they provide a much higher generalization capability.

A key measure for evaluating Web genre retrieval is the export accuracy. Using an independent corpus for the accuracy evaluation of a genre retrieval model gives consolidated findings and a significant model selection criterion. In this respect the GenreRich model is superior to the other genre retrieval models in our analysis. The high classifier agreement of the GenreRich model on corpus *A* and particularly on corpus *B* shows that the chance of being misled by the training set, and the overfitting risk, is low.

## 8.5 Implementing Genre-Enabled Web Search

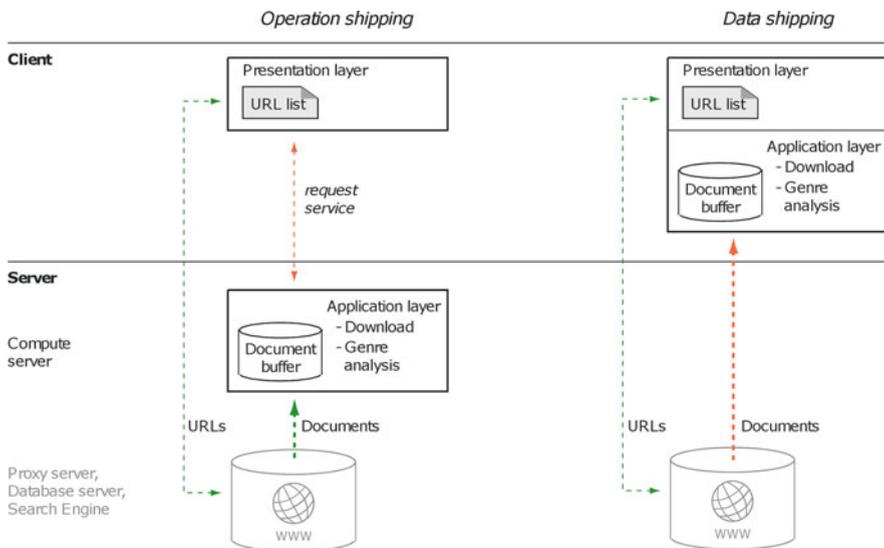
The aim of a genre-enabled Web search is to combine genre information with the standard list-based topic search. To implement a solution for this use case several design decision have to be made:

1. What are useful classes in a genre palette?
2. Shall a user be able to define new or own genre classes?
3. How shall genre information be integrated into the search process?
4. Is a distributed software architecture suited or necessary?

WEGA, a software for Web-based genre analysis that has been developed in our working group, can be characterized as follows: it implements the genre palette

shown in the fourth row in Table 8.1, and the classification results are integrated as genre labels into a standard result list (see Fig. 8.1). Based on such a kind of user interface, new and user-definable genre classes can be conveniently integrated, and a future version of WEGA shall provide this feature. Of course other visualization paradigms are conceivable: within the categorizing search engine AIssearch we employ a filter-based interface paradigm where documents are visualized as nodes of a hyperbolic graph, which can be faded in and out.<sup>6</sup> Presumably, a general way to combine genre and topic information cannot exist, and the information visualization paradigm must be tailored to the use case. In the remainder we concentrate on the fourth, software-technical question.

The first prototypes of WEGA were implemented according to the client-server-paradigm, simply because the sophisticated feature computation should not be carried out by the Web browser but by a powerful Web service. If the execution of high-level operations is shifted to a third party one speaks from “operation shipping” – in contrast to “data shipping” where even computationally intensive tasks are executed at the client site. Various issues are bound up with the decision to pursue the one or other strategy, and the advantages of one paradigm turn to disadvantages of the other [33]. Figure 8.7 illustrates the key difference between an operation shipping implementation and a data shipping implementation: in the former, presentation and application form a distributed system, while in the latter both are located at the client site.



**Fig. 8.7** The genre-enabled Web search WEGA was implemented according to two different software engineering paradigms: operation shipping (*left*) and data shipping (*right*)

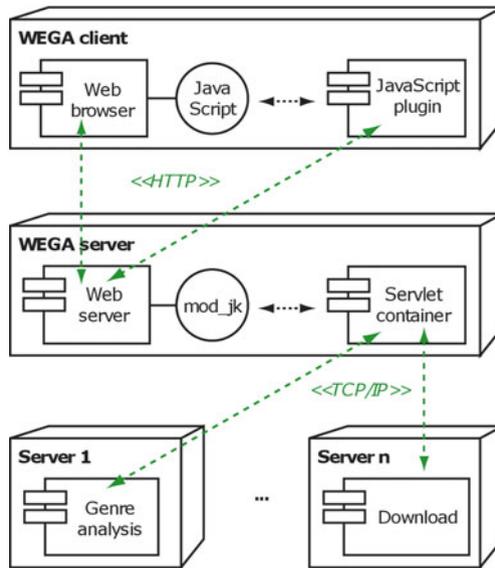
<sup>6</sup> <http://www.webis.de/research/projects/aisearch>

The actual version of WEGA follows the data shipping paradigm; it is realized as a Firefox add-on and implements a lightweight GenreRich model, i.e. the features of a GenreRich model without the part-of-speech features along with a linear discriminant analysis  $\gamma_{\mathcal{R}}$ . It can be downloaded from our Web site.<sup>7</sup> Under either paradigm the same functionality is realized in WEGA, however, by using different technical means:

- *Operation Shipping WEGA*. Presentation layer: DOM + AJAX (= Asynchronous JavaScript and XML). Retrieval model computation: Java servlet in servlet container. To learn more about the software architecture Figs. 8.8 and 8.9 provide an UML diagram for both the component deployment and the component interplay.
- *Data Shipping WEGA*. Presentation layer: DOM + AJAX. Retrieval model computation: JavaScript.

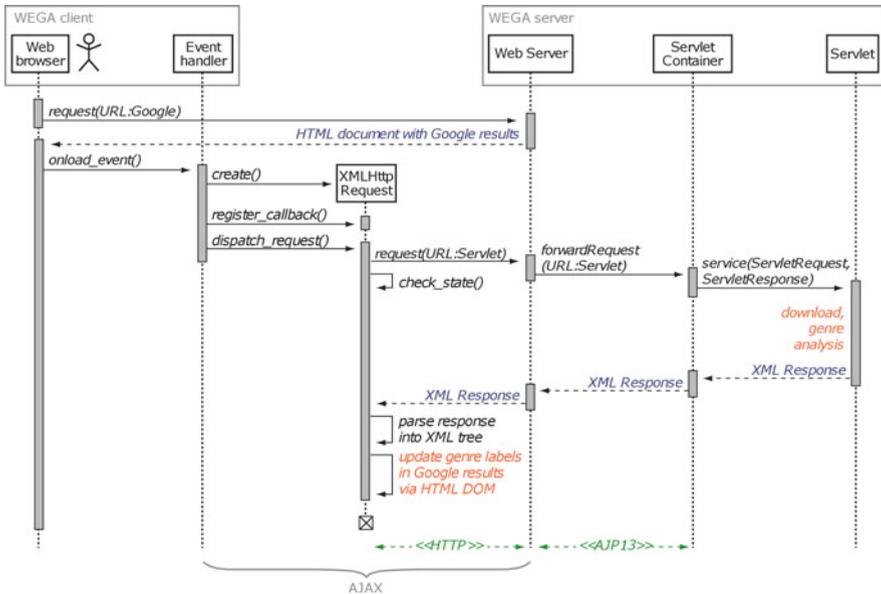
The data shipping paradigm came within the realms of possibility with the new elements for light-weight genre retrieval models, outlined in Section 8.3.2. However, the computationally means within a Web browser are inferior to that of the servlet technology; to get an idea of the effective difference Table 8.3 contrasts selected characteristics of the implementations.

Note that one of the biggest disadvantages of an operation shipping implementation for a Web genre analysis is the possible infringement of privacy and the



**Fig. 8.8** The deployment diagram of WEGA under the operation shipping paradigm. A servlet container provides the necessary components for the analysis service

<sup>7</sup> <http://www.webis.de/research/projects/wega>



**Fig. 8.9** The sequence diagram of WEGA under the operation shipping paradigm. The *middle part* of the diagram show the an synchronous interaction realized with AJAX

**Table 8.3** Operation shipping versus data shipping: comparison of computational characteristics of the associated implementations

Characteristic	Operation shipping	Data shipping
Language	Java	JavaScript
Code size	Medium	Small
<i>Runtime</i>		
Feature computation	342 [kB/s]	134 [kB/s]
Classification	<1 [d/ms]	<1 [d/ms]

implementation of adequate counter measures: private queries and search results are sent to a public server, a fact which will never be accepted by the majority of the users.

## 8.6 Conclusion

Web genre analysis has various applications – not only as a filter criterion for Web-based search, but also as preprocessing technology for advanced information extraction and document organization tasks. We use the term “genre retrieval model” as a collective term for the combination of a set of document representations  $\mathbf{D}$  and a classifier  $\gamma$  that maps a document representation  $\mathbf{d} \in \mathbf{D}$  on a set of genre class labels. Most of the existing genre retrieval models exploit high-level features, such

as part-of-speech, tailored text statistics, or information about the document structure. However, aside from the high computational effort a negative consequence is that the resulting genre retrieval models tend to generalize unsatisfactorily.

Especially because of the last point retrieval models for genre did not convince in the Web retrieval practice. Our research addresses this issue by providing a formal means to measure the generalization capability of genre retrieval models. We also propose a feature type which quantifies the concentration of genre-specific core vocabulary in a document, and which has the potential to improve the generalization capability of existing genre retrieval models. Our analysis shows that this new kind of feature class is successful in this respect.

The chapter discussed also software engineering aspects: the authors have developed and compared browser add-ons that implement genre-enabled Web search. Our implementation shows the feasibility of the technology and gives an idea of how genre information can be integrated into standard search technology.

## References

1. Antunes, P., C.J. Costa, and J. Ferreira Dias. 2001. Applying genre analysis to ems design: The example of a small accounting firm. In *Proceedings of the 7th International Workshop on Groupware, CRIWG 2001*, 74–81. Darmstadt: IEEE CS Press.
2. Bay, S.D. 2000. Multivariate discretization of continuous variables for set mining. In *KDD '00: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 315–319, New York, NY: ACM Press. ISBN 1-58113-233-6. doi: <http://doi.acm.org/10.1145/347090.347159>
3. Boese, E.S., and A.E. Howe. 2005. Effects of web document evolution on genre classification. In *Proceedings of the CIKM'05*, Nov 2005. ACM Press.
4. Bretan, I., J. Dewe, A. Hallberg, N. Wolkert, and J. Karlgren. 1998. Web-specific genre visualization. In *Proceedings of the Webnet World Conference on the WWW and Internet*.
5. Broder, A.Z. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36(2):3–10.
6. Crowston, K., and M. Williams. 2000. Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society* 16(3):201–216.
7. Deerwester, S.C., S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6):391–407.
8. Dewdney, N., C. VanEss-Dykema, and R. MacMillan. 2001. The form is the substance: Classification of genres in text. In *Proceedings of ACL Workshop on HumanLanguage Technology and Knowledge Management*. Toulouse, France.
9. Dimitrova, M., A. Finn, N. Kushmerick, and B. Smyth. 2002. Web genre visualization. In *Proceedings of the Conference on Human Factors in Computing Systems*. Minneapolis, Minnesota, USA.
10. Dougherty, J., R. Kohavi, and M. Sahami. Jul 1995. Supervised and unsupervised discretization of continuous features. In *Proceedings of the 12th International Conference on Machine Learning*, eds. A. Prieditis and S. Russell, 194–202, Menlo Park, CA: Morgan Kaufmann.
11. Fayyad, U.M., and K.B. Irani. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the International Joint Conference on Uncertainty in AI (IJCAI)*, 1022–1027. Chambery, France.
12. Finn, A., and N. Kushmerick. 2003. Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Acapulco, Mexico.

13. Freund, L., C.L.A. Clarke, and E.G. Toms. 2006. Towards genre classification for IR in the workplace. In *Proceedings of the 1st International Conference on Information Interaction in Context*, 30–36. New York, NY: ACM Press. ISBN 1-59593-482-0.
14. Karlgren, J., and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th. International Conference on Computational Linguistics, Coling 94*, vol. II, 1071–1075. Kyoto.
15. Kennedy, A., and M. Shepherd. 2005. Automatic identification of home pages on the web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, HICSS-38*. Big Island, Hawaii.
16. Kessler, B., G. Nunberg, and H. Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, eds. P.R. Cohen and W. Wahlster, 32–38. Somerset, NJ: Association for Computational Linguistics.
17. Lawrie, D., W.B. Croft, and A.L. Rosenberg. 2001. Finding topic words for hierarchical summarization. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 9–13 Sept 2001, 349–357. New Orleans, LA.
18. Lawrie, D.J., and W.B. Croft. 2003. Generating hierarchical summaries for web searches. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul 28–Aug 1, 2003, 457–458. Toronto, ON.
19. Lee, Y.-B., and S.H. Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 145–150. ACM Press. ISBN 1-58113-561-0. doi: <http://doi.acm.org/10.1145/564376.564403>
20. Lim, C.S., K.J. Lee, and G.C. Kim. 2005. Automatic genre detection of web documents. In *Proceedings of Natural Language Processing, IJCNLP 2004*, eds. K. Su, J. Tsujii, J. Lee, and O.Y. Kwong, 310–319. Springer.
21. Meyer zu Eiblen, S., and B. Stein. 2004. Genre classification of web pages: User study and feasibility analysis. In *KI 2004: Advances in Artificial Intelligence*, Sept 2004, eds. S. Biundo, T. Frühwirth, and G. Palm, LNAI of Lecture Notes in Artificial Intelligence, vol. 3228, 256–269, New York, NY: Springer. ISBN 0302-9743.
22. Mitchell, T.M. 1997. *Machine learning*. New York, NY: McGraw-Hill Higher Education. ISBN 0070428077.
23. Popescul, A., and L.H. Ungar. Automatic labeling of document clusters. <http://citeseer.nj.nec.com/popescul00automatic.html>, 2000
24. Quinlan, J.R. 1993. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers.
25. Rauber, A., and A. Müller-Kögler. 2001. Integrating automatic genre analysis into digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries*, 1–10. Roanoke, Virginia, USA.
26. Rehm, G. 2002. Towards automatic web genre identification. In *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS'02)*, Jan 2002. IEEE Computer Society.
27. Rendell, L.A. 1986. A general framework for induction and a study of selective induction. *Machine Learning* 1:177–226.
28. Robertson, S.E., and K. Sparck-Jones. 1976. Relevance weighting of search terms. *American Society for Information Science* 27(3):129–146.
29. Roussinov, D., K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu. 2001. Genre based navigation on the web. In *Proceedings of the 34th Hawaii International Conference on System Sciences*. Maui, Hawaii.
30. Salton, G., A. Wong, and C.S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18(11):613–620.
31. Santini, M. 2006. Common criteria for genre classification: Annotation and granularity. In *Proceedings of the ECAI-Workshop TIR-06*. Riva del Garda.

32. Santini, M. 2007. Automatic identification of genre in web pages. PhD thesis, University of Brighton.
33. Sellentin, J. 1999. Konzepte und Techniken der Datenversorgung für komponentenbasierte Informationssysteme. PhD thesis, University of Stuttgart, Stuttgart.
34. Stamatatos, E., N. Fakotakis, and G. Kokkinakis. 2000. Text genre detection using common word frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken.
35. Stein, B., and S. Meyer zu Eißén. 2004. Topic identification: framework and application. In *Proceedings of the 4th International Conference on Knowledge Management (I-KNOW 04)*, Graz, Austria, July 2004, eds. K. Tochtermann and H. Maurer, Journal of Universal Computer Science, 353–360. Graz: Know-Center.
36. Stein, B., and M. Busch. 2005. Density-based cluster algorithms in low-dimensional and high-dimensional applications. In *Proceedings of the 2nd International Workshop on Text-Based Information Retrieval (TIR 05)*, Fachberichte Informatik, Sept 2005, eds. B. Stein and S. Meyer zu Eißén, 45–56. Universität Koblenz-Landau.
37. Stein, B., and S. Meyer zu Eissen. 2006. Distinguishing topic from genre. In *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06)*, Graz, Sept 2006, Journal of Universal Computer Science, eds. K. Tochtermann and H. Maurer, 449–456. Springer.
38. Stein, B., and S. Meyer zu Eißén. 2008. Retrieval models for genre classification. *Scandinavian Journal of Information Systems (SJIS)* 20(1):91–117. ISSN 0905-0167.
39. Turney, P.D. 1995. Technical note: Bias and the quantification of stability. *Machine Learning* 20(1–2):23–33.
40. Utgoff, P.E. 1986. Shift of bias for inductive concept learning. In *Machine learning: An artificial intelligence approach*, eds. R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, vol. II, 107–148. Los Altos, CA: Kaufmann.
41. Yoshioka, T., and G. Herman. Coordinating information using genres. *CCS WP 214*. Cambridge, MA: Massachusetts Institute of Technology (MIT), Sloan School of Management, Aug 2000.



# Chapter 9

## Marrying Relevance and Genre Rankings: An Exploratory Study

Pavel Braslavski

### 9.1 Introduction

Recent years have shown a growing interest to automatic genre analysis of Web documents, especially in the context of Web search. As the amount of indexed documents grows, the specification of a few keywords is not enough to describe user information need. Many studies suggest looking at document genre as an additional non-topical retrieval criterion. The output of a genre classifier could be used in Web search both explicitly and implicitly. Explicit use implies at least three possibilities. First, a focused (“vertical”) search engine (SE) over documents belonging to a certain genre could be built. Second, the user can be given an opportunity to specify the desired genre in the query. Finally, the search engine results page (SERP) can be improved by enriching snippets with genre labels<sup>1</sup> or grouping the documents of the same genre together. However, all three options bring up issues.

If we look at successful vertical search services such as scientific paper search,<sup>2</sup> blog search,<sup>3</sup> news search engines,<sup>4</sup> or product search and comparison services<sup>5</sup> we notice that the task of gathering (or filtering out) content for services does not require especially sophisticated methods. Either the contributors are highly interested in providing their content to the service (scientific papers authors/publishers, on-line merchants), or the content is concentrated on several host sites in a certain form

---

P. Braslavski (✉)

Institute of Engineering Science RAS, 620219 Ekaterinburg, Russia  
e-mail: pb@imach.uran.ru; pb@yandex-team.ru

This paper expands the short paper presented at the workshop “Towards Genre-Enabled Search Engines: The Impact of NLP” [9].

<sup>1</sup> WEGA, a Firefox plug-in (see [31], Chapter 8 by Stein et al., this volume), exemplifies this approach.

<sup>2</sup> <http://citeseer.ist.psu.edu>, <http://scholar.google.com>

<sup>3</sup> <http://technorati.com>, <http://blog.yandex.ru>

<sup>4</sup> <http://news.google.com>, <http://news.yandex.ru>

<sup>5</sup> <http://shopping.yahoo.com>, <http://www.pricegrabber.com>

(like blog services, RSS feeds), or it can be found on the Web using simple surface features with high precision and satisfactory recall (e.g. scientific papers on authors' homepages).

Nowadays a simple search box and a ranked list of search results is a standard de facto for millions of search engine users. So the problem with the genre indicated explicitly in a query is that such advanced search option would be utilized by a marginal share of users.

The use of genre labels in search results presentation is somewhat questionable, too. Experiments [24] have shown that though most users expect genre information to be helpful for their Web search tasks, a straightforward implementation of genre-related hints does not improve user search effectiveness significantly. Moreover users can recognize distinct genres (such as catalog, FAQ, blog, or news) with high accuracy even from ordinary snippets [30].

The main problem in all three cases is to adapt or invent a suitable genre palette that is intuitively clear, complete, and unambiguous for the majority of users. Additionally, an appropriate interface needs to be designed. At present, this is just wishful thinking.

The approaches briefly described above imply an explicit use of genre within SEs. An alternative approach consists in using genre features in static (i.e. query-independent) ranking. Modern machine learning techniques allow for incorporating a wide variety of features to assess page quality regardless of the query (see [23] for an example). The features can be fairly heterogeneous and range from page popularity and pagerank to HTML well-formedness and color palette used. Some genres are less informative and convey mood or emotions rather than facts and information. One can try to incorporate this idea into the ranking scheme through machine learning. Another possible alternative is construction (as opposed to ranking) of SERP from documents of different genres. For example, if enough relevant documents are retrieved, there must be at least a news article, a product page, and a blog presented on the search results page. However, in this case too, one must decide an appropriate genre palette. A much more challenging task is to infer the genre from the query and return the documents of the implied genre to the user.

In this chapter, we suggest an additional alternative by incorporating genre information into relevance ranking.

There are different definitions of *genre* or *style* (we treat these terms interchangeably). Both terms are widely used in linguistics, literary studies, aesthetics, art history and fashion. An extensive overview of different approaches to definition of *genre* can be found in [27]. We treat genre of a text document as a concept opposite to the document topic, similarly to several studies, e.g. [22]. We accept the intuitive understanding that genre is mainly related to the form (*how*) whereas the topic – to the content (*what*) of a document. This simplified approach is justified since we do not perform *genre classification/categorization*.

The idea of the experiment is to make use of a simple continuous measure of document's genre (akin to readability score). The approach is similar to static ranking in the way that we use a query-independent page-level feature in ranking, however we employ a much more straightforward approach – merging ranked documents.

Although genre can be seen as “orthogonal” to topic (i.e. almost all topics can be expressed in different genres), in the framework of our experiment we hypothesize that formal documents are potentially more informative than less formal ones.

We take a third-party system run from the ad hoc retrieval track at the Russian information retrieval evaluation seminar (ROMIP) provided partially with relevance judgments. Then we re-rank the documents according to genre-related score, merge both rankings with different weights, and compare the new ranks with the original ones using relevance judgments. The main unit of the analysis is an individual text-rich document. Due to specifics of the corpus used at the initial stage of the experiment we exploit only textual features in the analysis neglecting Web-specific genres and document features such as HTML markup and structure, and URL tokens.

We conducted our experiments on Russian documents but the methods can be easily applied to a different language.

In the part of genre-related score extraction the study is rooted in our early experiments on genre categorization [7]. In contrast to our previous study on genre “admixture” in ranking [8] when we used an unsupervised approach, our current study employs a supervised method for extracting genre-related scores. The experiment is closely related to recent studies aimed at incorporating non-topical document facets into Web information retrieval (see Section 9.2).

The chapter is organized as follows. The next section surveys related work in the fields of genre classification, readability analysis, and previous experiments on incorporating genres into relevance ranking. Section 9.3 describes the data used in the experiment. Section 9.4 describes the extraction method and obtained formality score. Section 9.5 presents the produced rankings, evaluation metrics and final results. Section 9.6 concludes and outlines directions for future research.

## 9.2 Related Work

Our work is related to research in the fields of automatic genre classification, readability as well as information retrieval experiments on integrating genre-related features into relevance ranking schemata.

### 9.2.1 Genre Classification

After the pioneering work by Karlgren and Cutting [14] many papers on automatic genre classification have been published. The majority of the papers address the genre categorization problem and solve it using machine learning techniques. Set and structure of genre categories, corresponding learning sample, classification features, as well as learning technique constitute the diversity of the approaches. There are different sets of genres employed in the studies. The number of distinct genres ranges from binary classes (e.g. informative/imaginative; textual/non-textual) up to 16 multiple genre classes. Many researchers propose a hierarchical structure of genres. Some of them borrow an established set of genres, the others compile a

genre palette based on a user study. The variety of classification techniques includes discriminant analysis, naïve Bayes, logistic regression, neural networks, kNN, and SVM. A number of studies utilize existing corpora for learning and evaluation, the rest compile their own. As opposed to topical text categorization most genre categorization studies use mainly non-content features such as surface text statistics, function words count, POS and punctuation mark frequencies, etc. For more details one can refer to a comprehensive survey of the field [26]. At least two noteworthy papers appeared after the survey that primarily concentrate on analysis of Web documents. Meyer zu Eissen and Stein [19] conducted a user study spawning a set of eight Web genres useful for Web search, and built a corpus containing these genres (the KI-04 corpus). Along with linguistic features traditionally used in genre analysis, their study employs HTML-based features. Lim et al. [17] expanded this approach even further and made use of a wider range of features (326 in total), including various surface, lexical, syntactic, HTML, and URL features.

Automatic genre analysis is not restricted to genre categorization – there are some efforts on genre clustering. Rauber and Müller-Kögler [22] adapted an unsupervised technique for revealing genre-dependent similarities between documents. The self-organizing map (SOM) was used to cluster documents according to their various surface level text features. The results of analysis were incorporated into a content-based representation through coloring individual documents according to their location on the resulting SOM. Gupta et al. [13] applied the notion of Web site genre to improve web page cleansing methods (i.e. removal of ads, unnecessary images and extraneous links). Sites are clustered in word feature space using city-block distance. The distinction of the method is that sites are characterized not only by the words they contain but also by the words from snippets returned by several SEs in response to the web site domain name.

## 9.2.2 *Readability Scores*

Research on readability has its roots in psycholinguistics but in fact is very similar to automatic genre analysis. The aim is to obtain a simple measure to compare the comprehension complexity of texts conveying similar meaning using surface cues [11].

The “traditional” way to construct a readability formula is as follows. First, text complexity estimates are obtained experimentally. Second, text features that potentially contribute to its complexity are extracted. Third, text features and text complexity are tied together using regression analysis. There are different psycholinguistic techniques to measure text complexity: reading time (normalized by the individual reading skills), post-reading questionnaires assessing text comprehension, and cloze tests. There are different features used in readability formulæ: number of words from different word lists (such as “easy”, “hard”, “abstract”, “most frequent”, etc. word lists), word length, sentence length, number of sentences per paragraph, number of prepositional phrases, etc. In summary, all the features can be divided in two classes: semantic features reflecting the complexity of vocabulary

and structural features reflecting compositional complexity (usually on the sentence level, sometimes on the paragraph level).

There have been recent papers introducing a novel approach to readability that is very close to ours.

Si and Callan [28] and Collins-Thompson and Callan [10] re-formulate the readability prediction task as a categorization problem: they use labeled data (documents with assigned readability labels), tokens as features, and naïve Bayes classifier. Their approach emphasizes semantic features, i.e. difficulty of a text is defined entirely through its vocabulary. The method outperforms traditional readability measures on Web data.

A related study is described in [16]: a query-independent “familiarity classifier” is built upon several hundreds of documents manually tagged as “introductory” or “advanced” using random forest classifier. Three groups of feature are employed: (1) stop-words, (2) common readability features and traditional readability scores themselves, (3) features based on various characteristics of web page documents (e.g. anchor text count or Similarity of WordNet expansion of top 10% of document with remaining 90%). The authors show that traditional readability measures such as Fog index, Flesch readability score, and Flesch-Kincaid grade level capture the notion of familiarity poorly. However, the method does not consider topic relevance: top-20 documents returned by a search engine are all assumed to be relevant to the query, which seems to be a very strong assumption.

### 9.2.3 Genres in Relevance Ranking

Strzalkowski et al. performed stylistic analysis on TREC data already in 1995 [29]. Their idea was to find stylistic features that could discriminate relevant and non-relevant documents. Using previous TREC results, they found that relevant documents tend to be more complex on different levels – textual, syntactic, and lexical. A decision tree classifier was built upon labeled data, documents classified as non-relevant were to move to the end of the list. However this strategy did not gain in average precision: “The consequence is that to make use of stylistic variation for reliable relevance grading we need a query typology: each query must be identified for likely style preferences” [29]. As a matter of fact, our study reasserts these findings.

A High Accuracy Retrieval from Documents (HARD) track was organized within TREC campaign in 2003–2005 [3, 4]. The goal of the track was “to bring the user out of hiding, making him or her an integral part of both the search process and the evaluation” [3] as opposed to an abstract “average” user behind traditional TREC topics. TREC topics were provided with metadata including GENRE and FAMILIARITY items. In particular, in HARD 2004 track GENRE had values of *news-report*, *opinion-editorial*, *other*, or *any*; FAMILIARITY had a value of *little* or *much*. Within HARD track RELEVANT judgment means that the document is on topic *and* it satisfies the appropriate metadata. Attempts to utilize the available metadata, including GENRE and FAMILIARITY are exemplified by track reports

[1, 6]. Belkin et al. [6] used readability scores, average number of syllables per word, and abstractness/concreteness of the document’s vocabulary to model familiarity. Genres were modeled by language models; the Kullback-Leibler (KL) divergence determined whether a document was a member of the genre. Final rankings were obtained via weighted combination of baseline scores and metadata classifiers’ scores. Both genre and familiarity classifiers performed poorly. As the authors stated, “using language models to capture genre preference was a complete failure, presumably because the language models captured the topics of the training documents.”

Abdul-Jaleel et al. [1] were more successful at building genre classifier. They used linear SVM and 10K most frequent tokens in the corpus, subcollection tags, and various length measures of a document as features. Final rankings were produced by linear combination of the normalized outputs of both the retrieval and classifier outputs. Although genre classifier showed good performance, it did not leverage the retrieval effectiveness. Authors noticed that “many documents judged relevant clearly fall outside the requested metadata. Searchers know a relevant document when they see one, but a priori they do not fully know what metadata is required of a relevant document.”

In the following sections, we will describe experiments that complement the approaches described above.

## 9.3 Data

In this study we use two datasets of Russian documents: (1) a small corpus of five functional styles as a learning sample for extracting a genre-related score and (2) a subset of reference ROMIP Web collection for experimentation and evaluation purposes.

### 9.3.1 *Functional Styles Sample*

For our experiment we needed a simple measure that captured the formality or “seriousness” of the document akin to a text readability measure. Unfortunately, there is no widely accepted and use-proven readability index for Russian. For the purpose of obtaining a genre-related score we reused a functional styles sample that was employed in our previous experiments. The sample contains 50 federal acts (official functional style), 54 scientific papers in natural sciences (academic style), 61 online news articles (journalistic style), 79 short stories by modern Russian authors (literary style), and 61 fragments of online chats (everyday communication style) – 305 documents in Russian in total.

It is important to stress that our study is not aimed at building a functional styles classifier. The assumption is that formality progressively decreases from federal acts to chats, being federal act the most formal genre and chat the least formal.

### 9.3.2 ROMIP Collection

ROMIP stands for Russian Information Retrieval Evaluation Seminar which is a Russian TREC-like information retrieval evaluation initiative [25]. ROMIP Web collection contains about 600,000 HTML pages in Russian from the free Web hosting `narod.ru` and adequately reflects the diversity of Web genres. The collection is used in the ROMIP ad hoc retrieval track and is freely available upon request.

Along the documents the collection contains a list of about 20,000 queries taken from a real-life Web SE query log. Each participating system performs the whole set of queries over ROMIP collection. A small selection of queries (or topics in TREC terminology) is evaluated manually using a pooling method in each yearly cycle. A short description (close to TREC’s *narrative*) representing one of the possible query interpretations is provided to help assessors (Fig. 9.1). Many descriptions imply detailed and informative documents. This fact suggests that ranking “serious” documents higher may improve the overall search quality within the ROMIP framework. We implement this approach in our experimental framework, however it will not comply with all real-life information needs obviously.

In our previous stylistic experiments [8] we found out that menus, navigation, ads, authorship and copyright notices, etc. presented on the majority of HTML pages in the ROMIP collection significantly skew genre-related parameters. So we took the collection after template removal routine described in [2]. However, the difference from the original collection was not substantial since the ROMIP collection is compiled from free hosting pages and includes mainly sites with moderate number of pages which makes proper template detection and removal difficult. All documents were converted to Windows-1251 Cyrillic encoding and subsequently to *plain text*.

For our experiment, we took the results of one of the ROMIP’2006 participating systems which utilizes only text relevance features [12]:

- single query terms match;
- pairs of query terms match;

<p><i>Query arw13494: memory training</i>  <i>Description</i> Documents containing advice for human memory improvement, diverse techniques for memory training. Documents containing recipes of food supplements are useful. Especially important are documents containing detailed and precise instructions for those who want to train their memory.</p> <p><i>Query arw19003: are we alone in the universe?</i>  <i>Description:</i> The page must contain information on extraterrestrial intelligence research, existing hypotheses as well as different opinions on this issue.</p> <p><i>Query arw18885: why do the airplanes fly</i>  <i>Description</i> The page must contain information about airplanes, aerodynamics basics, wing lift.</p>
---

**Fig. 9.1** Sample ROMIP topics: query and its description (originally in Russian, descriptions are used on the evaluation stage only)

- exact phrasal match;
- all query terms appear in the document;
- a significant part of query terms appears within a sentence.

Additionally, pseudo-relevance feedback techniques were used. The system was trained on relevance judgments of two previous campaigns – ROMIP’2004 and ROMIP’2005.

This ROMIP subset contains 6,906 documents corresponding to 70 evaluated search topics (67 topics with 100 ranked documents per topic plus three topics with 23, 87, and 96 documents, respectively). The majority of these documents have binary relevance judgments: 5,393 documents (420 relevant + 4,973 non-relevant) with so-called “strong” judgments (i.e. all assessors agreed on judgment) and 5416 documents (1,105 relevant + 4,311 non-relevant) with “weak” relevance judgments (i.e. at least one assessor judged a document as “relevant”). Some topics have no corresponding relevant documents (13 in case of “strong” relevance and three in case of “weak” relevance). The rest of the documents have tag “can’t be judged” or do not fall into the evaluated document pool. The pool depth in ROMIP’2006 was 50, i.e. the first 50 documents from the participating systems’ runs were pooled and evaluated. At the 50 cut-off the statistics of the subset looks as follows: 3,473 documents in total, including 354 and 899 relevant documents (strong and weak judgments, respectively); topics with zero relevant documents – 15 and 4 (strong and weak judgments, respectively).

## 9.4 Formality Score

As we mentioned before, there is no widely accepted and use-proven readability score for Russian that would be appropriate for our aims. So we opted for building a “formality score” based on our previous research.

In our earlier experiments on genre categorization [7] we employed the concept of functional styles, which is well-established in Russian linguistics. There are five basic functional styles: *official*, *academic*, *journalistic*, *literary (fiction, belles-lettres)*, and *everyday communication style*. Functional styles have been the subject of an study on automatic stylistic analysis [20]. More details on the theory of functional styles can be found in [15].

Our approach is rather operational. We consider five functional styles simply as text classes of gradually decreasing formality. We use this small sample only for building a genre-related score and then “throw this ladder away after climbed up it”. The quantitative characteristics of the functional styles sample confirm appropriateness of the approach (Fig. 9.2). Such features as average word length (one of the most commonly used features in different readability formulae) and POS distribution change monotonically over five styles.

We use *canonical discriminant analysis* to extract the formality score. The method is illustrated in Fig. 9.3: feature space transformation is performed in order to find a direction (a weighted sum of initial features) with the best separating ability

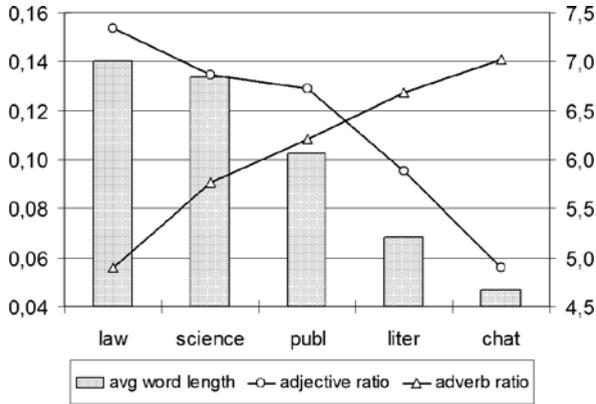


Fig. 9.2 Selected characteristics of the functional styles sample

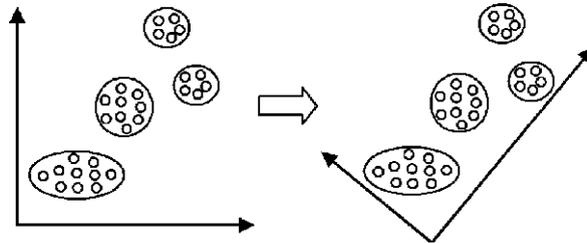


Fig. 9.3 The idea of canonical discriminant analysis

between classes. The method is similar to *principal components analysis* (PCA); the difference is that class structure is taken into account. After this point we abandon the set of discrete genres and proceed to an continuous index.

We experimented with different genre-related easily computable textual features. As mentioned before we use exclusively real-value textual features implying relatively long (since all features are averages) and coherent text documents. The features we used are quite similar to those used in our previous experiments and other genre analysis studies: surface features like word and sentence length (the latter is based on a simple rule for sentence boundary detection), punctuation and functional words counts, and POS ratios (using *mystem* POS tagger [21] without any disambiguation). Feature selection process was guided by the percentage of explained variance, analysis of variance over five classes, as well as considerations on feature semantics. After a series of trials we opted for a combination of nine features. The formula for the first canonical root that we treat as formality score is as follows (standardized values, greater values correspond to lesser formality):

$$S = -0.49x_1 + 0.27x_2 + 0.46x_3 + 0.04x_4 + 0.24x_5 + 0.32x_6 - 0.48x_7 + 0.32x_8 - 0.11x_9,$$

where

- $x_1$  – average word length;
- $x_2$  – smiley count;
- $x_3$  – finite verb count;
- $x_4$  – adjective count;
- $x_5$  – first person pronoun count;
- $x_6$  – expressive punctuation count;
- $x_7$  – neuter noun count;
- $x_8$  – adverb count;
- $x_9$  – genitive chain count.

The first canonical root explains 84% of sample’s variance. Fig. 9.4 shows that although the classes are not smoothly separable in this 2D space, they line up along X axis, preserving their “formality order” in general.

The obtained index is fairly similar to a readability score: average word length, a component of almost all readability measures, enters into the formula with negative weight, the same way as genitive chain count (reflects syntactic complexity) and neuter noun count (neuter nouns tend to be more abstract in Russian). In contrast smiley, expressive punctuation and first person counts enter into the formula with positive weights, reflecting text informal flavor. For convenience we mapped the obtained canonical root onto [0, 1] interval with lesser values corresponding to lesser formality.

The applied corpus-based approach is low-cost, flexible, and easily adjustable compared to traditional methods for building readability scores based on reading

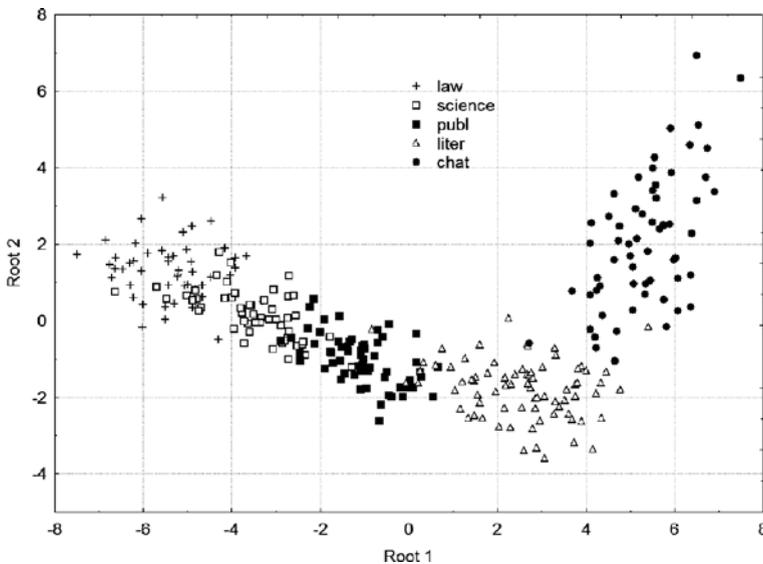


Fig. 9.4 Scatter-plot of the learning sample in the 1st and 2nd canonical roots

tests. A thorough examination of the obtained index and comparison with the existing readability indices will be addressed in a separate study.

## 9.5 Results

### 9.5.1 Genre-Related Rankings

We calculated formality scores for documents in our ROMIP subset. We performed a selective comparison of documents and can estimate that obtained index reflects formality perception accurately. Relevant documents appeared to be somewhat “more formal”: averaged formality scores for our ROMIP subset are 0.62 and 0.59 for relevant and non-relevant documents, respectively (the difference is significant at  $p < 0.005$ ). Distribution of formality score values over Web documents sample is presented on Fig. 9.5. One can see that distribution is fairly smooth, “neutral” documents constitute the majority of the sample.

The obtained formality score similarly to readability indices implies coherent text. There are many types of non-textual web documents such as link and price lists, input forms, photo galleries, home pages with predominantly presentational content, etc. In order to filter out such documents as far as possible using simple methods, we introduced two restrictions for documents to be re-ranked: (1) longer than five sentences and (2) finite verb/sentence ratio greater than threshold (a simple signal of text coherence, threshold is selected empirically).

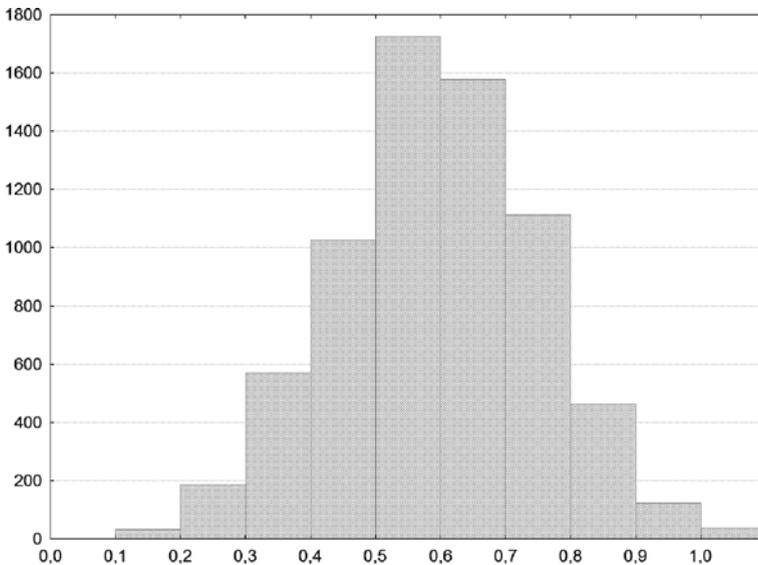


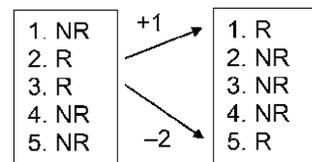
Fig. 9.5 Distribution of mapped formality values over ROMIP sample (6,848 documents)

All evaluated documents meeting the restrictions have been ranked according to the genre-related score in descending order within each topic (more ‘formal’ documents on the top); all other documents preserved their initial positions. We obtained four initial genre-related rankings:

1. *T100*: all documents longer than five sentences (4,846 documents processed);
2. *T100C*: additionally, finite verb/sentence ratio  $\geq 0.6$  (3,823 documents processed);
3. *T50*: top-50 documents in each topic longer than five sentences (3,030 documents processed);
4. *T50C*: additionally, finite verb/sentence ratio  $\geq 0.6$  (2,332 documents processed).

In the next step, we aggregated the obtained genre-related ranks ( $R_G$ ) with the initial keyword-relevance ranks ( $R_Y$ ) (see [12] for details on  $R_Y$ ). We used a straightforward approach to aggregation: new rank was computed based on a linear combination of text relevance and genre-related ranks, i.e.  $R_Y + \alpha R_G$  ( $\alpha$  is weight of genre-related ranking,  $\alpha \in [0, 1]$ ). This scheme can be referred to as a simple case of weighted Borda method that is widely used in different areas, including rank aggregation for metasearch. It is important to note that we did not aim at finding an optimal  $\alpha$  for the rank combination. Although the number of processed documents is appreciable, the number of topics with relevant documents does not allow us to test our results properly and generalize well. The proposed re-ranking method is fairly conservative. Apart from the fact that many short and presumably incoherent documents preserve their positions since we are not confident enough to assign them a formality score, small  $\alpha$  values prevent documents from distant jumps.

For evaluation of the aggregated ranks we use *rank displacement of relevant documents* ( $D_R$ ) – a metric introduced in [5] for evaluation of data fusion effects in information retrieval (Fig. 9.6).  $D_R$  sums the ups and downs of relevant documents in the new list in comparison to the original one. Note that small movements in the top of the list “cost” the same as in the bottom. Furthermore, we count up absolute number of tasks with positive and negative values of  $D_R$ . Additionally, we use official ROMIP metrics: *mean average precision* (*MAP*, calculated for the top-50 documents), *p1*, and *p10* (precision at levels 1 and 10, respectively). Note, that average precision (*AP*) is highly sensitive to ranking of relevant documents in contrast to  $D_R$ , thus little movements of relevant documents in the bottom of the ranked list have almost no effect on this metric; conversely small drops of relevant documents in the top of the list impair the metric value significantly.



**Fig. 9.6** Rank displacement of relevant (*R*) documents (for this example  $D_R = -1$ )

### 9.5.2 Merged Rankings

The most illustrative results are obtained on weak relevance judgments. Figs. 9.7, 9.8, 9.9 and 9.10 show both macro- and micro-averaged  $D_R$  values and absolute numbers of topics with positive vs. negative changes depending on genre-related rank's weight for T100, T100C, T50, and T50C rankings. Standard ROMIP metrics for T100C and T50C rankings are shown in Fig. 9.11.

As one can see a small admixture of genre-related scores can slightly improve relevance ranking in terms of  $D_R$  metric. As Fig. 9.8 shows, in the best case approximately every second relevant document in each topic climbs one position higher in average. A simple criteria for text coherence based on finite verb ratio increases maximum macro-averaged  $D_R$  and broadens the range of its positive values and at the same time flattens the difference between topics with positive and negative effects. In case of the pool-deep re-ranking (T50C) the use of this criteria keeps the macro-averaged  $D_R$  values in the positive half-plane and positive changes majorize negative changes at topic level.

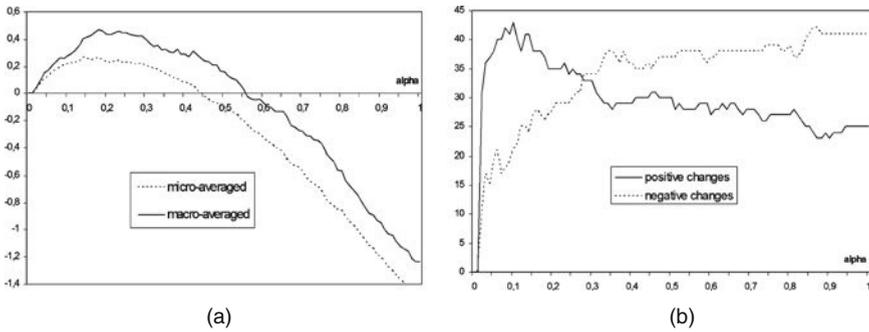


Fig. 9.7 T100 re-ranking results: **a** averaged rank displacement; **b** number of tasks with positive and negative  $D_R$

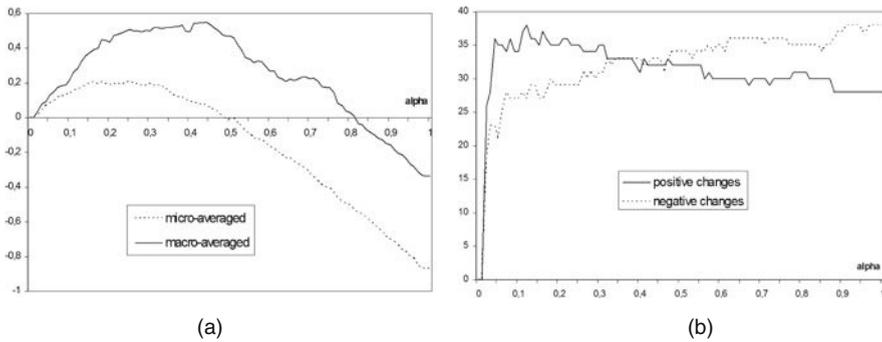


Fig. 9.8 T100C re-ranking results: **a** averaged rank displacement; **b** number of tasks with positive and negative  $D_R$

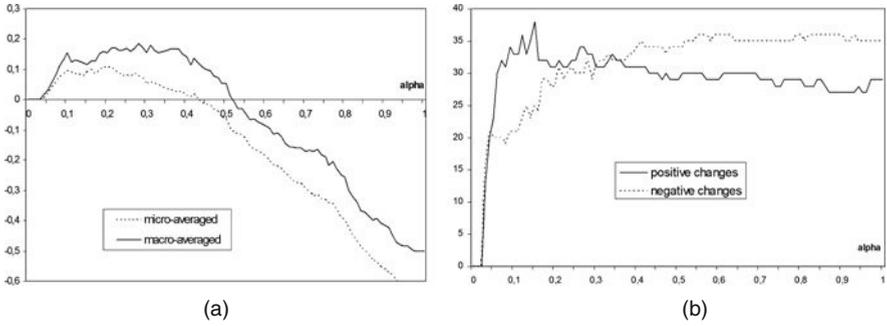


Fig. 9.9 T50 re-ranking results: **a** averaged rank displacement; **b** number of tasks with positive and negative  $D_R$

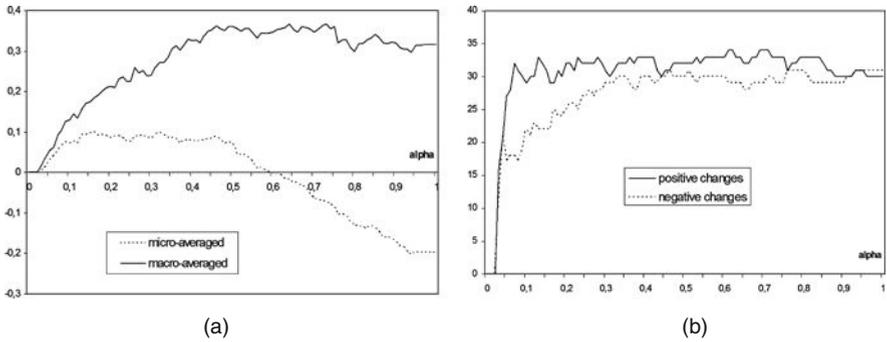


Fig. 9.10 T50C re-ranking results: **a** averaged rank displacement; **b** number of tasks with positive and negative  $D_R$

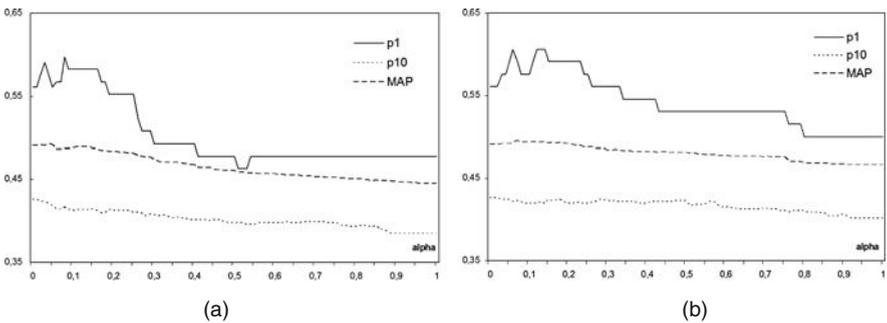
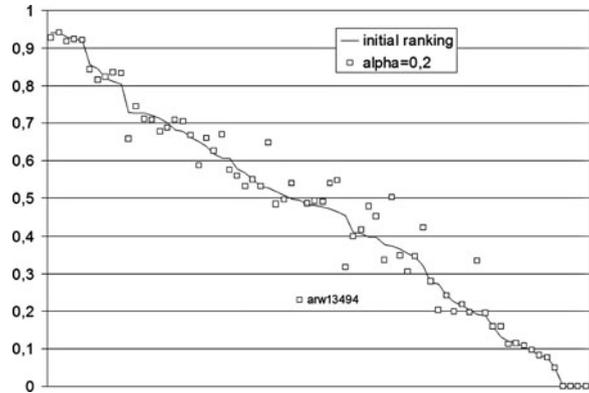


Fig. 9.11 Standard ROMIP metrics: **a** T100C; **b** T50C

However, these positive effects are not reflected in the standard ROMIP metrics except for an insignificant growth of  $MAP$  (less than 1%, Fig. 9.11b) and some occasional splashes on  $p1$  plot (Fig. 9.11). Figure 9.12 illustrates difference in average precision between initial ranking ( $\alpha = 0$ ) and merged one ( $\alpha = 0.2$ ) by topic (the outlying topic is *arw13494: memory training*).

**Fig. 9.12** Initial vs. new ( $\alpha = 0.2$ ) average precision (T50C), topics sorted by initial AP



If we take a look at individual topics, we can find approximately 25 of them that are responsive to mixing genre ranks with traditional keyword-relevance ranks in almost every proportion. The examples are (originally in Russian):

- *arw17563: what to feed a cat on*
- *arw2000: meals in the fast*
- *arw5608: quantum computer*
- *arw10947: tv commercial creation*
- *arw2755: all about al capone*

We were unable to find a reliable pattern for these topics based on mean and standard deviation of the formality score, number of relevant documents, etc. According to the subjective observation descriptions of these topics might represent a more rigorous interpretation within ROMIP evaluation than a common one. But at the same time, mean formality score designates “serious” topics with confidence. For example, these five topics with maximum mean formality scores consist mainly of legal, financial, medical, and popular scientific documents (originally in Russian):

- *arw12162: contract-based [military] service*
- *arw2538: magnetic field effects on humans*
- *arw18557: harmful effects of polluted air on respiratory apparatus*
- *arw16263: what is a promissory note*
- *arw7927: national income*

### 9.6 Conclusion

In this chapter we investigated different options for using genre-related information in Web search and consider the implicit use of such information most promising.

We conducted an experiment on merging genre-related and text-relevance rankings using reference ROMIP Web collection. To this end we proposed a method for automatic extraction of formality score using canonical discriminant analysis

applied to a small sample of functional styles. Evaluation of the aggregated ranks shows that we can achieve moderate improvements on our experimental data set in average by mixing in a small fraction of genre-related rank. Notably, there is a subset of queries that is quite responsive to mixing genre ranks with traditional keyword-relevance ranks. These findings confirm previous results on incorporating genre information into relevance ranking.

Our study suggests that a promising direction for future research could be incorporating genre information into static ranking. To this end a comprehensive study of distinctive genres' usefulness has to be carried out.

Another possible direction could be inferring of the expected genre (or *genre range* when thinking of continuous genre index) of the answer based on query processing. To the best of our knowledge the sole study on predicting user's education level based on a query is paper by Liu et al. [18]. The study demonstrates good quality in classifying queries according to student grade. The approach uses SVM and various features derived solely from queries, including sentence and word length features, percentage of part-of-speech tags, various readability indices, as well as frequency of numerous 1-, 2-, and 3-word sequences. Yet the paper deals with natural language questions rather than real Web SE queries and the problem remains open. A more reliable way could be click data analysis for frequent queries in order to estimate most expected document genres for those queries.

A further option could be accounting for genres in the personalized search framework. The problem is that a user's genre expectations vary from topic to topic, and drift unevenly with time.

**Acknowledgments** We would like to thank Mikhail Ageev and Andrei Tselishchev for their help with data processing. We also thank Yandex ([www.yandex.ru](http://www.yandex.ru)) for providing us with the experimental data. Many thanks to Matthew McCool and volume editors for their valuable comments on the draft.

## References

1. Abdul-Jaleel, N., J. Allan, W.B. Croft, F. Diaz, L. Larkey, X. Li, M.D. Smucker, and C. Wade. 2005. UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC 2004*.
2. Ageev, M., I. Vershinnikov, and B. Dobrov. 2005. Extraction of the significant part of web pages for information retrieval (in Russian) [Izvlечение značimoj informacii iz web-stranic dlja zada informacionnogo poiska]. In *Internet-Matematika*, 283–301. Available online: [http://company.yandex.ru/grant/2005/07\\_Ageev\\_102942.pdf](http://company.yandex.ru/grant/2005/07_Ageev_102942.pdf)
3. Allan, J. 2004. HARD track overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of TREC-2003*, 24–37.
4. Allan, J. 2005. HARD track overview in TREC 2004: High accuracy retrieval from documents. In *Proceedings of TREC-2004*, 25–35.
5. Beitzel, S.M., E.C. Jensen, A. Chowdhury, D. Grossman, O. Frieder, and N. Goharian. 2004. Fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology (JASIST)* 55(10):859–868.
6. Belkin, N., I. Chaleva, M. Cole, Y.-L. Li, L. Liu, Y.-H. Liu, G. Muresan, C. Smith, Y. Sun, X.-J. Yuan, and X.-M. Zhang. 2005. Rutgers' HARD track experiences at TREC 2004. In: *Proceedings of TREC-2004*.

7. Braslavski, P. 2004. Document style recognition using shallow statistical analysis. In *Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, 1–9. Nancy. Available online: <http://esslli2004.loria.fr/content/readers/36.pdf>
8. Braslavski, P., and A. Tselishchev. 2005. Style-dependent document ranking. In: *Proceedings of the 7th Russian Conference on Digital Libraries (RCDL'2005)*, 159–164. Available online: [http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7\\_1\\_paper.pdf](http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7_1_paper.pdf)
9. Braslavski, P. 2007. Combining relevance and genre-related rankings: An Exploratory Study. In *Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP"*, 1–4, Borovets, Bulgaria. Available online: <http://kansas.ru/pb/paper/ranlp2007.pdf>
10. Collins-Thompson, K., and J.P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL*, 193–200.
11. DuBay, W.H. 2004. The principles of readability. Available online: <http://www.nald.ca/fulltext/readab/readab.pdf>
12. Gulin, A., M. Maslov, and I. Segalovich. 2006. Yandex' algorithm for text relevance ranking at ROMIP'2006 (in Russian) [Algoritm tekstovogo ran'zirovaniya Jandeksa na ROMIP'2006]. In *Proceedings of ROMIP'2006*, 40–51. Suzdal. Available online: [http://www.romip.ru/romip2006/03\\_yandex.pdf](http://www.romip.ru/romip2006/03_yandex.pdf)
13. Gupta, S., G. Kaiser, S. Stolfo, and H. Becker. 2005. Genre classification of websites using search engine snippets. In *Proceedings of SIGIR'2005 Workshop "Stylistic Analysis of Text For Information Access"*. Salvador, Bahia.
14. Karlgren, J., and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, 1071–1075.
15. Kožina, M.N. 1968. Foundations of the functional stylistics (in Russian) [K osnovaniyam funkcional'noi stilistiki], Perm.
16. Kumaran, G., R. Jones, and Madani, O. 2005. Biasing web search results for topic familiarity. In *Proceedings of CIKM'05*, 271–272.
17. Lim, C.S., K.J. Lee, and G.C. Kim. 2005. Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management* 41:1263–1276.
18. Liu, X., W.B. Croft, P. Oh, and D. Hart. 2004. Automatic recognition of reading levels from user queries. In *Proceedings of SIGIR'2004*, 548–549.
19. Meyer zu Eissen, S., and B. Stein. 2004. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004)*, 256–269. Ulm.
20. Michos, S., E. Stamatatos, N. Fakotakis, G. Kokkinakis. 1996. Categorizing texts by using a three level functional style description. In *Artificial intelligence: Methodology, systems, applications, frontiers in artificial intelligence and applications*, ed. A.M. Rasmay, vol. 35. Available online: <http://slt.wcl.ee.upatras.gr/papers/michos2.pdf>
21. Mystem Tool. <http://company.yandex.ru/technology/mystem/>
22. Rauber, A., and A. Müller-Kögler. 2001. Integrating automatic genre analysis into digital libraries. In *Proceedings of the JCDL'2001*, 1–10.
23. Richardson, M., A. Prakash, and E. Brill. 2006. Beyond PageRank: Machine learning for static ranking. In *Proceedings of WWW'2006*, 707–715.
24. Rosso, M.A. 2005. Using genre to improve web search. PhD thesis, University of North Carolina, Chapel Hill, NC.
25. Russian Information Retrieval Evaluation Seminar (ROMIP). <http://romip.ru>
26. Santini, M. 2004. State-of-the-art on automatic genre identification. Technical Report ITRI-04-03, Information Technology Research Institute, University of Brighton, Brighton. Available online: <ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf>
27. Santini, M. 2007. Automatic identification of genre in web pages. PhD thesis, University of Brighton, Brighton.
28. Si, L., and J. Callan. 2001. A statistical model for scientific readability. In *Proceedings of CIKM'2001*, 574–576.

29. Strzalkowski, T., L. Guthrie, J. Karlgren, J. Leistensnider, F. Lin, J. Perez-Carballo, T. Straszheim, J. Wang, and J. Wilding. 1996. Natural language information retrieval: TREC-5 Report. In *Proceedings of TREC'1995*.
30. Stubbe, A., C. Ringlstetter, and R. Goebel, R. 2007. Elements of a learning interface for genre qualified search. In *Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP"*, 21–28. Borovets, Bulgaria.
31. WEGA: Web Genre Analysis Project. <http://www.uni-weimar.de/cms/medien/webis/research/projects/wega.html>



**Part IV**  
**Structure-Oriented Models of Web Genres**

# Chapter 10

## Classification of Web Sites at Super-Genre Level

Christoph Lindemann and Lars Littig

### 10.1 Introduction

The World Wide Web has developed into a central source of information, a very important marketplace, a highly noticed presentation platform, and a frequented meeting place, to mention only some. Furthermore, the ever-growing number of users and content creators leads to a rapid evolution and emergence of different Web sites. As a consequence, it is more and more difficult to identify the Web sites providing the information and services of interest.

However, while Web sites differ in their design and content, many Web sites are created for the same purpose so that they are related like the Web sites of two universities or two competing corporations. This observation directly corresponds to our notion of genre since we think that a genre is defined as a category assigned on the basis of external criteria such as purpose [2, 22]. As a consequence, classification into genres has to focus on the purpose the unit of analysis is created for as central point of investigation. While the concept of genre is often associated with a grouping of texts based on external criteria due to early and important work in this field of research, e.g. [2], we deal with the concept of Web genre. This concept transfers the idea of classification by purpose from texts to the Web introducing new opportunities and challenges. Thus, we have to transfer and extend appropriate methods from the field of text classification in order to embrace these opportunities and to face these challenges. The latter comprise in first place the heterogeneity of the Web and the emergence of new categories like blogs which cannot be found in traditional genres that are usually instantiated on paper. Opportunities arise especially from the exploitation of the link structure and other structural features which are useful for the classification of these new categories. Therefore, we analyze the structure of Web sites in order to examine how it reflects the purpose a Web site is created for. We also show that it is necessary to take content-related features that also reflect this purpose into account. Consequently, a Web site can be classified into a Web genre based on structure and content.

---

C. Lindemann (✉)

Department of Computer Science, University of Leipzig, Leipzig, Germany  
e-mail: cl@rvs.informatik.uni-leipzig.de

The concept of genre is in general very useful in the context of classification tasks because it facilitates the identification of categories. Thus, categories can be identified by considering which purposes can be pursued or are dominant in a certain area. The granularity of the unit of analysis and the genre granularity are other important aspects of consideration. Analysis at the level of Web sites is in general attracting more and more interest [20] for a variety of techniques like spam and duplicate detection. One reason for this trend is based on the fact that there is a more frequent change of information and availability on page-level compared to site-level data so that the latter constitutes a solid foundation for research in several domains. Therefore, a complete Web site is the unit of analysis of our work. Considering the genre granularity, Web genres can be accounted for at subgenre, genre and super-genre level. In our opinion, the super-genre level relates to coarse-grained, general, and dominant purposes and consequently broad categories which comprise several fine-grained sub-categories. Since we focus on dominant purposes Web sites can be created for, we account for the Web genre of a Web site at a super-genre level, i.e. we assign a Web site to one of eight Web genres, namely *Academic*, *Blog*, *Community*, *Corporate*, *Information*, *Nonprofit*, *Personal*, and *Shop*. Another interesting approach for the identification of a compact system of genres is given in [27]. In this study, genres are not identified by considering the purpose a Web site is created for but by analyzing the major aims of text production. Nevertheless, the adapted typology, which includes six categories, exhibits close relations to the Web genres analyzed in this chapter.

The ability to identify to which of the considered Web genres a Web site belongs clearly improves the capability of search engines to present high quality search results. Yahoo! Mindset [32], a former research project, distinguished between search results from Web sites of two different Web genres, i.e. between Web sites with a more commercial or more informational background. An adaptive ranking could be created by favoring those results that correspond to the intent behind the search. This is especially useful for ambiguous queries, e.g. single-word queries, and very popular keywords. While this simple application already motivates the consideration of Web genres at super-genre level, further applications that arise from the classification of Web sites at this level including personalized ranking, the ability to present widespread results, and site tagging strengthen this motivation. A personalized ranking can be performed by favoring results of a certain Web genre that is of interest to a user with respect to her search history, e.g. a user that has frequently clicked on results related to academic Web sites is likely to conceive results of this Web genre among the top hits in future search queries. Widespread results can be presented by including results from several Web genres in the top 10 hits in case of ambiguous queries. Site tagging allows for labeling the results with the corresponding Web genre so that the user can easily choose from the results. Besides the discussion of existing and new technology for the construction of Web genre retrieval models, implementation aspects for a genre-enabled Web search including this labeling of search results on a per page basis are outlined in [28].

These examples show that the combination of relevance and genre rankings is promising for the improvement of search results. Several options regarding the

utilization of genre-related information in Web search are also discussed in [4]. Another application that benefits from Web site classification at super-genre level is the creation of inverted indexes by selecting Web sites of a certain Web genre. Furthermore, the change ratio of a Web page correlates with the genre of its corresponding Web site [7], e.g. Web pages of corporate Web sites change more frequently than academic Web pages. As a consequence, the freshness of an index can be increased by adjusting the refreshing strategy in accordance to these genre-related observations. All applications described above can be used to improve the search quality and therefore constitute the motivation for Web site classification at super-genre level. In addition to this, genre considerations at super-genre level establish the basis for analysis at subgenre level. An exploratory empirical investigation of genre connectivity in an academic Web space is outlined in [3]. This investigation considers fine-grained genres of academic Web pages, which are divided into the two main categories personal and institutional pages. However, resource discovery in the sense of identifying academic Web sites in order to analyze the related Web pages in the presented way is a prerequisite especially for large-scale studies. Therefore, any approach focusing on one of the eight examined Web genres at subgenre level can be considered as an extension and continuation of our studies.

In this chapter, we present an approach for the classification of Web sites at super-genre level. This approach combines the utilization of structure and content of Web sites in order to classify them into one of eight relevant Web genres. While the analysis of structure and content of Web sites has been proven to be appropriate in many research areas before, e.g. for fine-grained classification [14], for spam-detection [15] and Web usage mining [8], the main contributions of our work are: Firstly, we derive a rich set of effective features for classification focusing on size, organization, composition of URLs, technical realization, and link structure of a Web site. Secondly, we analyze structural and content-based features in a large measurement-based study to reveal the strengths and weaknesses of classification solely based on structure or on content. Thirdly, we show how a content-based classification with standard techniques can be effectively combined with a structure-based classification of Web sites into three sets of aggregated Web genres to considerably improve precision and recall. These sets comprise the Web genres (1) Academic and Information, (2) Blog, Community, and Shop, and (3) Corporate, Nonprofit, and Personal. We evaluate the effectiveness of the presented approach on a dataset consisting of 16,256 Web sites with 20,731,273 crawled and 100,321,069 known pages. These known pages include the crawled pages and further discovered but not actually downloaded pages. The dataset is derived from a restricted crawl of the German part of the Web, i.e. top-level domain .de, and is therefore concentrated on Web sites and Web pages with German content. Our approach achieves an accuracy of 92% for the classification of Web sites from this dataset.

The remainder of this chapter is organized as follows. Section 10.2 summarizes related work on the classification of Web sites. In Section 10.3, we explain the experimental framework and the underlying dataset. A detailed analysis of the identified features for classification utilized in the presented approach is given in Section 10.4.

Section 10.5 presents the results of classification using either only structure, only content, or the combination of both. Finally, we summarize our findings.

## 10.2 Related Work

Research on the classification of Web sites aims at discovering useful knowledge for establishing structure in the Web. It can be divided into two sub areas that differ in the targeted objective and granularity of classification. Coarse-grained classification, i.e. classification at super-genre level, seeks to discern the purpose a Web site is created for in order to improve the quality and ranking of search results, e.g. [1, 24]. Fine-grained classification deals with the automated categorization of Web sites in order to build Web directories, e.g. [14, 29].

Amitay et al. [1] showed that the structure of a Web site reflects the purpose it is created for. As a consequence, they utilized 16 structural properties as features for coarse-grained classification. Three of these properties were based on the distribution of pages into the distinct levels of a Web site and the remaining properties focused on its link structure. In particular, they were based on the external links to and from a Web site, e.g. number of outlinks per page, and on the internal linkage patterns within a Web site, e.g. number of crosslinks per page. Amitay et al. examined these structural properties for 202 Web sites and achieved a precision of up to 59% for their classification into classes that relate to the purpose of these Web sites. Opposed to [1], which mainly focused on the link structure of a Web site, we aimed at gaining a deeper insight into the relation between structure and purpose [23]. Therefore, we identified and analyzed different aspects of structural properties in a measurement-based study. The results of this study were used for the coarse-grained classification of Web sites by their structural properties into five classes (Academic, Blog, Corporate, Personal, and Shop). Lindemann and Littig [23] showed promising results with respect to an understanding of the relation between structure and purpose of Web sites and to their succeeding classification solely based on structural properties. As a consequence, we expand our studies by considering a more comprehensive set of structural properties on a larger dataset in order to reveal the potentiality and limitations of structure-based classification of Web sites. Building upon [23] and the condensed results of [24], we present an approach for classifying Web sites at super-genre level utilizing both structure and content. The classification of Web sites into eight Web genres introduced in this chapter achieves a considerably higher accuracy than previous approaches. Furthermore, the evaluation of our approach is based upon an 80-times larger dataset compared to [1].

In the field of fine-grained classification, Pierre [26] stressed several issues related to the content-dependent classification of Web sites. He presented a superpage-based approach for the classification of Web sites into industry categories, which considered especially metatags. In [21], Kwon and Lee utilized a k-nearest neighbor approach for text categorization and proposed a connectivity analysis for the classification of Web sites. Thus, they classified a Web site based upon the category and weight of its pre-classified Web pages. Kriegel et al. [14, 19]

presented various schemes for the fine-grained classification of Web sites which exploited semantic structure and local context information. In [14], they represented a Web site as a tree of Web pages with different topics. They employed a k-order Markov tree classifier and evaluated their approach based upon a testbed of 82,842 Web pages of 207 corporate Web sites. Building upon [14, 19] proposed the classification of Web sites as sets of feature vectors. Tian et al. [29] presented an approach that used hidden Markov trees for modeling the DOM tree of each Web page and the page tree of all Web pages of a Web site. They introduced a two-phase algorithm for Web site classification through fine-to-coarse recursion. Kumar et al. [20] considered the identification and segmentation of topically cohesive regions within the URL tree of large Web sites.

The outlined approaches for fine-grained classification [14, 19–21, 26, 29], focus on different small fractions of the Web or specialized problems. Furthermore, some of these approaches are only applicable to large Web sites or their evaluation was based upon small datasets. In contrast to this, the approach presented in this chapter allows for considering well-established Web genres at super-genre level and is effectively applied to a large dataset to ensure its practical applicability.

### 10.3 Dataset

Due to rapidly evolving Web genres that occur especially on fine-grained levels, a complete classification of Web sites constitutes an infeasible task. This problem is related to an outcome of internet-based communication and publication which often conflates general types of purpose resulting in the hybridising of previously discrete genres [5]. As a consequence, it makes sense to concentrate on well established Web genres at super-genre level. Thus, we consider the following set of Web genres:

- *G1 – Academic*: Universities and research institutions,  
e.g. <http://www.uni-leipzig.de>, <http://www.tu-freiberg.de>
- *G2 – Blog*: Web logs,  
e.g. <http://www.ard-sportblog.de>, <http://www.stoersignale.de>
- *G3 – Community*: Chats, forums, and online communities,  
e.g. <http://www.medien-foren.de>, <http://www.bastelforen.de>
- *G4 – Corporate*: Enterprises,  
e.g. <http://www.staudestahl.de>, <http://www.peschke-kainz.de>
- *G5 – Information*: Information portals, news, and media sites,  
e.g. <http://www.waz.de>, <http://www.zdf.de>
- *G6 – Nonprofit*: Foundations, schools, theatres, etc.,  
e.g. <http://www.preussenstiftung.de>, <http://www.landesmuseum-bs.de>
- *G7 – Personal*: Individuals and small groups,  
e.g. <http://www.harald-sandner.de>, <http://www.galbrecht.de>
- *G8 – Shop*: Online shops,  
e.g. <http://www.amazon.de>, <http://www.linoshop.de>

Similar genres are considered in other research studies [1, 16] which distinguish Web sites by the purpose they are created for. However, this set of Web genres cannot be complete due to the tremendous size of the World Wide Web as mentioned before. We intentionally omit spam Web sites in order to have a clean dataset because such sites occur in numerous shapes and dedicated methods for detecting them have been proposed, e.g. [15]. Furthermore, we do not focus on search engines and Web directories in accordance to our motivation as these Web sites are normally used for Web search and not searched for themselves. Nevertheless, we believe that the Web genres in our flat list are very relevant as the majority of Web sites can be clearly assigned to one of these Web genres. This still holds true in the presence of especially large Web sites which might have been created for multiple purposes or include subdomains that differ from the main purpose. However, the analysis of our dataset shows that there is almost always a dominant purpose which allows for a conclusive classification. As a consequence, our approach for classification is concentrated on single-label classification.

The process of mining Web content and structure is only focused on the top level domain .de (Germany) due to several reasons: Firstly, we want to avoid noisy data because of possible influences by national distinctions on Web design that might lead to differences in the structure of a Web site. Secondly, this part of the Web is a very good example of the Web of an industrialized country where Web sites of different Web genres reside in the same top level domain. Thirdly, the specific dictionaries for the different Web genres of Web sites, which are introduced in Section 10.4.2 are obviously language-specific. Nevertheless, we believe that our methodology is applicable to the top level domains of other industrialized countries since only the collection of Web sites has to be changed but not the Web mining methodology or the approach for Web site classification into Web genres at super-genre level.

For each of the considered eight Web genres we select Web sites from publicly available Web directories like the Open Directory Project [10] or dedicated commercial directories. This is achieved by randomly choosing URLs from appropriate categories with respect to the definition of the considered Web genres, i.e. selecting Web sites for Web genre Information from a listing of newspapers and enterprise Web sites from a specialized directory that solely includes such sites. We only rely on directories with manually edited entries to ensure the reliability of our dataset. In order to further increase this reliability, we make an effort in verifying the correctness of the disjoint sets of Web sites of each Web genre with the help of annotators who inspect a large number of random samples. Only if all annotators agree upon the classification of a Web site as given by the directory this Web site is taken into account. The annotators can fulfill their task more efficiently due to their knowledge of the German Web. Thus, this constitutes another reason for focusing on the top level domain .de.

A Web site, which constitutes the basic unit of analysis, is considered as the set of Web pages which belong to the same domain, e.g. uni-leipzig.de. Thus, the Web pages located in a subdomain, e.g. informatik.uni-leipzig.de, also belong to this Web site. This definition is made in accordance to related work in the field of Web site

classification, e.g. [14]. All Web sites were crawled in August 2006 by employing a search engine software developed by our group. Our search engine is capable of crawling and indexing more than 50,000 pages per hour on a Linux dual-processor PC server with 3.0 GHz Intel Pentium IV Xeon processors and 6 GB RAM. Crawling of a Web site always starts with the entry or homepage respectively and is performed in breadth-first-search manner following the internal links. The whole page tree is traversed in this way which allows for exactly determining the level of a Web page within the tree. This method is much more accurate than estimating the level of a page from the number of slashes within the URL path. Since our crawl is strictly focused on the pre-selected Web sites, external links are analyzed during the crawl but not followed. The content of a Web page is stored in a repository along with some information about the Web page. This allows for a later inspection of the content to derive features for classification. In order to reduce the traffic placed on the servers hosting the selected Web sites, we crawled at most 10,000 pages per Web site. This boundary introduces no bias since most Web sites comprise less than 10,000 pages and we are able to analyze structural properties of a Web site from Web pages which have not been downloaded. These facts are described in the next section. In addition to this restriction, our crawler follows the robots exclusion protocol and obeys the netiquette, i.e. our crawler requests at most one page per second from a Web server. We use several mechanisms in order to cope with usual crawler problems like crawler traps or large files. These mechanisms include an upper boundary for the downloadable file size and a restriction of the maximum level of a Web page to escape infinite loops. Collecting the data is completed when no further Web pages can be retrieved obeying these restrictions.

In our studies, we only examine a Web site if at least 30 pages have been correctly crawled. This approach minimizes measurement errors due to flash intros and redirections. Furthermore, 30 pages define the smallest size of a statistically relevant sample [31]. The resulting dataset includes 16,256 Web sites of the eight considered Web genres with a total amount of 20,731,273 crawled and 100,321,069 known pages as shown in Table 10.1. The known pages include the crawled pages and further discovered but not actually downloaded pages, i.e. in addition to the 20 million crawled pages we further discovered and analyzed about 80 million pages of the considered Web sites.

**Table 10.1** Statistics of dataset of restricted crawl

Web genre	No. of web sites	No. of crawled pages	No. of known pages
G1 – Academic	139	938,883	4,975,199
G2 – Blog	904	2,244,303	8,238,526
G3 – Community	515	2,019,507	9,813,806
G4 – Corporate	6,943	2,615,352	7,931,760
G5 – Information	243	1,691,492	9,658,511
G6 – Nonprofit	2,482	1,535,818	5,737,498
G7 – Personal	960	742,008	1,570,985
G8 – Shop	4,070	8,943,910	52,394,784
Total	16,256	20,731,273	100,321,069

## 10.4 Features for Classification

### 10.4.1 Features Derived from Structure

As the classification of Web sites at super-genre level is related to the purpose they are created for, it is crucial to identify the structural properties of a Web site that best reflect this purpose. Therefore, we deal with the identification of properties that describe the structure of a Web site with respect to several aspects. In particular, we focus on size, organization, technical realization, link structure, and URL composition.

Deriving features for the classification of Web sites from a Web crawl is a process of knowledge discovery from data. Due to the heterogeneity of the Web and its lack of structure, the initial step in this process is data preprocessing. It includes data cleaning and data reduction. As data reduction serves as a direct preparation for the classification, it is outlined in Section 10.5. Here, we first focus on descriptive data summarization, which provides the analytical foundation for data preprocessing [17]. It defines the transformation and summarization, respectively, of our measured data from the Web to structural properties of Web sites. Thus, descriptive summarization provides the basics for measuring the central tendency and the dispersion of our data. Based upon these measurements, the identification of structural properties includes two steps. Firstly, we define properties which are supposed to be of value. Secondly, we analyze these properties in order to investigate whether they really reflect the purpose of a Web site so that they can be used as features for classification.

Tables 10.2, 10.3, 10.4, 10.5 and 10.6 list the so identified structural properties of each of the five considered aspects. First of all, we analyze the size of a Web site in terms of page count, i.e. number of known pages, and amount of available data, i.e. average document size. The latter property can obviously only be derived from crawled Web pages. However, we identify many structural properties that can be inferred from all known Web pages of a Web site as depicted in the second column of Tables 10.2, 10.3, 10.4, 10.5 and 10.6. Since it is much more expensive to download a remote Web page than to perform in-memory classification operations [14, 29], the ability to derive knowledge of the structure of a Web site from non-crawled pages is a major advantage and heavily increases our underlying dataset.

The second considered aspect of a Web site is its organization. We concentrate on the fraction and number of different file types, the level of pages within the page tree, and the number of subdomains. In particular, we determine the file type the creator of a Web sites uses for the single Web pages by inspecting the file extension within the URL. Considering the file types in this way is different from extracting

**Table 10.2** Features considering the size of web sites

Feature	Derivation	Information gain
Number of known pages	Known	0.36
Average document size	Crawled	0.25

**Table 10.3** Features considering the organization of web sites

Feature	Derivation	Information gain
Fraction of PDF/PS files	Known	0.34
Fraction of HTML files	Known	0.29
Fraction of pages on densest level	Known	0.21
Number of different file types	Known	0.20
Average level	Known	0.20
Maximum level	Known	0.20
Fraction of scripts	Known	0.19
Number of subdomains	Known	0.11

**Table 10.4** Features considering the technical realization of web sites

Feature	Derivation	Information gain
Fraction of URLs with session ID	Known	0.31
Fraction of pages with javascript	Crawled	0.31
Number of different servers	Crawled	0.10

**Table 10.5** Features considering the link structure of web sites

Feature	Derivation	Information gain
Average external site outdegree	Crawled	0.54
Average external outdegree	Crawled	0.53
Average outdegree	Crawled	0.46
Average external leaf outdegree	Crawled	0.44
Average internal outdegree	Crawled	0.37
Number of external links on homepage	Crawled	0.34
Number of internal links on homepage	Crawled	0.29
Fraction of links to homepage	Crawled	0.28
External/internal linkage ratio	Crawled	0.27
Fraction of external links on densest level	Crawled	0.26
Downlink/uplink ratio	Crawled	0.19
Average external linkage level	Crawled	0.14
Fraction of side links	Crawled	0.13

**Table 10.6** Features considering the URL composition of web sites

Feature	Derivation	Information gain
Average number of digits in URL path	Known	0.29
Average number of slashes in URL path	Known	0.22
Domain length	Known	0.21
Average path length	Known	0.19

the content-type from the HTTP header since our approach should only focus on a Web site creator's choice for a file type and not on the actual content type. The approach allows for determining the fraction of HTML files, i.e. Web documents with the file extension .html or .htm, the fraction of PDF and PS files, which often constitute reference material, and the fraction of scripts, e.g. .asp, .php, .jsp, and .pl, within a Web site. Furthermore, we count the number of different file types, which are used within one Web site, and the number of subdomains. In the latter case,

we add up the number of different host parts within the URLs of a Web site. The homepage of a Web site is considered as the root of the page tree and has level 0. All pages that are linked from the homepage have level 1 and so forth. We determine the average and the maximum level of all pages. In addition to this, we calculate how many of all pages are situated on the densest level.

In order to analyze the technical realization of a Web site we explore whether session IDs and javascript are used. Session IDs are detected within the URLs and the use of javascript is detected by checking the source code of a page for certain tags. However, we want to note that session IDs can be stored in multiple ways, e.g. with cookies, so that they cannot always be detected from a URL. In addition to this, we determine the number of different servers the content is hosted on by inspecting the HTTP header.

A good source with plenty of information about the structure of a Web site is provided by its link structure. We exploit this source by considering the average internal, i.e. the average number of links to pages within the same Web site, and external outdegree. Furthermore, we distinguish between the average external, external site, and external leaf outdegree. The average external site outdegree is defined by the average number of links to different external Web sites instead of Web pages. The average external leaf outdegree describes the average number of links emanating from leaf pages within the page tree. Duplicated links within a Web page are counted only once in all cases. Furthermore, we count the number of external and internal links on the homepage, determine the fraction of links that point to the homepage, and calculate the ratio of external to internal links. In addition, we also compute the fraction of links on the densest level within the page tree and the average level from which external links are emanating. Finally, we determine the ratio of downlinks to uplinks and the fraction of sidelinks. All of these link types are internal links. Sidelinks point to pages on the same level within the page tree, while downlinks and uplinks point to pages on a lower and higher level, respectively.

Further properties describing the general composition of the URLs of a Web site can be directly derived from the URLs. We determine the average number of slashes and digits within the URL path, whereas successive slashes are counted only once, and the average length of this path. Furthermore, we ascertain the domain length, i.e. the length of the top domain without subdomains. All properties regarding the URL composition and the organization can be derived from the known pages.

The Tables 10.2, 10.3, 10.4, 10.5 and 10.6 list only those identified structural properties that possess discriminative power and reflect the purpose of a Web site. Therefore, they are declared as features for classification. In order to analyze their usability for distinguishing between Web sites of different Web genres, we compute the information gain of each single feature. The information gain is a feature selection measure which is usually used for selecting a splitting criterion that best separates a given dataset [25]. The higher the information gain the less amount of information is still required to finish classification.

We observe from Tables 10.2, 10.3, 10.4, 10.5 and 10.6 that all aspects of structural properties provide features with a high information gain, i.e. these features possess a high discriminative power so that they are very suitable for classification.

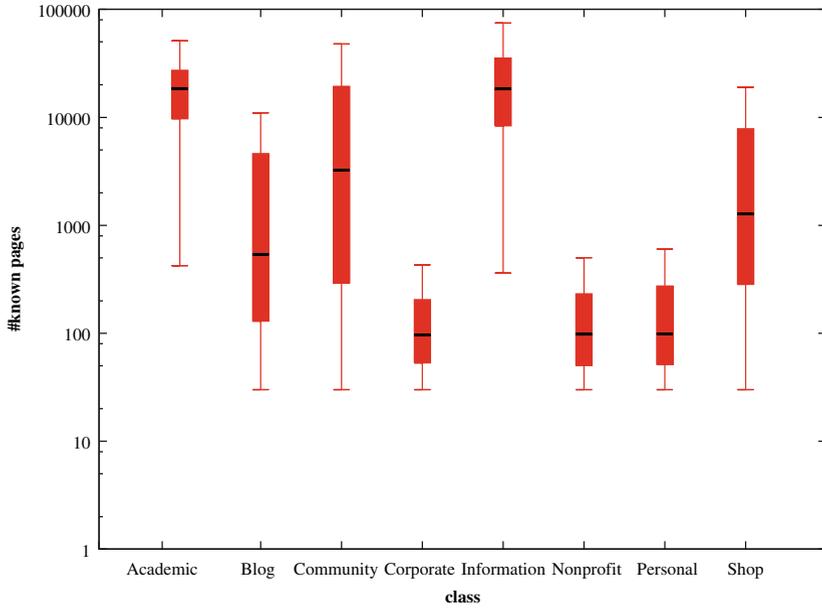
Therefore, the consideration of all aspects describing the structure of a Web site is equally important, although some features regarding the link structure yield an especially high value. Consequently, other studies extend the investigation of the link structure by considering a graph-based perspective for automatically analyzing Web genre data [9]. This is achieved by mining further graph patterns representing web-based hypertext structures.

We further delve into the relation between structure and purpose by analyzing the differences between Web sites of different Web genres with respect to certain structural properties. For illustration purposes, we concentrate on the most impressive examples of each of the five considered aspects. We visualize these differences by creating box-and-whisker plots, which are often used in the comparison of several sets of compatible data [17]. These plots incorporate the five-number summary of a distribution that includes the median, the first and third quartile, and the smallest and largest observations that are less than 1.5 times of the interquartile range (IQR) beyond the quartiles. This boundary of  $1.5 \times \text{IQR}$  is also used for data cleaning during data preprocessing. We filter out Web sites if they are detected as outliers with respect to the boundary for at least 10 of the 30 identified features. Furthermore, we do not consider Web sites with less than 30 crawled pages as mentioned before. Nevertheless, the statistics of our dataset as depicted in Table 10.1 present the number of Web sites after this data cleaning process.

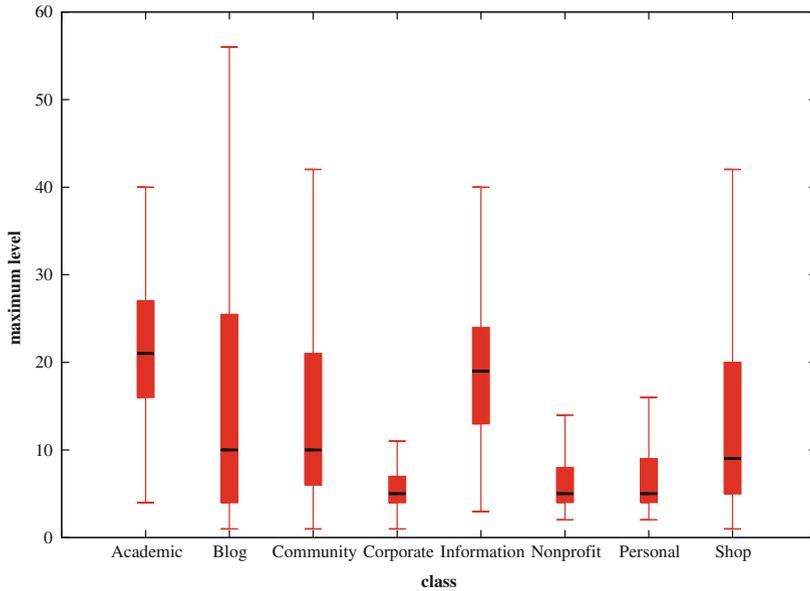
Figure 10.1 compares the Web genres with respect to the number of known pages. This feature focuses on the size of a Web site. We observe that the largest Web sites belong to the Web genres Academic and Information as 50% of these Web sites have more than 10,000 pages. This is indicated by the median, drawn as a horizontal black line within the boxes, which is the 50th percentile. In contrast to this, the smallest Web sites are members of the Web genres Corporate, Nonprofit, and Personal. In summary, we perceive a relation between the Web genres (1) Academic and Information, (2) Blog, Community, and Shop, and between the Web genres (3) Corporate, Nonprofit, and Personal. This trend can also be observed from Fig. 10.2 which shows the maximum level of the page tree as a representative of the features reflecting the organization of a Web site. Here, the Web genres Academic and Information stand out again since they possess the deepest page tree.

The average number of digits, which is the feature of type URL composition with the highest information gain, is shown in Fig. 10.3. We figure out further differences between the Web sites of different Web genres, e.g. Web sites of the Web genres Information and Shop comprise more digits within the URLs of their pages compared to the other Web genres. Although these simple structural properties focusing on the URL composition were not promising for reflecting the purpose of a Web site in first place, their information gain and the observation from Fig. 10.3 underline their usability as features for classification.

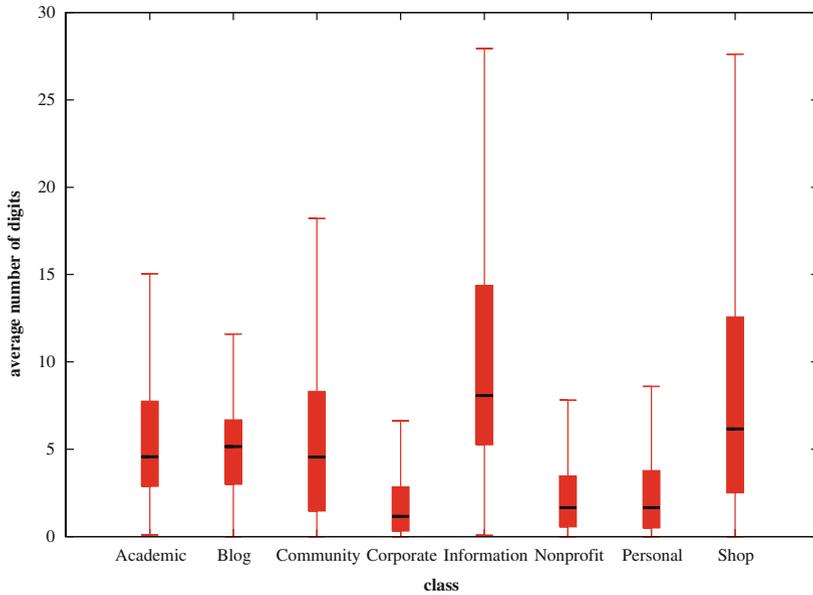
Figure 10.4 further proves the relation between structure and purpose. We observe from this figure that especially the Web genres Blog, Community, and Shop differ from the other Web genres as more than 50% of the Web sites of these Web genres use javascript on every Web page. Finally, the Figs. 10.5 and 10.6 relate to features describing the link structure of Web sites. Figure 10.5 denotes that Web sites



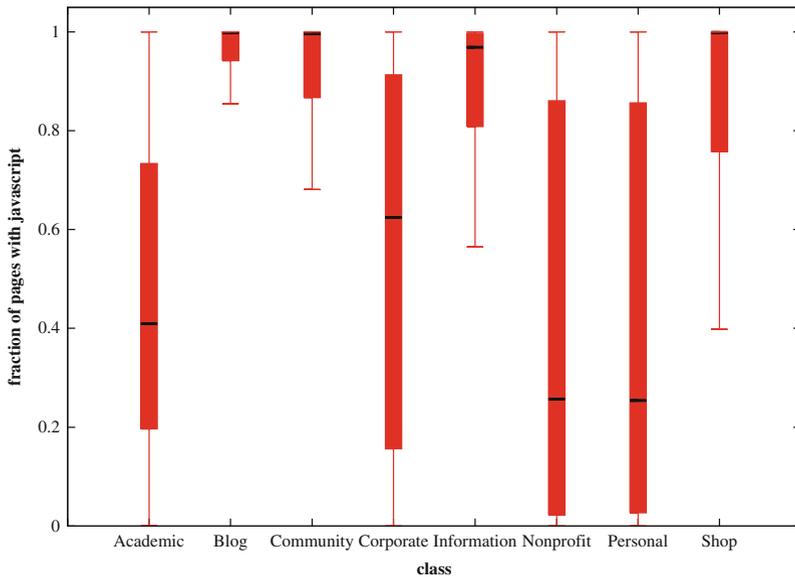
**Fig. 10.1** Number of known pages. Web sites of Web genres Academic and Information comprise the largest amount of pages. A relation between the Web genres Academic and Information – Blog, Community, and Shop – Corporate, Nonprofit, and Personal can be clearly perceived



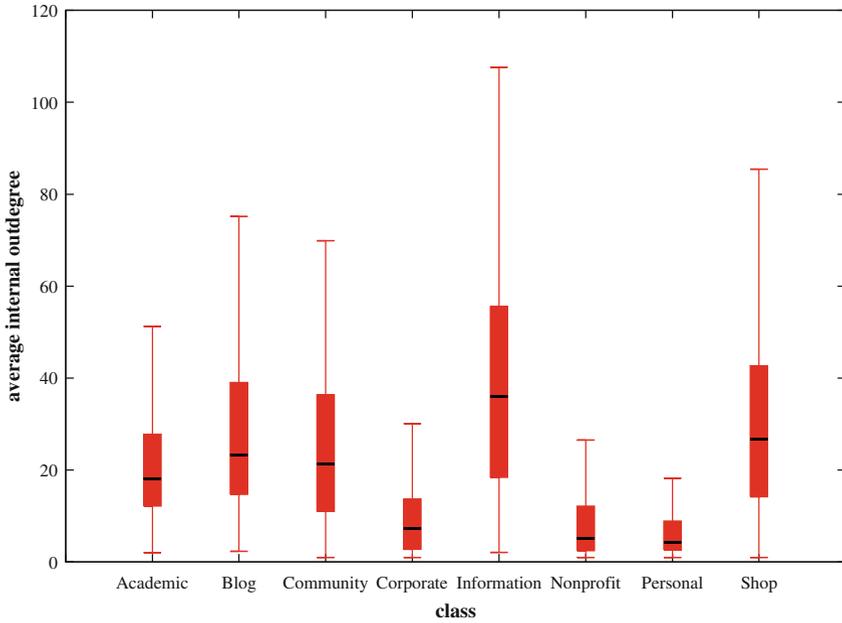
**Fig. 10.2** Maximum level of page tree. The Web genres Academic and Information stand out again since they possess the deepest page tree



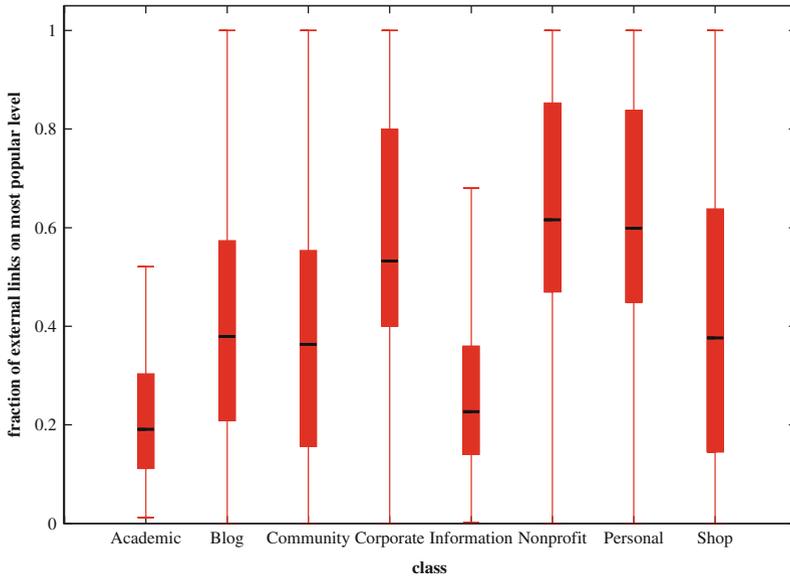
**Fig. 10.3** Average number of digits in URL path. Simple structural properties reflecting the URL composition are nevertheless useful as features for classification because differences between Web sites of different Web genres can be observed



**Fig. 10.4** Fraction of pages with javascript. The Web genres Blog, Community, and Shop differ from the other Web genres as more than 50% of the corresponding Web sites use javascript on every page



**Fig. 10.5** Average internal outdegree. Web sites of the Web genres Information and Shop have the strongest internal link structure



**Fig. 10.6** Fraction of external links on densest level of page tree. Relation between three sets of Web genres is again clearly revealed

of the Web genres Shop and Information have the strongest internal link structure. Here, more than 50% of these Web sites have on average more than 25 and 35 internal links per page. The fraction of external links on the densest level, depicted in Fig. 10.6, reveals a relation between three sets of Web genres again. Fifty percent of the Web sites of the Web genres Corporate, Nonprofit, and Personal include about 60% of all external links within pages on the densest level. Therefore, they clearly differ from the Web genres Blog, Community, and Shop which in turn differ from the Web genres Academic and Information.

All things considered, the observations from the plots are twofold. On the one hand, we reveal substantial differences between the structure of Web sites of different Web genres with respect to all aspects of identified properties. This allows for their utilization as features for classification by structure. On the other hand, we observe a relation between three sets of Web genres whose Web sites obviously exhibit a similar structure.

### ***10.4.2 Features Derived from Content***

The before mentioned similarities in the structure of Web sites of some Web genres fuel the necessity to explore further features for classification, which are not related to structure. Therefore, we aim at deriving features from the content of Web sites by creating a specific thesaurus for each Web genre. A specific thesaurus contains the most representative terms of a corpus that is specific for a Web genre. These terms stand out as prominent features for the corresponding Web genre.

We build such a thesaurus by utilizing a knowledge-weak technique for automatically learning relevant terminology [30]. This technique is especially approved in the context of text mining so that we customize it for the application to Web data. It is knowledge-weak because it simply relies on statistical analysis of term occurrences to point out relevant terms. Therefore, the only knowledge required to derive features from the content of a Web site is the assignment of the Web sites within the training data to a particular class and the ability to identify stopwords and HTML-tags as explained below. This is a great advantage as knowledge is in general costly in terms of time and effort.

The approach is based on a standard feature selection technique exploiting term statistics. In order to learn the terminology of a specific corpus, the distribution of terms within this target corpus is compared to the distribution in a general corpus. This general corpus is polythetic and is created by summarizing the terms of Web sites of all analyzed Web genres. It is used as a background filter. Prior to compiling term statistics, we strip off HTML-tags and ignore stop words like “and”. Stop words can be removed because these words are in general the most frequent words which consequently occur in Web sites of every considered Web genre. Therefore, stop words will not stand out as prominent features for a Web genre. A term is considered as relevant to a Web genre if it occurs significantly more often in the target corpus than in the background corpus. We determine the frequency of a term by counting it only once per Web site opposed to the original technique used for text mining. This method has turned out to be the most effective one since it is less

susceptible to words that are frequently mentioned only on single Web sites and it is not biased to the number of pages per Web site. Furthermore, it reduces the influence of advertising and especially template material within Web pages.

After these preprocessing steps, we use the log likelihood ratio statistics for a two-by-two contingency table to measure the correlation of terms with a target corpus. The log likelihood ratio is reported to be more effective in differentiating the relevance of a term for a specific category than other statistics like mutual information or information gain [17]. The computed log likelihood ratio scores can be mapped to a Chi-square distribution [13] in order to measure their significance. The most significant terms are finally inserted into the specific thesaurus of a Web genre. A two-by-two contingency table has one degree of freedom. Therefore, we insert a term into a specific thesaurus only if its Chi-square score is above 10.827 with a 0.1% chance of a Type I error. Furthermore, we have to assure that the relative frequency of a term is higher in the target than in the background corpus in order to prevent terms from the background corpus to get into the thesaurus. The resulting thesaurus is sorted descending by the score of the included terms. This approach selects on average about 1,000 specific terms per Web genre.

The top 5 terms with respect to their relevance for the corresponding Web genre are depicted in Table 10.7 along with their English translation to receive an impression of the specific thesauri. As can be seen from the table, the approach extracts descriptive terms for every considered Web genre.

**Table 10.7** Top 5 relevant terms per web genre

Web genre	Term/translation
G1 – Academic	Studium/study, Forschung/research, Student/student, Universität/university, Alumni/alumni
G2 – Blog	Blog/blog, Pingback/pingback, RSS/rss, Kommentar/comment, Weblog/weblog
G3 – Community	Chat/chat, Forum/forum, Community/community, Flirt/flirt, Mitglied/member
G4 – Corporate	Unternehmen/company, Lösungen/solutions, Referenzen/references, Fertigung/manufacturing, Industrie/industry
G5 – Information	Wirtschaft/economy, Tourismus/tourism, Wetter/weather, Kultur/culture, Stadt/city
G6 – Nonprofit	Theater/theatre, Gymnasium/secondary school, Museum/museum, Stiftung/foundation, Schule/school
G7 – Personal	Homepage/homepage, Gästebuch/guestbook, Photo/photo, Bild/picture, Familie/family
G8 – Shop	Shop/shop, Warenkorb/shopping cart, Versand/shipping, Onlineshop/online shop, Konto/account

## 10.5 Classification of Web Sites

Since we have derived features for the classification of Web sites at super-genre level from their structure and content, we seek to compare the classification accuracy of both approaches. In addition to this, we analyze strengths and weaknesses of these

approaches. We evaluate the accuracy of classification in terms of precision, recall, and F1 score [6]. Micro-averaging and macro-averaging aggregate these measures into an overall measure. Furthermore, we employ 10-fold cross validation [6, 25]. This method involves dividing our dataset randomly into 10 partitions of equal size. In each of 10 steps a classifier is learned on 9 of these partitions and tested on the remaining partition so that each Web site of the dataset is classified exactly once. A Web site is assigned to the Web genre with the highest probability. Subsequently, we determine in each step the achieved precision and recall for the considered Web genre individually and average them over all steps of the cross validation.

### *10.5.1 Classification by Structure*

We employ a naive Bayesian classifier for the classification by structure. This classifier is known to be simple but nevertheless very efficient and accurate in many domains [11, 12]. Since we do not aim at evaluating the best suited classifier for the classification task at hand, the naive Bayesian classifier defines a good benchmark in this scenario.

As mentioned before, data preprocessing is an essential step in Web Mining especially due to the heterogeneity of the Web. In preparation for the classification by structure we deal with data reduction which involves unsupervised data discretization, data transformation, and feature subset selection. These techniques are applied to obtain a reduced representation of the dataset that closely maintains the integrity of the original data but allows for much more efficiency in classification [17]. First of all, we divide the range of our continuous-valued features into intervals or bins by employing weighted proportional k-interval discretization [33]. Subsequently, we further transform our data by replacing each bin value by the bin mean, i.e. we carry out smoothing by bin means. After these steps of numerosity reduction we conduct feature subset selection in order to remove irrelevant or redundant features. This is achieved by utilizing the wrapper approach by Kohavi and John [18]. In this approach, a feature subset selection algorithm exists as a wrapper around an induction algorithm, which is considered as a black box and which is used to induce a classifier. The feature subset selection algorithm conducts a search for a good subset using the induction algorithm itself as part of the function evaluating feature subsets. The induction algorithm is run on the training data with different sets of features removed from the data. Finally, the feature subset with the highest evaluation is chosen for the classification of an independent test set that was not used during the search. We apply each step of data reduction in every turn of the cross validation to the nine partitions currently used as training data in order to evaluate the accuracy of classification by structure as accurately as possible.

The results of classification by structure are given in Table 10.8. We observe that this approach for classification yields a micro-averaged F1 score of 70%. A F1 score of 79% for the Web genres Academic, Corporate, and Shop indicates that Web sites of these Web genres can be best identified while it is more difficult to correctly classify Web sites of the Web genres Community and Personal.

**Table 10.8** Results of classification by structure

Web genre	Precision	Recall	F1 score
G1 – Academic	0.83	0.75	0.79
G2 – Blog	0.65	0.73	0.69
G3 – Community	0.26	0.29	0.27
G4 – Corporate	0.86	0.73	0.79
G5 – Information	0.72	0.75	0.73
G6 – Nonprofit	0.58	0.59	0.59
G7 – Personal	0.28	0.54	0.37
G8 – Shop	0.79	0.78	0.79
Micro-averaged	0.70	0.70	0.70
Macro-averaged	0.62	0.65	0.63

These results underline that structural properties are appropriate as features for classification. However, the achievable classification accuracy is limited due to similarities in the structure of some Web sites of different Web genres as outlined in Section 10.4.1. We seek to shed further light on these weaknesses of the approach by analyzing the confusion matrix depicted in Table 10.9. The rows and columns of the matrix are labeled with the shortcuts for the genres as defined in Section 10.3. The rows of the table show for each considered Web genre the actual number of Web sites that have been assigned to the different targeted Web genres given in the columns. As a consequence, the diagonal highlights the number of correctly classified Web sites. Tables 10.13 and 10.16 are depicted in the same way. We observe that Web sites of the Web genres Academic (G1) and Information (G5) are confused, e.g. 31 out of 139 academic Web sites are falsely assigned to Web genre Information. Similar observations can be made for the Web genres Blog, Community, and Shop as well as for the Web genres Corporate, Nonprofit, and Personal.

These observations match the results outlined in Section 10.4.1. As a consequence, we aggregate the corresponding Web genres into three sets. The classification of Web sites into these sets of Web genres results in a high accuracy which is highlighted by a F1 score of 94% and a low amount of misclassified sites as shown in Tables 10.10 and 10.11. While this approach is not valuable in accordance to our motivation of improving search results, it is an essential step towards the intended approach that combines classification by structure and content.

**Table 10.9** Confusion matrix of classification by structure

	G1	G2	G3	G4	G5	G6	G7	G8
G1	104	1	1	0	31	1	0	1
G2	0	664	33	0	0	0	12	195
G3	0	37	149	10	0	9	51	259
G4	0	61	139	5,043	26	789	709	176
G5	21	6	14	0	183	1	0	18
G6	0	20	55	540	8	1,470	258	131
G7	0	16	33	181	4	155	523	48
G8	0	214	158	116	3	98	313	3,168

**Table 10.10** Results of aggregated classification by structure

Aggregated web genre	Precision	Recall	F1 score
A (Academic, Information)	0.87	0.92	0.89
B (Blog, Community, Shop)	0.91	0.92	0.92
C (Corporate, Nonprofit, Pers.)	0.96	0.95	0.95
Micro-averaged	0.94	0.94	0.94
Macro-averaged	0.91	0.93	0.92

**Table 10.11** Confusion matrix of aggregated classification by structure

Aggregated web genre	A	B	C
A (Academic, Information)	352	28	2
B (Blog, Community, Shop)	10	5,051	428
C (Corporate, Nonprofit, Pers.)	44	466	9,875

### 10.5.2 Classification by Content

We apply an approach for matching unknown text against our specific thesauri in order to classify Web sites by their content. The first step in this process is to create a specific thesaurus for each considered Web genre based upon the nine sets assigned as training data in each turn of the cross validation. We classify all Web sites of the test data, i.e. of the remaining set, by computing a score for every Web genre. The central parameter of this score is the size of the intersection between the terms of a Web site and the terms of the specific thesaurus of a Web genre. The more terms the Web site has in common with the thesaurus of a certain Web genre, the more likely it is that the Web site belongs to that Web genre. However, it is important to take several other parameters into account in combination with this basic intuition. These parameters comprise the number of terms  $n$  found within a Web site, the number of known terms  $k$ , i.e. how many of the  $n$  terms appear in at least one thesaurus, and consequently the number of unknown terms  $u$ , where  $u = n - k$ . The size of the intersection is weighted by the inverse of  $n$  and by the ratio of  $k$  and  $u$ . In addition to this, it is weighted by the size of the currently considered thesaurus. Furthermore, the size of the intersection is not measured by simply counting the terms that appear in both sets but by summing up the weight of each term. This weight is defined by the inverse of the product of the number of thesauri it belongs to and its rank within the considered thesaurus. As a consequence, a Web site achieves a high score for a Web genre if it has a moderate number of terms most of which are very relevant for at best only one Web genre whose thesaurus is not too large itself. Finally, a Web site is assigned to the Web genre with the highest score.

Table 10.12 highlights the results of classification by content. Since the created thesauri are very descriptive for the Web genres, we achieve a micro-averaged F1 score of 84%. We observe that the classifier faces difficulties in correctly classifying Web sites of the Web genres Blog, Community, Information, and Personal as the F1 scores for all of these genres are less or equal 72%. This observation is further

**Table 10.12** Results of classification by content

Web genre	Precision	Recall	F1 score
G1 – Academic	0.82	0.89	0.85
G2 – Blog	0.58	0.92	0.72
G3 – Community	0.76	0.63	0.69
G4 – Corporate	0.86	0.93	0.90
G5 – Information	0.83	0.45	0.58
G6 – Nonprofit	0.85	0.74	0.79
G7 – Personal	0.79	0.60	0.68
G8 – Shop	0.90	0.83	0.86
Micro-averaged	0.84	0.84	0.84
Macro-averaged	0.80	0.75	0.77

**Table 10.13** Confusion matrix of classification by content

	G1	G2	G3	G4	G5	G6	G7	G8
G1	124	1	0	10	0	2	1	1
G2	0	836	11	11	2	0	8	36
G3	1	66	323	30	1	6	16	72
G4	6	77	3	6,477	7	179	17	177
G5	3	55	1	33	109	18	3	21
G6	8	115	2	453	3	1,829	29	43
G7	6	124	7	115	5	88	576	39
G8	4	156	79	359	4	30	77	3,361

approved by the confusion matrix given in Table 10.13, e.g. 124 out of 960 Web sites of Web genre Personal (G7) are misclassified as members of Web genre Blog (G2). These problems are also encountered by Web sites of the Web genres Community (G3) and Information (G5). A closer look on these Web genres reveals that they usually cover a wide spectrum of topics while Web sites of the other Web genres are more focused on certain fields like research or shopping.

### 10.5.3 Classification by Structure and Content

Considering the strengths and weaknesses of the classification by structure and by content we conclude that both approaches are very appropriate for the classification of Web sites at super-genre level. However, classification by structure suffers from Web sites that possess a similar structure although being created for different purposes. In addition to this, classification by content is affected by Web sites which do not generally focus on a certain topic. Therefore, we aim at developing a novel method for classification that utilizes structure and content in combination. On the one hand we overcome the disadvantages of classification by structure because we concentrate on three sets of aggregated Web genres, i.e. we reduce the difficulties of classification to solving a three class problem. On the other hand we overcome the disadvantages of classification by content by selectively combining the results of classification by structure and content.

**Table 10.14** Algorithm for classification by structure and content

---

1:	measure structural properties of Web sites of training data
2:	perform data preprocessing
3:	determine best feature subset with wrapper approach
4:	create specific thesaurus for each genre based upon training data
5:	for each Web site of the test data do
6:	analyze terms of the Web site
7:	compute score for every Web genre
8:	determine maximum score
9:	if max. score < 0.75 or prediction = G2, G3, G5, or G7 then
10:	analyze features of determined subset
11:	use naive Bayesian classifier to determine aggregated set
12:	combine results of classification by structure and content
13:	assign Web site to a Web genre based on combined votes
14:	else
15:	assign Web site to a Web genre based on maximum score
16:	end if
17:	end for

---

The complete algorithm for the approach that utilizes structure and content in combination is outlined in Table 10.14. The preparation for classification by structure, i.e. lines 1–3, and by content, i.e. line 4, constitutes the first step of the algorithm. As mentioned before, this step is focused only on the training data while the classification itself is processed on independent test data that is not included in the training data, see line 5. The comparison of the results of classification by structure only and by content only, as described in Section 10.5.1 and 10.5.2 and outlined in Table 10.17, reveals a higher potentiality of classification by content. Therefore, we classify a Web site based on its content by computing a score which reflects the probability of the Web site to belong to one of the considered eight Web genres, i.e. lines 6–8. If this approach does not result in a certain decision for one Web genre, i.e. the highest probability is below 0.75, or if it predicts one of the before identified problematic Web genres regardless of the score, we continue with classification by structure, i.e. lines 10–11. Otherwise, the Web site is directly assigned to the predicted Web genre, i.e. line 15. The threshold for the score, that defines the minimum probability of a certain decision, is determined experimentally. Blog (G2), Community (G3), Information (G5), and Personal (G7) are the problematic Web genres as outlined in Section 10.5.2 since the content-based classification is complicated by the fact that they usually cover various topics. The structure-based classification assigns the currently processed Web site to one set of the aggregated Web genres. Afterwards, the results of this classification approach are combined with the before computed results of classification by structure, i.e. line 12. This combination is achieved by taking the average of the probability determined by content-based classification and the probability of the predicted set of aggregated Web genres for the Web genres within this set. The probability of the remaining Web genres is determined by reducing the before computed probability of content-based classification. The resulting values for each Web genre are normalized so that they sum up to 1. As a consequence, the combined classification usually results in a

higher probability for the Web genres within the aggregated set predicted before. Thus, the classification by content is most of the time refined to choose from these Web genres which clearly leads to a reduced error rate. Finally, the Web site is assigned to the Web genre with the highest probability based upon this approach, i.e. line 13.

The results of the classification by combining structure and content can be observed from Table 10.15. The presented approach yields a micro-averaged F1 score of 92%. Furthermore, our approach achieves a F1 score of at least 75% for every Web genre at super-genre level. The confusion matrix given in Table 10.16 further approves that the introduced approach is very appropriate because we observe that most elements are represented along the diagonal which shows the number of correctly classified Web sites for each genre. Table 10.17 compares the results of classification by structure, by content, and by the combination of both. We observe that the combined approach results in a significantly higher F1 score for all considered Web genres. The micro-averaged F1 score is increased from 70 to 84% for the classification by structure and by content respectively to 92% for the combined approach. Thus, its accuracy clearly surpasses the achievements of classification solely based on structure or content.

**Table 10.15** Results of classification by structure and content

Web genre	Precision	Recall	F1 score
G1 – Academic	0.84	0.99	0.91
G2 – Blog	0.83	0.92	0.87
G3 – Community	0.79	0.71	0.75
G4 – Corporate	0.94	0.98	0.96
G5 – Information	0.98	0.79	0.88
G6 – Nonprofit	0.99	0.82	0.89
G7 – Personal	0.77	0.87	0.82
G8 – Shop	0.93	0.92	0.93
Micro-averaged	0.92	0.92	0.92
Macro-averaged	0.88	0.88	0.88

**Table 10.16** Confusion matrix of classification by structure and content

	G1	G2	G3	G4	G5	G6	G7	G8
G1	137	0	0	2	0	0	0	0
G2	0	834	18	2	0	0	0	50
G3	2	26	367	4	0	0	40	76
G4	2	6	0	6,835	1	20	23	56
G5	10	8	7	8	193	0	0	17
G6	5	48	9	290	1	2,023	59	47
G7	4	38	10	33	1	8	834	32
G8	3	49	56	90	0	2	123	3,747

**Table 10.17** Comparison of classification results in terms of F1 score

Web genre	Structure-based	Content-based	Combined
G1 – Academic	0.79	0.85	0.91
G2 – Blog	0.69	0.72	0.87
G3 – Community	0.27	0.69	0.75
G4 – Corporate	0.79	0.90	0.96
G5 – Information	0.73	0.58	0.88
G6 – Nonprofit	0.59	0.79	0.89
G7 – Personal	0.37	0.68	0.82
G8 – Shop	0.79	0.86	0.93
Micro-averaged	0.70	0.84	0.92
Macro-averaged	0.63	0.77	0.88

## 10.6 Conclusion

We presented an approach for the classification of Web sites at super-genre level. This approach consisted of a combination of classification by structure and content. Classification by structure was used for a classification of Web sites into three sets of aggregated Web genres exploiting 30 structural properties. Classification by content utilized specific thesauri for the classification of Web sites into eight well-established Web genres. The proposed combination of classification by structure and content significantly improved the overall accuracy of classification. In particular, we found out that on the one hand classification by structure suffers from Web sites that possess a similar structure although being created for different purposes. On the other hand, the accuracy of classification by content is limited for Web sites that deal with various topics while they are still clearly related to a certain purpose, e.g. Web sites of Web genre Blog. As a consequence, the micro-averaged F1 score increased from 70% for classification by structure and 84% for classification solely based on content to 92% for the presented approach combining structural and content-specific features.

## References

1. Amitay, E., D. Carmel, A. Darlow, R. Lempel, and A. Soffer. 2003. The connectivity sonar: Detecting site functionality by structural patterns. In *Proceedings of the 14th Conference on Hypertext and Hypermedia*. Nottingham.
2. Biber, D. 1988. *Variation across speech and writing*. Cambridge, MA: Cambridge University Press.
3. Björneborn, L. 2010. Genre connectivity and genre drift in a web of genres. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini. Dordrecht: Springer.
4. Braslavski, P. 2010. Marrying relevance and genre rankings: An Exploratory Study. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini. Dordrecht: Springer.

5. Bruce, I. 2010. Evolving genres in online domains: The hybrid genre of the participatory news article. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, M. Dordrecht: Springer.
6. Chakrabarti, S. 2003. *Mining the web*. San Francisco, CA: Morgan Kaufmann.
7. Cho, J., and H. Garcia-Molina. 2000. The evolution of the web and its implications for an incremental crawler. In *26th Conference on Very Large Data Bases*. Cairo.
8. Cooley, R. 2003. The use of web structure and content to identify subjectively interesting web usage patterns. *ACM Transactions on Internet Technology* 3(2):93–116.
9. Dehmer, M., and F. Emmert-Streib. 2010. Mining graph patterns in web-based systems: A conceptual view. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini. Dordrecht: Springer.
10. DMOZ. Open directory project, <http://www.dmoz.org>
11. Domingos, P., and M. Pazzani. 1997. On the optimality of the bayesian classifier under zero-one loss. *Machine Learning* 29:103–137.
12. Duda, R., P. Hart, and D. Stork. 2001. *Pattern classification*, 2nd Ed. New York, NY: Wiley.
13. Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19:61–74.
14. Ester, M., H.-P. Kriegel, and M. Schubert. 2002. Web site mining: A new way to spot competitors, customers and suppliers in the World Wide Web. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*. Edmonton.
15. Fetterly, D., M. Manasse, and M. Najork. 2004. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases*. Paris.
16. Gibson, D., K. Punera, and A. Tomkins. 2005. The volume and evolution of web page templates. In *Proceedings of the 14th International World Wide Web Conference*. Chiba.
17. Han, J., and M. Kamber. 2006. *Data mining*, 2nd Ed. San Francisco, CA: Morgan Kaufmann.
18. Kohavi, R., and G. John. 1997. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324.
19. Kriegel, H.-P., and M. Schubert. 2004. Classification of websites as sets of feature vectors. In *International Conference on Databases and Applications*. Innsbruck.
20. Kumar, R., K. Punera, and A. Tomkins. 2006. Hierarchical topic segmentation of websites. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA.
21. Kwon, O.-W., and J.-H. Lee. 2003. Text categorization based on k-nearest neighbor approach for web site classification. *Information Processing and Management* 39:25–44.
22. Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC Jungle. *Language Learning & Technology* 5:37–72.
23. Lindemann, C., and L. Littig. 2006. Coarse-grained classification of web sites by their structural properties. In *Proceedings of the 8th International Workshop on Web Information and Data Management*. Arlington, VA.
24. Lindemann, C., and L. Littig. 2007. Classifying web sites. In *Proceedings of the 16th International World Wide Web Conference*. Banff.
25. Liu, B. 2007. *Web data mining: Exploring hyperlinks, contents and usage data*. Heidelberg: Springer.
26. Pierre, J.M. 2001. On the automated classification of web sites. *Linköping Electronic Articles in Computer and Information Science* 6.
27. Sharoff, S. 2010. In the garden and in the jungle: Comparing genres in the BNC and internet. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini. Dordrecht: Springer.
28. Stein, B., S. Meyer zu Eissen, and N. Lipka. 2010. Web genre analysis: Use cases, retrieval models, and implementation issues. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini. Dordrecht: Springer.
29. Tian, Y.-H., T. Huang, and W. Gao. 2004. Two-phase web site classification based on hidden Markov tree models. *Web Intelligence and Agent Systems* 2:249–264.

30. Vogel, D. 2003. Using generic corpora to learn domain-specific terminology. In *Workshop on Link Analysis for Detecting Complex Behavior*. Washington, DC.
31. Weiss, N.A. 2002. *Introductory Statistics*, 6th Ed., Greg Tobin. Reading MA: Addison Wesley.
32. Yahoo! Mindset, <http://mindset.research.yahoo.com>
33. Yang, Y., and Webb, G. 2003. Weighted proportional k-interval discretization for naive-bayes classifiers. *Artificial Intelligence* 2637:501–512.



# Chapter 11

## Mining Graph Patterns in Web-Based Systems: A Conceptual View

Matthias Dehmer and Frank Emmert-Streib

### 11.1 Introduction

The task of applying Data Mining methods [38] to web-based hypertexts is often referred to as Web Mining [16]. In view of the steadily increasing complexity of web data sources and the huge amount of information available online, Web Mining has been an important and fruitful research topic [16, 46]. Generally, Web Mining can be divided into the following categories:

1. Web Content Mining: Web Content Mining provides methods for automatically extracting information from web-based data sources. Important problems are data extraction and analysis by using, e.g., Text Mining methods [53].
2. Web Structure Mining: Web Structure Mining deals with exploring structural properties of web-based hypertexts, e.g., investigating internal and external link structures of web-based documents [16] or exploring hypertext structure types using graph-based models [55]. Moreover, there are a lot of earlier contributions rooted in complex network theory [29] dealing with analyzing mathematical growth-properties of the web graph and web subgraphs by using stochastic models [1, 34, 40, 48, 63]. Often, these methods aim to improve web-search and information extraction algorithms in Web Mining [14, 45].
3. Web Usage Mining: Web Usage Mining [73] deals with exploring and analyzing patterns reworked from web logs to analyze behavior of hypertext users. Such an analysis can be in particular useful to optimize business websites, to analyze their quality and to detect effectiveness features, see, for example [64].

In this chapter, we put the emphasis on discussing methods (in the context of Web Structure Mining) to analyze graph-based hypertext patterns. To tackle our problem, we discuss a graph-theoretic framework for exploring graph-based patterns representing web-based hypertext structures. Besides modeling document structures as

---

M. Dehmer (✉)

Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Vienna, Austria; Institute for Bioinformatics and Translational Research, Hall in Tyrol, Austria  
e-mail: matthias.dehmer@univie.ac.at; mdehmer@geometrie.tuwien.ac.at;  
Matthias.Dehmer@umit.at

graphs [57] that means that in the sense of consistent graph similarity measuring, we apply a method to measure the structural similarity of graphs (see Section (11.4.2)) to approach problems in Web Structure Mining, for example:

1. Computing the cumulative similarity distribution  $\Theta$  of a web genre [50] corpus containing graph-based documents (see Section 11.5). A possible interpretation of  $\Theta$  addresses the important question how structurally distributed the graph-based documents in the given corpus are.
2. Structural filtering of web-based units: By measuring the structural similarity of the document structures and then applying clustering techniques, we obtain clusters which contain structurally similar web-based units.

The main contribution of this conceptual chapter is to shed light on the task of automatically analyzing web genre data by using a method for structurally comparing graph-based hypertexts [18, 22, 27]. We use the term “web genre” and “web genre data” in the sense of Mehler et al. [59] where web genres are considered as hypertext types, see, e.g. [59, 65]. Also, we want to emphasize that we do not use the vector space model [31, 52] to represent a web-based document structure [18, 24]. Instead, we use a special graph class called generalized trees (GTs) [25, 57] for modeling our web-based documents [57].

Basically, Mehler focuses on webgenres not from the point of view of a bag-of-features model [56]. Rather, this approach conceives instances of webgenres as complex signs that have a characteristic structure due to their membership to a certain genre. This contrasts genre modeling with topic modeling in Information Retrieval [2] where a topic is represented by a set of lexical units that are typically used to manifest that topic. Rather, Mehler’s approach is linguistic in the sense that instances of a certain text type are seen to have a characteristic topical structure *and* a characteristic generic structure. Take the example of a newspaper article in contrast to, say, a personal letter: although in both cases the universe of topics is certainly open, we can nevertheless expect that instances of both types depart with respect to the topical areas they typically deal with. Moreover, the differences between these text types are also manifested in structural terms: the structure of a letter significantly differs from that of most newspaper articles. *So why not exploring text structure [28], document structure [62] or even layout structure [75] to get insights into the webgenre (or hypertext type) of a webpage or of a website?*

Interestingly, many webgenre models oversee this structural source of the characteristics of webgenres. Consequently, they tend to rely on some extension or simply on some application of the bag-of-features or vector space model. However, such an approach disregards a central characteristic of web units as instances of webgenres, that is, their hyperlink structure, which is genuine web-based. From this point of view, a website is seen to be identifiable as an instance of a webgenre by means of its hypertextual structure – beyond its textual structure. Mehler et al. [59] have shown that because of many aspects of informational uncertainty this hypertextual structure is – by analogy to its textual counterpart – not immediately accessible: neither can we simply read-out this structure from HTML tags or URLs, nor is it manifested by hyperlinks only. Rather, this *hidden* hypertext document structure needs first to be explored as this is done with its counterpart in the form of document structure [62].

In this paper, we propose a structural approach of webgenres and webgenre classification that builds upon a webgenre-related hypertext structure model. More specifically, we utilize a certain graph model (in the form of generalized trees) that has been found to be the structural kernel of many complex linguistic aggregates [54]. Our task is to add a computational model that deals with this class of graphs as a model of webgenre structure. In this sense, we propose an algorithmic model that integrates a recent structural model of linguistic units by example of webgenres with their computational processing.

The graph similarity-based approach we want to discuss in this chapter operates on generalized trees representing hierarchical and directed graphs. We notice that generalized trees are more general than ordinary rooted trees because a generalized tree contains an ordinary rooted tree as a special case. For practical applications, this implies that a generalized tree captures more structural information of the underlying document structure than an usual DOM-tree [15] represented by a directed rooted tree. The classical DOM-tree model has been also applied for measuring the structural similarity of underlying hypertext structures by [13, 42].

The chapter is organized as follows: Section 11.2 presents some mathematical preliminaries. In Section 11.3, we briefly discuss the problem of deriving structural properties of graphs to characterize them structurally. Besides outlining existing methods for measuring the similarity of web-based document structures in Section 11.4, this section also discusses a graph similarity-method that operates on generalized trees. In Section 11.5, we outline resulting applications in Web Structure Mining and Web Usage Mining. The chapter finishes with a short summary in Section 11.6.

## 11.2 Mathematical Preliminaries

First, we introduce some mathematical preliminaries [25, 37, 39].

**Definition 1**  $G = (V, E)$ ,  $|V| < \infty$ ,  $E \subseteq \binom{V}{2}$  is called a finite undirected graph.  $G = (V, E)$ ,  $|V| < \infty$ ,  $E \subseteq V \times V$  represents a finite directed graph.

**Definition 2** Let  $G = (V, E)$  be a graph.  $\tilde{G} = (\tilde{V}, \tilde{E})$  is called a subgraph iff  $\tilde{V} \subseteq V$  and  $\tilde{E} \subseteq E$ . Moreover, if it holds  $\tilde{E} = E \cap (\tilde{V} \times \tilde{V})$ , then we call  $\tilde{G}$  the induced subgraph of  $G$ .

**Definition 3** An isomorphism class denotes the set of graphs which are isomorphic to a given graph  $G$ .

**Definition 4** A tree is a connected, acyclic undirected graph. A tree  $T = (V, E)$  with a distinguished vertex  $r \in V$  is a rooted tree.  $r$  is called the root of the tree. The level of a vertex  $v$  in a rooted tree  $T$  equals the length of the path from  $r$  to  $v$ . The maximum path length  $d$  from the root  $r$  to any vertex in the tree is called the depth of  $T$ . A leaf is a vertex incident to exactly one edge in a tree.

**Definition 5** Let  $G = (V, E)$  be a finite, directed graph. Then, we define the following sets and quantities:

$$\begin{aligned} \mathcal{N}^+(v) &= \{w \in V \setminus \{v\} \mid (v, w) \in E\}, \\ \mathcal{N}^-(v) &= \{w \in V \setminus \{v\} \mid (w, v) \in E\}, \\ \delta_{\text{out}}(v) &= |\mathcal{N}^+(v)|, \\ \delta_{\text{in}}(v) &= |\mathcal{N}^-(v)|. \end{aligned}$$

We call  $\delta_{\text{out}}(v)$  and  $\delta_{\text{in}}(v)$  out-degree and in-degree of  $v \in V$ , respectively.

**Definition 6** A directed acyclic graph  $T$  is called a *directed rooted tree* if there is an unique vertex  $r$  satisfying  $\delta_{\text{in}}(r) = 0$  from which any other vertex of  $T$  is reachable by a unique path.

**Definition 7** Let  $T = (V, E_1)$  be a directed rooted tree. The vertex set is defined by

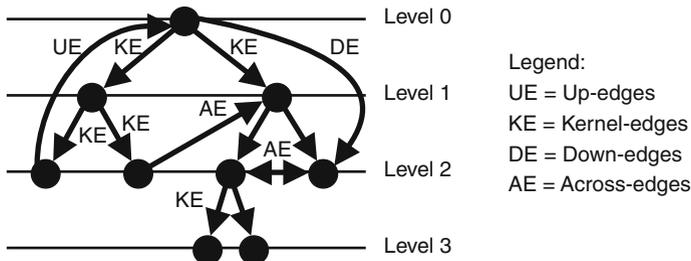
$$V := \{v_{0,1}, v_{1,1}, v_{1,2}, \dots, v_{1,|V_1|}, v_{2,1}, v_{2,2}, \dots, v_{2,|V_2|}, \dots, v_{d,1}, v_{d,2}, \dots, v_{d,|V_d|}\}, \tag{11.1}$$

and we assume  $|V| < \infty$ .  $|L|$  denotes the cardinality of the level set  $L = \{l_0, l_1, \dots, l_d\}$ . The surjective mapping  $\mathcal{L} : V \rightarrow L$  is called a multi level function that assigns to every vertex an element of the level set  $L$ . It holds  $d = |L| - 1$ .  $v_{i,j}$  denotes the  $j$ -th vertex on the  $i$ -th level,  $0 \leq i \leq d, 1 \leq j \leq |V_i|$ .  $|V_i|$  denotes the number of vertices on level  $i$ . The edge set  $E_{GT} := E_1 \cup E_2 \cup E_3 \cup E_4$  of a finite generalized tree  $H = (V, E_{GT})$  is defined as [57]:

- $E_1$  forms the edge set of the underlying directed rooted tree  $T$ . These edges are called Kernel-edges.
- $E_2$ : *Up-edges* associate analogously vertices of the tree hierarchy with one of their (dominating) predecessor vertices.
- $E_3$ : *Down-edges* associate vertices of the tree hierarchy with one of their (dominated) successor vertices in terms of that tree hierarchy.
- $E_4$ : *Across-edges* associate vertices of the tree hierarchy, none of which is an (immediate) predecessor of the other in terms of the tree hierarchy.

Figure 11.1 shows a generalized tree exemplarily.

**Definition 8** We define some metrical properties of graphs.  $d(u, v)$  denotes the distance between  $u \in V$  and  $v \in V$  representing the minimum length of a



**Fig. 11.1** A generalized tree with its edge types

path between  $u, v$ . Note that  $d(u, v)$  is an integer metric. We call the quantity  $\sigma(v) = \max_{u \in V} d(u, v)$  the eccentricity of  $v \in V$ .  $\rho(G) = \max_{v \in V} \sigma(v)$  and  $r(G) = \min_{v \in V} \sigma(v)$  is called the diameter and radius of  $G$ , respectively.

### 11.3 Structural Graph Measures

Graphs can be considered as powerful and generic models to describe complex relational objects which appear in a large number of scientific areas, e.g., computer science, chemistry, sociology, cognitive sciences and biology [17, 33, 76]. Apart from using graphs for modeling real world problems, an important problem is also to quantify structural information by inferring structural properties of a graph in question. This problem addresses the task of characterizing graphs based on graph measures. To give a short overview on such structural network measures, we present the listing as follows:

1. Degree distributions  $P(i)$ , e.g., see [29].
2. Exponent of degree distributions, i.e., it holds  $P(i) \sim i^{-\gamma}$ , e.g., see [29].
3. Total number of vertices  $|V|$  and edges  $|E|$ .
4. Distance matrix  $(d(v_i, v_j))_{v_i, v_j \in V}$ .
5. Metrical properties of graphs, e.g.,  $\sigma(v)$ ,  $\rho(G)$  and  $r(G)$ , e.g., see [70].
6. Clustering coefficient, modularity and network motifs, e.g., see [3, 8].
7. Vertex centrality measures, e.g., see [9, 51, 76].
8. Eigenvector measures, e.g., see [47, 51].

Another method to characterize graphs is based on quantifying structural information using information-theoretic measures. This problem relates to determine the structural complexity of a graph. Entropic measures to determine the so-called structural information content of a graph have been developed by [7, 6, 19, 20, 30]. A task that is also related to determine structural features of graphs is to identify stylistic properties. For example, a stylistic property can be understood as a characteristic structural feature of a graph that manifests a graph class, e.g., a hierarchy, an undirected edge set, a directed edge set etc. To identify such features exemplarily, we consider Fig. 11.2. The depicted graphs from different application domains manifest different styles of graphs. More precisely, graph (A) represents a directed rooted tree to model a DOM-structure. Graph (B) shows a more complex website structure representing a generalized tree. Graph (C) is a chemical structure represented by an undirected and vertex labeled graph. A different definition of a style that aims to compare such styles structurally (this lead to a generalization of the classical graph similarity problem [26]) has been already expressed in [26]. In [26], a style was defined as a set of graphs with impressed structural properties. Finally, we compared the styles by using a method which is based on the definition of a median graph [26, 58].

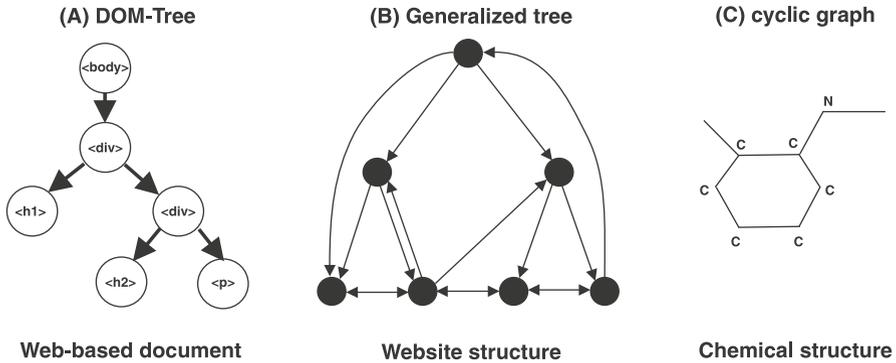


Fig. 11.2 Graph styles from different application domains

## 11.4 Graph Similarity Measures for Web Mining

### 11.4.1 Classical Similarity and Distance Measures for Graphs

The problem of measuring the similarity (or distance) between structures representing networks occur in numerous scientific disciplines [5, 13, 22, 68]. Usually, graph similarity measures are based on incorporating structural features of given graphs, e.g., degree sequences, subgraphs, and other metrical properties of graphs [70]. Also, the task of measuring the structural similarity of graphs is often referred to as *graph matching* [12]. There exist basically two major paradigms for matching graphs structurally which have been intensely discussed in the scientific literature: *exact graph matching* and *inexact graph matching* [12].

Exact graph matching is mainly based on the principle of finding a graph or a subgraph of a given graph that matches a graph or subgraph structure of another graph exactly. With other words, one has to determine if two graphs are isomorphic [39], i.e., structurally equivalent. It is known that even classical graph similarity measures belonging to the exact graph matching paradigm are based on determining isomorphic and subgraph isomorphic relations, see, e.g., [43, 71, 72, 77]. A prominent example of a classical graph metric represents the well-known Zelinka-distance [77]; two graphs are more similar, the bigger the common induced (isomorphic) subgraph is. This implies that graphs which have a large common induced subgraph have a small distance and vice versa. It is worth mentioning that Zelinka [77] was the first who introduced such a measure for unlabeled graphs of same order. The key result is as follows [71, 72, 77].

**Theorem 1** Let  $H = (V_H, E_H)$  and  $G = (V_G, E_G)$  be unlabeled graphs without reflexive and multiple edges and it holds  $|V_H| = |V_G| = n$ .  $\overline{SUB}_m(H)$  denotes the set of induced subgraphs of order  $m$ .  $H^*$  denotes the isomorphism classes of such graphs in which  $H$  lies and let

$$SUB_m(H) := \{H^* \mid H \in \overline{SUB}_m(H)\}. \quad (11.2)$$

$SUB_m(H)$  is just the set of isomorphism classes in which the induced subgraphs of  $H$  with order  $m$  lie. Then,

$$d_Z(H, G) := n - SIM(H, G), \quad (11.3)$$

is a graph metric, where

$$SIM(H, G) := \max\{m \mid SUB_m(H) \cap SUB_m(G) \neq \emptyset\}. \quad (11.4)$$

A more general version of this theorem was introduced by Sobik [71, 72]. The following assertion states that the measure  $d_S(H, G)$  for determining the structural similarity of arbitrary and also labeled graphs represents a graph metric.

**Theorem 2** Let  $H := (V, E, f_V, f_E, A_V, A_E)$  be a finite and labeled graph.  $A_V, A_E$  denote finite, non-empty vertex and edge alphabets and  $f_V : V \rightarrow A_V, f_E : E \rightarrow A_E$  the associated vertex and edge labeling functions. Now, let  $H$  and  $G$  be finite, labeled graphs of arbitrary orders, respectively. Then,

$$d_S(H, G) := \max\{|H|, |G|\} - SIM(H, G) \quad (11.5)$$

is a graph metric.

Now, we want to briefly discuss inexact graph matching. The most prominent measure from inexact graph matching is the so-called *graph edit distance* (GED) developed by Bunke [10]. It can be considered as a powerful extension of the Levenshtein-distance [49]. GED is mainly based on the idea to define graph edit operations such as insertion or deletion of an edge/vertex or relabeling of a vertex along with costs associated with each such operation [10]. Moreover, Bunke [10] calls an optimal inexact match a sequence of edit operations which transforms a graph  $G$  into  $H$  by producing minimal transformation costs. If  $m_1, m_2, \dots, m_n$  are assumed to be all possible transformations mapping  $G$  to  $H$ , then the optimal inexact match [10]  $m'$  is defined by

$$c(m') = \min\{c(m_i) \mid 1 \leq i \leq n\}. \quad (11.6)$$

Finally, the graph edit distance between two graphs is the minimum cost associated with a sequence of edit operations. Further, the optimal error-correcting graph isomorphism is defined as the resulting isomorphism after obtaining this optimal sequence of edit operations [10]. The original result of Bunke [10] can be now expressed as follows.

**Theorem 3** Let  $d(H, G)$  be the costs for determining the optimal inexact match between  $H$  and  $G$ . Then,  $d(H, G)$  is a graph metric.

Many other graph similarity or distance measures and methods can be found in, e.g. [4, 17, 44, 60, 67, 71, 72].

### 11.4.2 Graph Similarity Measures Based on Trees

In this section, we outline graph similarity measures applied to web-based document structures. As follows, we express a listing of graph similarity measures which have been applied to DOM-trees [13]:

1. Similarity measures which are based on tree edit measures, e.g., see [41, 69, 74].
2. Similarity measures based on the frequency of tag labels, e.g., see [13].
3. Similarity measures based on Fourier transformation, e.g., see [32].
4. Similarity measures based on path similarity, e.g., see [42].

A major problem of these measures is that they only operate on ordinary rooted trees which do not capture the structural information properly represented by a complex hyperlink structure associated to a graph-based document. Especially the measures based on tag frequencies, see, e.g., [13] are restrictively interpretable because a rearrangement of the tag order does not necessarily imply a variation of the corresponding similarity measure. Moreover, the sketched measures do not provide the option to emphasize certain structural properties when measuring the structural similarity of graphs because the measures are non-parameterized. In contrast, parameterized similarity measures would give us the possibility to learn the parameters by using appropriate data sets. In Section 11.4.3, we express the definition of such a parameterized measure for determining the structural similarity of generalized trees. An in-depth treatment of graph similarity measures can be found in [11, 12, 18, 22].

### 11.4.3 Structural Similarity of Generalized Trees

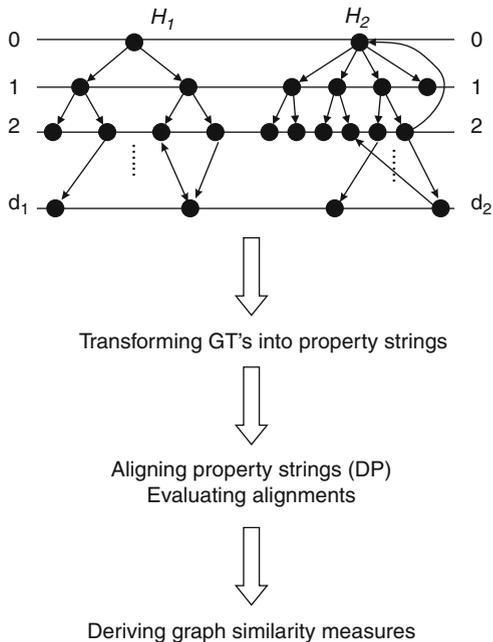
This section aims to repeat the construction principle of a method for measuring the structural similarity of generalized trees, see, e.g. [18, 22, 27]. The main construction steps can be stated as follows [18, 22, 27]:

- We start with two generalized trees,  $H_1$  and  $H_2$ .
- Derive their formal string representations and transform them into linear integer strings which are called property strings.
- Perform string alignments of the derived property strings by using a dynamic programming (DP) algorithm. From each such alignment (on each level  $i$ ), a similarity score will be obtained.
- By cumulating up the derived similarity scores, a final graph similarity measure can be obtained. Hence, the problem of comparing two generalized trees structurally is then equivalent with determining optimal property string alignments.

These key steps are also visualized in Fig. 11.3. We start repeating the construction by stating some definitions [18, 21, 22].

**Definition 9** Let  $X$  be a set. A positive function  $s : X \times X \rightarrow [0, 1]$  is called similarity measure if

**Fig. 11.3** Key steps to infer a graph similarity measure for generalized trees



- $s(x, y) > 0 \quad \forall x, y \in X.$
- $s(x, y) = s(y, x) \quad \forall x, y \in X.$
- $s(x, y) \leq s(x, x) = 1 \quad \forall x, y \in X.$

**Definition 10** Let  $X$  be a set. A positive function  $\omega : X \times X \rightarrow [0, 1]$  is called distance measure if

- $\omega(x, y) \geq 0 \quad \forall x, y \in X.$
- $\omega(x, y) = \omega(y, x) \quad \forall x, y \in X.$
- $\omega(x, x) = 0 \quad \forall x \in X.$

**Definition 11** Let  $H$  be a generalized tree. We call the set

$$S^H := \left\{ v_{0,1}^H, v_{1,1}^H \circ v_{1,2}^H \circ \dots \circ v_{1,|V_1|}^H, \dots, v_{d,1}^H \circ v_{d,2}^H \circ \dots \circ v_{d,|V_d|}^H \right\}, \quad (11.7)$$

the formal string representation of  $H$ . The symbol  $\circ$  denotes usual string concatenation.

**Definition 12** Let  $H$  be a generalized tree. We call

$$S_{\text{out}}^H := \left\{ \delta_{\text{out}}(v_{0,1}^H), \delta_{\text{out}}(v_{1,1}^H) \circ \delta_{\text{out}}(v_{1,2}^H) \circ \dots \circ \delta_{\text{out}}(v_{1,|V_1|}^H), \dots, \delta_{\text{out}}(v_{d,1}^H) \circ \delta_{\text{out}}(v_{d,2}^H) \circ \dots \circ \delta_{\text{out}}(v_{d,|V_d|}^H) \right\}, \quad (11.8)$$

the set of out-degree property strings and

$$S_{\text{in}}^H := \left\{ \delta_{\text{in}} \left( v_{0,1}^H \right), \delta_{\text{in}} \left( v_{1,1}^H \right) \circ \delta_{\text{in}} \left( v_{1,2}^H \right) \circ \cdots \circ \delta_{\text{in}} \left( v_{1,|V_1|}^H \right), \dots, \right. \\ \left. \circ \delta_{\text{in}} \left( v_{d,1}^H \right) \circ \delta_{\text{in}} \left( v_{d,2}^H \right) \circ \cdots \circ \delta_{\text{in}} \left( v_{d,|V_d|}^H \right) \right\}, \quad (11.9)$$

the set of in-degree property strings of  $H$ .

Define  $r_k^{H^k} := v_{0,1}^{H^k}$ ,  $k \in \{1, 2\}$ . Let  $H^1$  be a given GT and  $v_{i,j}^{H^1}$ ,  $0 \leq i \leq d_1$ ,  $1 \leq j \leq \sigma_i$  denotes the  $j$ -th vertex on the  $i$ -th level of  $H^1$ . Analogously, this also holds for  $v_{i,j}^{H^2} \in H^2$ . As mentioned above, the task of measuring the structural similarity between  $H^1$  and  $H^2$  is equivalent to determine the optimal alignment of

$$S_1 = v_{0,1}^{H^1} \circ v_{1,1}^{H^1} \circ v_{1,2}^{H^1} \circ \cdots \circ v_{d_1,\sigma_{d_1}}^{H^1}, \\ S_2 = v_{0,1}^{H^2} \circ v_{1,1}^{H^2} \circ v_{1,2}^{H^2} \circ \cdots \circ v_{d_2,\sigma_{d_2}}^{H^2},$$

with respect to their associated property strings and to a cost function  $\alpha$ .  $S_k[i]$  denotes the  $i$ -th position of the sequence  $S_k$  and it holds  $S_1[n] = v_{d_1,\sigma_{d_1}}^{H^1}$ ,  $S_2[m] = v_{d_2,\sigma_{d_2}}^{H^2}$ ,  $\mathbb{N} \ni n, m \geq 1$ ,  $S_k[1] = r_k^{H^k}$ ,  $k \in \{1, 2\}$ . The algorithm for finding the optimal alignment of  $S_1$  and  $S_2$  generates a matrix  $(\mathcal{M}(i, j))_{i,j}$ ,  $0 \leq i \leq n$ ,  $0 \leq j \leq m$ . We find that its time complexity is  $O(|\hat{V}_1| \cdot |\hat{V}_2|)$ , see [18, 23]. To determine optimal alignment of the derived property strings, we state the following algorithm [18, 23]:

$$\begin{aligned} \mathcal{M}(0, 0) &:= 0, \\ \mathcal{M}(i, 0) &:= \mathcal{M}(i-1, 0) + \alpha(S_1[i], -) : 1 \leq i \leq n, \\ \mathcal{M}(0, j) &:= \mathcal{M}(0, j-1) + \alpha(-, S_2[j]) : 1 \leq j \leq m, \end{aligned}$$

and

$$\mathcal{M}(i, j) := \min \begin{cases} \mathcal{M}(i-1, j) + \alpha(S_1[i], -) \\ \mathcal{M}(i, j-1) + \alpha(-, S_2[j]) \\ \mathcal{M}(i-1, j-1) + \alpha(S_1[i], S_2[j]) \end{cases}$$

for  $1 \leq i \leq n$ ,  $1 \leq j \leq m$ . Here, the derived property strings will be aligned on two levels: globally and locally. To evaluate the alignments, we need the preliminary assertion as follows.

**Lemma 1** Let  $\omega(x, y) := 1 - e^{-\frac{1}{2} \frac{(x-y)^2}{\sigma^2}}$ .  $\omega : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$  is a distance measure.

*Proof* From the definition of  $\omega(x, y)$  we infer  $\omega(x, y) \in [0, 1]$ ,  $\forall x, y \in \mathbb{R}$  and  $\omega(x, x) = 1 - 1 = 0$ ,  $\forall x \in \mathbb{R}$ . Since  $(x-y)^2 = (y-x)^2$ ,  $\forall x, y \in \mathbb{R}$ , the symmetry condition holds.

Now, we define

$$\alpha^{\text{out}}\left(v_{i_1, j_1}^{H^1}, v_{i_2, j_2}^{H^2}\right) := \begin{cases} \omega^{\text{out}}\left(\delta_{\text{out}}\left(v_{i_1, j_1}^{H^1}\right), \delta_{\text{out}}\left(v_{i_2, j_2}^{H^2}\right), \sigma_{\text{out}}^1\right) & : i_1 = i_2 \\ +\infty & : \text{else,} \end{cases}$$

$0 \leq i_k \leq d_k, 1 \leq j_k \leq \sigma_{i_k}, k \in \{1, 2\}$ , where  $\omega^{\text{out}}(x, y, \sigma_{\text{out}}^k) := 1 - e^{-\frac{1}{2}(x-y)^2/(\sigma_{\text{out}}^k)^2}$ ,  $x, y, \sigma_{\text{out}}^k \in \mathbb{R}$ , and

$$\begin{aligned} \alpha^{\text{out}}\left(v_{i, j_1}^{H^1}, -\right) &:= \omega^{\text{out}}\left(\delta_{\text{out}}\left(v_{i, j_1}^{H^1}\right), \xi, \sigma_{\text{out}}^2\right), \\ \alpha^{\text{out}}\left(-, v_{i, j_2}^{H^2}\right) &:= \omega^{\text{out}}\left(\xi, \delta_{\text{out}}\left(v_{i, j_2}^{H^2}\right), \sigma_{\text{out}}^2\right). \end{aligned}$$

$\xi > 0$  prevents an alignment between two leaves being better evaluated as an alignment between a leaf and a gap (“-”) [22]. By  $\omega^{\text{in}}(x, y, \sigma_{\text{in}}^k) := 1 - e^{-\frac{1}{2}(x-y)^2/(\sigma_{\text{in}}^k)^2}$ , we define analogously  $\alpha^{\text{in}}\left(v_{i_1, j_1}^{H^1}, v_{i_2, j_2}^{H^2}\right)$ ,  $\alpha^{\text{in}}\left(v_{i, j_1}^{H^1}, -\right)$  and  $\alpha^{\text{in}}\left(-, v_{i, j_2}^{H^2}\right)$ .

To evaluate the alignments of the property strings locally (i.e., on each generalized tree level), we express the mapping [18, 22]

$$\text{align}\left(v_{i, j_1}^{H^1}\right) := \begin{cases} v_{i, j_2}^{H^2} & : \text{align}^{-1}\left(v_{i, j_2}^{H^2}\right) = v_{i, j_1}^{H^1} \\ - & : \text{else.} \end{cases}$$

For  $v_{i, j_1}^{H^1}$ , the mapping determines the vertex  $v_{i, j_2}^{H^2}$  during the trace-back [18]. Moreover, we define the functions

$$\begin{aligned} \gamma_{H^k}^{\text{out}}(i) &:= \frac{\sum_{j=1}^{\sigma_i^k} \hat{\alpha}_{\text{out}}\left(v_{i, j}^{H^k}, \text{align}\left(v_{i, j}^{H^k}\right)\right)}{\sigma_i^k}, \\ \gamma_{H^k}^{\text{in}}(i) &:= \frac{\sum_{j=1}^{\sigma_i^k} \hat{\alpha}_{\text{in}}\left(v_{i, j}^{H^k}, \text{align}\left(v_{i, j}^{H^k}\right)\right)}{\sigma_i^k}, \end{aligned}$$

$k \in \{1, 2\}$ , which provide similarity values of the alignments of out-degree and in-degree property strings. Finally, by analogously defining the functions  $\hat{\alpha}_{\text{out}}$  and  $\hat{\alpha}_{\text{in}}$ , we obtain the normalized and cumulative functions

$$\begin{aligned} \gamma^{\text{out}}\left(i, \hat{\sigma}_{\text{out}}^1, \hat{\sigma}_{\text{out}}^2\right) &:= 1 - \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^1} \hat{\alpha}_{\text{out}}\left(v_{i, j}^{H^1}, \text{align}\left(v_{i, j}^{H^1}\right)\right) \right\} \\ &\quad - \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^2} \hat{\alpha}_{\text{out}}\left(v_{i, j}^{H^2}, \text{align}\left(v_{i, j}^{H^2}\right)\right) \right\}, \quad (11.10) \end{aligned}$$

and

$$\gamma^{\text{in}}\left(i, \hat{\sigma}_{\text{in}}^1, \hat{\sigma}_{\text{in}}^2\right) := 1 - \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^1} \hat{\alpha}^{\text{in}}\left(v_{i,j}^{H^1}, \text{align}\left(v_{i,j}^{H^1}\right)\right) \right\} \\ - \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^2} \hat{\alpha}^{\text{in}}\left(v_{i,j}^{H^2}, \text{align}\left(v_{i,j}^{H^2}\right)\right) \right\}, \quad (11.11)$$

which detect the similarity of an out-degree and in-degree alignment on a level  $i$ .  $\hat{\sigma}_{\text{out}}^1, \hat{\sigma}_{\text{out}}^2$  and  $\hat{\sigma}_{\text{in}}^1, \hat{\sigma}_{\text{in}}^2$  are the parameters of  $\hat{\alpha}^{\text{out}}$  and  $\hat{\alpha}^{\text{in}}$ , respectively. By using the defined quantities, it can be proven that the resulting comparative measure is a graph similarity measure (i.e., the measure satisfies the properties of Definition (9)) [18, 22].

**Theorem 4** *Let  $H_1, H_2$  be two generalized trees,  $0 \leq i \leq \mu$ ,  $\mu := \max(d_1, d_2)$ . Then,*

$$s(H_1, H_2) := \frac{(\mu + 1)}{\sum_{i=0}^{\mu} \gamma^{\text{fin}}\left(i, \hat{\sigma}_{\text{out}}^1, \hat{\sigma}_{\text{out}}^2, \hat{\sigma}_{\text{in}}^1, \hat{\sigma}_{\text{in}}^2\right)} \prod_{i=0}^{\mu} \gamma^{\text{fin}}\left(i, \hat{\sigma}_{\text{out}}^1, \hat{\sigma}_{\text{out}}^2, \hat{\sigma}_{\text{in}}^1, \hat{\sigma}_{\text{in}}^2\right), \quad (11.12)$$

is a graph similarity measure where  $\gamma^{\text{fin}}$  is defined by

$$\gamma^{\text{fin}} = \gamma^{\text{fin}}\left(i, \hat{\sigma}_{\text{out}}^1, \hat{\sigma}_{\text{out}}^2, \hat{\sigma}_{\text{in}}^1, \hat{\sigma}_{\text{in}}^2\right) \\ := \zeta \cdot \gamma^{\text{out}} + (1 - \zeta) \cdot \gamma^{\text{in}}, \quad \zeta \in [0, 1].$$

## 11.5 Applications

In the following, we outline existing and future applications of our presented approach which we have stated in Section 11.4.3. Here, we represent websites as a graph-based model [57] where we map each document structure to a generalized tree. In [22], a family of graph similarity measures was evaluated based on a corpus containing 500 conference websites from mathematics and computer science created by Mehler et al. [57]. Finally, the conference websites were inferred from the web and transformed into generalized trees by using the tool HyGraph [35, 36].

One of the main ideas is to apply a comparative analysis to a corpus consisting of graph-based web units. Now, for automatically analyzing web genre data, we propose the following evaluation steps:

1. Because the graph similarity measure outlined in Section 11.4.3 is parameterized, one can emphasize structural features of the graphs under consideration when measuring their structural similarity [27, 22]. This can be done by varying the parameters  $(\zeta, \hat{\sigma}_{\text{out}}^1, \hat{\sigma}_{\text{out}}^2, \hat{\sigma}_{\text{in}}^1, \hat{\sigma}_{\text{in}}^2)$ . For example in [27, 22], we have shown

that by setting  $\zeta$  equal to 1 or 0, we either consider the alignments of out-degree or in-degree property strings only. To set  $\zeta = \frac{1}{2}$  means that we weight the out-degree and in-degree property strings equally [22, 27].

2. We calculate the complete similarity matrix by computing the pairwise similarity scores of the given generalized trees. For this, we use the graph similarity measure presented in Section 11.4.3 with a fixed parameter set [22]. Moreover, we can compute the so-called cumulative similarity distribution  $\Theta$  usually depicted as a two-dimensional plot.  $\Theta$  can be used for expressing the percentage of generalized trees which possess a similarity value less or equal  $s \in [0, 1]$  and, hence, to answer the question how structurally different the document structures of a given corpus are [22, 27]. Generally, we consider the study of  $\Theta$  as a preliminary step for automatically analyzing web genre data that already led to a better understanding of the problem of comparing web-based hypertexts structurally [18, 22, 27].
3. Starting from a computed similarity matrix, one can additionally apply multivariate analysis methods, e.g., clustering techniques to filter web-based documents. By determining such clusters one identifies websites of similar structure, i.e., these clusters contain structurally similar web pages [18].

From the just outlined steps, it should be clear that this approach can also be used for analyzing data sets of hypertext structures inferred from other Web Mining areas. For example, if it would be possible to transform weblog data sets into sets of generalized trees, we could apply the approach analogously. This would result to novel applications in Web Usage Mining. In [18, 27] it has been sketched that the focus of such a study would be to analyze the navigation behavior of hypertext users [61, 66]. Generally, navigation patterns can be described by graphs [61, 66]. Particularly in our case, we would describe those by generalized trees. Each cluster we could determine by using the above stated approach then contains generalized trees which reflect a similar navigation behavior of a specific user. As we have already outlined in [18, 27], a possible interpretation of these clusters can lead to study psychological features of hypertext users.

## 11.6 Conclusion

The main goal of this conceptual chapter was to present an approach for automatically analyzing web genre data representing graphs. Instead of using the well-known vector space model for modeling document structures, we applied a graph-based representation model proposed by Mehler et al. [57]. A notable feature of this model is that the document structures represented by generalized trees capture more structural information than DOM-trees [18, 36, 57]. In Section 11.4.2, we briefly reviewed methods to measure the structural similarity of web-based documents which operate on tree structures only. In contrast to this, in Section 11.4.3 we repeated an approach for measuring the structural similarity of generalized trees. A key feature of this method is that the graphs will be transformed into linear integer

strings. By applying a string alignment algorithm, we weighted these alignments and finally derived a graph similarity measure for generalized trees. Hence, we solved a graph similarity problem by transforming it into a string similarity problem. Section 11.5 presented an overview of possible evaluation steps for automatically analyzing web genre data representing graphs. Moreover, existing applications of this approach were discussed.

**Acknowledgments** We are thankful to Alexander Mehler for fruitful discussions on this topic.

## References

1. Albert, R., H. Jeong, and A.L. Barabási. 1999. Diameter of the world wide web. *Nature* 401:130–131.
2. Baeza-Yates, R., and B. Ribeiro-Neto, eds. 1999. *Modern information retrieval*. Reading, MA: Addison-Wesley.
3. Barabási, A.-L., and Z.N. Oltvai. 2004. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.
4. Basak, S.C., V.R. Magnuson, G.J. Niemi, and R.R. Regal. 1988. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Applied Mathematics* 19:17–44.
5. Batagelj, V. 1988. Similarity measures between structured objects. In *Proceedings of an International Course and Conference on the Interfaces between Mathematics, Chemistry and Computer Sciences*. Dubrovnik, Yugoslavia.
6. Bonchev, D. 1979. Information indices for atoms and molecules. *MATCH* 7:65–113.
7. Bonchev, D. 1983. *Information theoretic indices for characterization of-chemical structures*. Chichester: Research Studies Press.
8. Bornholdt, S., and H.G. Schuster. 2003. *Handbook of graphs and networks. From the genome to the Internet*. Weinheim: Wiley-VCH.
9. Brandes, U., and T. Erlebach. 2005. *Network analysis*. Lecture Notes in Computer Science. Heidelberg: Springer.
10. Bunke, H. 1983. What is the distance between graphs? *Bulletin of the EATCS* 20:35–39.
11. Bunke, H. 2000a. Recent developments in graph matching. In *Proceedings of the 15th International Conference on Pattern Recognition* 2:117–124.
12. Bunke, H. 2000b. Graph matching: Theoretical foundations, algorithms, and applications. In *Proceedings of Vision Interface 2000*, 82–88. Montreal, Canada.
13. Buttler, D. 2004. A short survey of document structure similarity algorithms. In *International Conference on Internet Computing*, 3–9. Los Vegas, Nevada, USA.
14. Carrière, S.J., and R. Kazman. 1997. Webquery: Searching and visualizing the web through connectivity. *Computer Networks and ISDN Systems* 29(8–13):1257–1267.
15. Chakrabarti, S. 2001. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proceedings of the 10th International World Wide Web Conference*, May 1–5, 211–220. Hong Kong.
16. Chakrabarti, S. 2002. *Mining the web: Discovering knowledge from hypertext data*. San Francisco, CA: Morgan Kaufmann.
17. Cook, D., and L.B. Holder. 2007. *Mining graph data*. Weinheim: Wiley-Interscience.
18. Dehmer, M. 2006. *Strukturelle analyse web-basierter Dokumente. Multimedia und Telekooperation*. Wiesbaden: Deutscher Universitäts Verlag.
19. Dehmer, M. 2008a. Information-theoretic concepts for the analysis of complex networks. *Applied Artificial Intelligence* 22(7 and 8):684–706.
20. Dehmer, M. 2008b. Information processing in complex networks: graph entropy and information functionals. *Applied Mathematics and Computation* 201:82–94.

21. Dehmer, M., and F. Emmert-Streib. 2007. Structural similarity of directed universal hierarchical graphs: A low computational complexity approach. *Applied Mathematics and Computation* 194:7–20.
22. Dehmer, M., and A. Mehler. 2007. A new method of measuring similarity for a special class of directed graphs. *Tatra Mountains Mathematical Publications* 36:39–59.
23. Dehmer, M., A. Mehler, and R. Gleim. 2004. Aspekte der Kategorisierung von Webseiten. In *Proceedings des Multimediaworkshops der Jahrestagung der Gesellschaft für Informatik*, eds. P. Dadam und M. Reichert, Lecture Notes in Computer Science, vol. 2, 39–43, Berlin: Springer.
24. Dehmer, M., F. Emmert-Streib, and J. Kilian. 2006. A similarity measure for graphs with low computational complexity. *Applied Mathematics and Computation* 182:447–459.
25. Dehmer, M., A. Mehler, and F. Emmert-Streib. 2007. Graphtheoretical characterizations of generalized trees. In *Proceedings of the International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07)*. Las Vegas, NV.
26. Dehmer, M., F. Emmert-Streib, and T. Gesell. 2008. A comparative analysis of multidimensional features of objects resembling sets of graphs. *Applied Mathematics and Computation* 196:221–235.
27. Dehmer, M., F. Emmert-Streib, A. Mehler, and J. Kilian. 2006. Measuring the structural similarity of web-based documents: A novel approach. *International Journal of Computational Intelligence* 3(1):1–7.
28. Dimter, M. 1981. *Textklassenkonzepte heutiger Alltagssprache*. Tübingen: Niemeyer.
29. Dorogovtsev, S.N., and J.F.F. Mendes. 2003. *Evolution of networks. From biological networks to the internet and WWW*. Oxford: Oxford University Press.
30. Emmert-Streib, F., and M. Dehmer. 2007. Information theoretic measures of UHG graphs with low computational complexity. *Applied Mathematics and Computation* 190:1783–1794.
31. Ferber, R. 2003. *Information retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. Heidelberg: dpunkt.verlag.
32. Flesca, S., G. Manco, E. Masciari, L. Pontieri, and A. Pugliese. 2002. Detecting structural similarities between XML documents. In *Proceedings of the International Workshop on the Web and Databases (WebDB 2002)*. Madison, Wisconsin, USA.
33. Foulds, L.R. 1992. *Graph theory applications*. New York, NY: Springer.
34. Gibson, D., R. Kumar, K.S. McCurley, and A. Tomkins. 2007. Dense subgraph extraction. In *Mining graph data*, eds. D. Cook and L.B. Holder, 411–441. Hoboken, NJ: Wiley-Interscience.
35. Gleim, R. 2004. Integrierte Repräsentation, Kategorisierung und Strukturanalyse Webbasierter Hypertexte. Master's thesis, Technische Universität Darmstadt, Fachbereich Informatik, Sept 2004.
36. Gleim, R. 2005. HyGraph: Ein Framework zur Extraktion, Repräsentation und Analyse webbasierter Hypertexte. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, eds. B. Fisseni, H.-C. Schmitz, B. Schröder, and P. Wagner, 42–53. Frankfurt a.M.: Lang.
37. Halin, R. 1989. *Graphentheorie*. Berlin: Akademie Verlag.
38. Han, J., and M. Kamber. 2001. *Data mining: Concepts and techniques*. New York, NY: Morgan and Kaufmann Publishers.
39. Harary, F. 1969. *Graph theory*. Reading, MA: Addison Wesley Publishing Company.
40. Huberman, B., and L. Adamic. 1999. Growth dynamics of the world-wide web. *Nature*, 399:130.
41. Jiang, T., L. Wang, and K. Zhang. 1994. Alignment of trees – an alternative to tree edit. In *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, 75–86, London: Springer-Verlag.
42. Joshi, S., N. Agrawal, R. Krishnapuram, and S. Negi. 2003. A bag of paths model for measuring structural similarity in web documents. In *KDD '03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 577–582, New York, NY.

43. Kaden, F. 1982. Graphmetriken und Distanzgraphen. *ZKI-Informationen, Akademie der Wissenschaften der DDR* 2(82):1–63.
44. Kaden, F. 1986. Graphmetriken und Isometrie Probleme zugehöriger Distanzgraphen. *ZKI-Informationen, Akademie der Wissenschaften der DDR* 1(P6):1–100.
45. Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5):604–632.
46. Kosala, R., and H. Blockeel. 2000. Web mining research: A survey. SIGKDD explorations: Newsletter of the Special Interest Group (SIG) on knowledge discovery & data mining, *ACM* 2(1):1–15.
47. Koschützki, D., K.A. Lehmann, L. Peters, S. Richter, D. Tenfelde-Podehl, and O. Zlotkowski. 2005. Clustering. In *Centrality indices*, eds. U. Brandes and T. Erlebach, Lecture Notes of Computer Science, 16–61. Berlin: Springer.
48. Kumar, R., P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal. 2000. The web as a graph. In *PODS '00: Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 1–10, New York, NY: ACM Press.
49. Levenstein, V.I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics – Doklady* 10(8):707–710, Feb 1966.
50. Lindemann, C., and L. Littig. 2010. Classification of web sites at super-genre level. In *Genres on the web: Computational models and empirical studies*, eds. A. Mehler, S. Sharoff, and M. Santini, Text, Speech and Language Technology. Dordrecht: Springer.
51. Mason, O., and M. 2007. Verwoerd. Graph theory and networks in biology. *IET Systems Biology* 1(2):89–119.
52. Mehler, A. 2001. *Textbedeutung. Zur prozeduralen Analyse und Repräsentation struktureller Ähnlichkeiten von Texten*, volume 5 of *Sprache, Sprechen und Computer/Computer Studies in Language and Speech*. Frankfurt a. M.: Peter Lang.
53. Mehler, A. 2004. Textmining. In *Texttechnologie. Perspektiven und Anwendungen*, eds. H. Lobin and L. Lemnitzer, 83–107. Tübingen: Stauffenburg.
54. Mehler, A. 2009. Generalized shortest paths trees: A novel graph class applied to semiotic networks. In *Analysis of complex networks: From biology to linguistics*, eds. M. Dehmer and F. Emmert-Streib, 175–220. Weinheim: Wiley-VCH.
55. Mehler, A. 2010. Structure formation in the web. toward a graphtheoretical model of hypertext types. In *Linguistic modelling of information and markup languages*, eds. A. Witt and D. Metzger, 225–247. Dordrecht: Springer.
56. Mehler, A., and R. Gleim. 2006. The net for the graphs – towards webgenre representation for corpus linguistic studies. In *WaCky! Working papers on the web as corpus*, eds. M. Baroni and S. Bernardini, 191–224. Bologna: Gedit.
57. Mehler, A., M. Dehmer, and R. Gleim. 2004. Towards logical hypertext structure – A graphtheoretic perspective. In *Proceedings of the Fourth International Workshop on Innovative Internet Computing Systems (I2CS '04)*, eds. T. Böhme and G. Heyer, Lecture Notes in Computer Science, vol. 3473, 136–150, Berlin/New York: Springer.
58. Mehler, A., R. Gleim, and M. Dehmer. 2005. Towards structure-sensitive hypertext categorization. In *Proceedings of the 29th Annual Conference of the German Classification Society*, LNCS, Mar 9–11. Universität Magdeburg, Berlin/New York, NY: Springer.
59. Mehler, A., R. Gleim, and A. Wegner. 2007. Structural uncertainty of hypertext types. An empirical study. In *Proceedings of the Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP"*, 30 Sept 2007, 13–19, in conjunction with RANLP 2007. Borovets, Bulgaria.
60. Messmer, B.T., and H. Bunke. 1998. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(5):493–504.
61. Noller, S., J. Naumann, and T. Richter. 2001. LOGPAT – Ein webbasiertes Tool zur Analyse von Navigationsverläufen in Hypertexten. <http://www.psych.uni-goettingen.de/congress/gor-2001>
62. Power, R., D. Scott, and N. Bouayad-Agha. 2003. Document structure. *Computational Linguistics* 29(2):211–260.

63. Raghavan, P. 2000. Graph structure of the web: A survey. In *LATIN 2000: Theoretical Informatics. Proceedings of 4th Latin American Symposium*, 123–125. Punta del Este, Uruguay.
64. Rahm, E. 2002. Web usage mining. *Datenbank-Spektrum* 2(2):75–76.
65. Rehm, G. 2007. *Hypertextsorten. Definition – Struktur – Klassifikation*. Norderstedt: Books on Demand.
66. Richter, T., J. Naumann, and S. Noller. 2003. Logpat: A semi-automatic way to analyze hyper-text navigation behavior. *Swiss Journal of Psychology* 62:113:120.
67. Schädler, C. 1999. Die Ermittlung struktureller Ähnlichkeit undstruktureller-Merkmale bei komplexen Objekten: Einkonnektionistischer Ansatz und seine Anwendungen. PhD thesis, Technische Universität Berlin.
68. Scsibrany, H., K. Karlovits, W. Demuth, F. Müller, and K. Varmuza. 2003. Clustering and similarity of chemical structures represented by binary substructure descriptors. *Chemometrics and Intelligent Laboratory Systems* 67:95–108.
69. Selkow, S.M. 1977. The tree-to-tree editing problem. *Information Processing Letters* 6(6):184–186.
70. Skorobogatov, V.A., and A.A. Dobrynin. 1988. Metrical analysis of graphs. *MATCH* 23:105–155.
71. Sobik, F. 1982. Graphmetriken und Klassifikation strukturierter Objekte. *ZKI-Informationen, Akademie der Wissenschaften der DDR* 2(82):63–122.
72. Sobik, F. 1986. Modellierung von Vergleichsprozessen auf der Grundlage von Ähnlichkeitsmaßen für Graphen. *ZKI-Informationen, Akademie der Wissenschaften der DDR* 4:104–144.
73. Spiliopoulou, M. 2000. Web usage mining for web site evaluation. *Communications of the ACM* 43(8):127–134.
74. Tai, K.C. 1979. The tree-to-tree correction problem. *Journal of the ACM* 26(3):422–433. ISSN 0004-5411.
75. Waltinger, U., A. Mehler, and A. Wegner. 2009. A two-level approach to web genre classification. In *Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST '09)*, 23–26 Mar 2009. Lisboa.
76. Wasserman, S., and K. Faust. 1994. *Social network analysis: Methods and applications*, Structural Analysis in the Social Sciences. Cambridge, MA: Cambridge University Press.
77. Zelinka, B. 1975. On a certain distance between isomorphism classes of graphs. *Časopis pro řest. Matematiky* 100:371–373.



# Chapter 12

## Genre Connectivity and Genre Drift in a Web of Genres

Lennart Björneborn

### 12.1 Introduction

The World-Wide Web may be conceived as a web of genres; a network of different web page genres connected by links. The chapter outlines an exploratory empirical investigation of genre connectivity in an academic web space. The investigation formed part of a webometric<sup>1</sup> study [3] concerned with what types of web links, web pages and web sites function as cross-topic connectors affecting so-called small-world phenomena in the shape of short link distances along link paths connecting different topical domains in an academic web space. To the author's knowledge, this is one of the first studies to include genre connectivity on the Web.

Most links within and between web sites connect web pages with similar topics (e.g. [8]) leading to topically clustered aggregations of interlinked web pages and web sites. At the same time, some links connect different topical clusters. Such cross-topic links – and web page genres containing cross-topic links – were in focus in the study outlined in this chapter as they function as small-world shortcuts between different web clusters.

Small-world theory stems from research in social network analysis on short distances between two arbitrary persons through intermediate chains of acquaintances in social networks [12]. Watts and Strogatz [19] revived small-world theory by introducing a small-world network model characterized by a combination of highly clustered network nodes and short average path lengths between arbitrary pairs of network nodes. Subsequent research has revealed small-world properties in a large variety of networks, including biochemical, neural, ecological, technical, social, economical, and informational networks. For example, scientific collaboration networks, citation networks and semantic networks have small-world features [13, 15].

---

L. Björneborn (✉)  
Royal School of Library and Information Science, Copenhagen, Denmark  
e-mail: lb@iva.dk

<sup>1</sup> Webometrics is the study of quantitative aspects of information resources, structures and technologies on the Web, drawing on bibliometric and informetric approaches from Library and Information Science [5].

Containing both high local clusterization and short global separation, small-world networks simultaneously have small local *and* small global distances that facilitate high efficiency in disseminating information, ideas, contacts, signals, energy, viruses, etc., both on a local and global scale in the concerned networks. On the web, small-world link structures affect reachability structures [4], that is, how web users and web crawlers may reach and retrieve web resources when following links from web page to web page – from genre to genre. Small-world implications of genre connectivity are discussed in this chapter in Section 12.3.4.

## 12.2 Methodology

In the chapter, focus is not on small-world link structures as such, but on genre connectivity. However, as the data set used for investigating genre connectivity formed part of a study [3] with special focus on small-world link structures, the overall methodology in that study will be briefly outlined here.

A network analytic approach was used in the study with regard to collecting, extracting and analyzing web data. The data set in the study was harvested in July 2001 from 109 UK university web sites by a special web crawler [16] and comprised all identified 7,669 university subsites, i.e. departments, research groups, etc., with derivative university domain names (e.g., *www.geog.plym.ac.uk*: Geography Department, University of Plymouth).

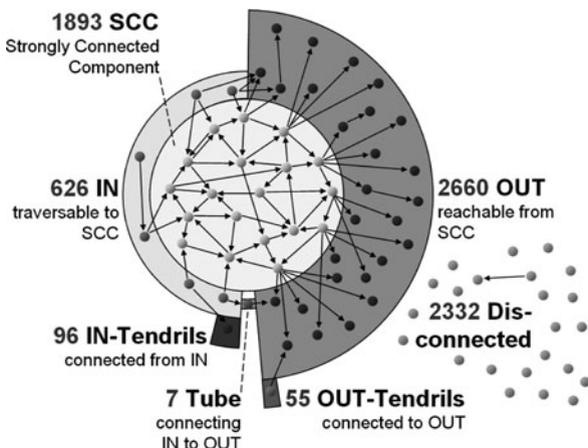
A five-step methodology was developed in the study to sample and identify small-world properties by zooming stepwise into more and more fine-grained web node levels and link structures in the data set.

The first step identified overall web graph components and reachability structures among the 7,669 subsites as shown in Fig. 12.1. In the strongly connected component (SCC), any subsite node can be reached from any other node along intermediate link paths. As illustrated in Fig. 12.1, there are link paths in both directions between any pair of SCC nodes; enabling small-world features in the shape of short link distances (degrees of separation) between these nodes. In the other components, there are only link paths in one direction – or nodes are disconnected from the other components.<sup>2</sup>

From the 1,893 SCC subsites identified in the first methodological step, a random sample of 189 subsites was examined in the second step in order to classify overall subsite topics. In the third step, a stratified sample was extracted consisting of five SCC subsites from different domains in humanities and social sciences randomly paired with five SCC subsites from natural sciences and technology. Adding the reverted order of start and end nodes, this step resulted in 10 “path nets” (Fig. 12.2) comprising all shortest link paths between SCC subsites belonging to dissimilar topics. The 10 path nets were constructed to function as investigable small-world

---

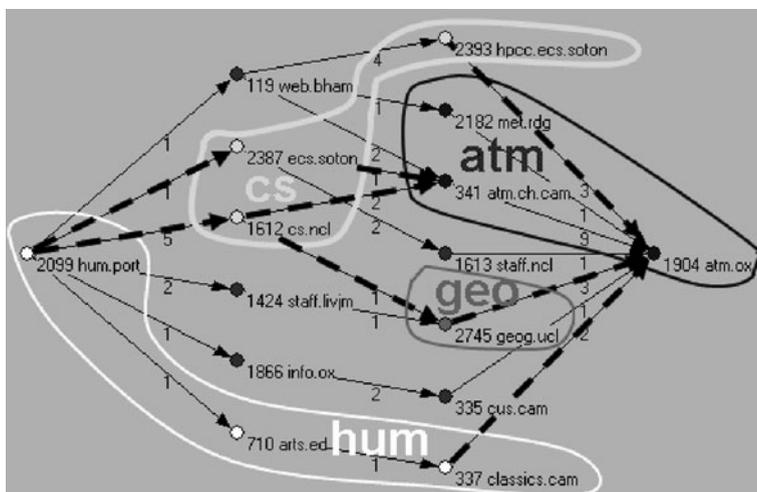
<sup>2</sup> See [3] and [4] for more details on the graph components in the model.



**Fig. 12.1** “Corona” model [3] of web graph components and reachability structures in the data set of 7,669 UK university subsites with numbers of subsites in each component and outlining some possible link paths within and between the components. The model is modified after Broder et al. [6]

link structures – as “mini small worlds” [4] – generated by deliberate juxtaposition of topically dissimilar seed nodes.

In the fourth step, genres of source and target pages along link paths in the 10 path nets were classified (cf. Section 12.2.2). The objective of all these steps was to lead up to the final step concerned with identifying what types of web links, pages,



**Fig. 12.2** Path net with topical areas humanities (hum), computer science (cs), geography (geo) and atmospheric sciences (atm) (see Appendix 10: path net “HN01” in [3] for affiliations). Non-enclosed nodes are campus-wide generic-type subsites. Counts of links are noted. Cross-topic links are marked with *dashed bold*

and sites function as cross-topic connectors in small-world link structures across an academic web space, the main research question in the study. The data subset consisting of 10 path nets was thus also used to investigate genre connectivity in the study.<sup>3</sup>

The network analysis software *Pajek*<sup>4</sup> was used to construct the path nets as subgraphs based on the adjacency matrix of how the 7,669 subsites were interconnected in the data set. As previously noted, a path net contains all shortest link paths between two given nodes. Figure 12.2 shows one of the 10 path nets. All nodes are subsites at different UK universities and the path net in the figure shows all the shortest link paths (path length 3) in the data set between node 2,099, the Faculty of Humanities and Social Sciences in Portsmouth ([www.hum.port.ac.uk](http://www.hum.port.ac.uk)), and node 1904, the Atmospheric Science subsite in Oxford ([www.atm.ox.ac.uk](http://www.atm.ox.ac.uk)). One of the link paths includes the start node and passes two other subsites in the humanities (*hum*). Two of the neighbor nodes to the end node in the path net also are atmospheric science subsites (*atm*). Three computer-science subsites (*cs*) function as connectors on some of the shortest link paths between the start node and the end node. There is also one geography subsite (*geo*) being a connector node. The dashed links denote cross-topic links connecting one topical area with another in the path net. For example, there is a link from an institutional link list at the Faculty of Classics in Cambridge (node 337) to an oceanographic researcher at the atmospheric science subsite in Oxford having a hobby web page devoted to an ancient Greek trireme warship.

### 12.2.1 Source Pages and Target Pages

In Fig. 12.2 links are shown between nodes on the *subsite level*. Figure 12.3 shows an excerpt of another of the 10 path nets zoomed into the *page level*. The figure gives an illustration of actual links in the data set between source pages and target pages along shortest link paths between the start node and the end node.

A total of 530 web pages comprising 281 source pages (providing outlinks to other subsites in the data set) and 249 target pages (receiving inlinks) were manually examined in the 10 path nets and web page genres were classified by the author. All the web pages were in English. There were 352 links connecting the pages. The dissimilar numbers of pages and links are due to the fact that source pages may have outlinks to more than one target page and target pages may receive inlinks from more than one source page, as may be seen in Fig. 12.3.

---

<sup>3</sup> As noted, the overall study in this chapter had focus on small-world aspects of link structures in academic web spaces. Of course, network analytic approaches to genre connectivity could also be based on data from non-academic web spaces and without focus on small-world aspects of these web spaces.

<sup>4</sup> Available at <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>



**Table 12.1** Genre classes of academic web pages: 9 institutional genres (*i.*) and 8 personal (*p.*)

Genre class	Definition and examples
i.archive	Institutional archive or database (e.g. British National Corpus (linguistics))
i.conf	Conference pages, workshops, seminars, etc. (e.g. program, sessions, lists of delegates, abstracts <sup>a</sup> )
i.generic	Non-research-related, campus-wide service web pages (e.g. Web server statistics)
i.hp	Homepages of academic institutions
i.list	Institutional pages with link lists as main content (e.g. staff lists, publication lists, link-rich (text-sparse) resource guides <sup>b</sup> )
i.proj	Joint and single research groups, including specific projects (e.g. research project descriptions)
i.publ	Institutional publications (e.g. reports, journals)
i.soft	Institutional software programs, download pages, demos, documentation, manuals, FAQs, tutorials
i.teach	Institutional teaching-focused web pages (e.g. course pages, tutorials)
p.archive	Personally maintained archive or database (e.g. mailing list archive, discussion group archive)
p.hobby	Personal hobby web pages (researcher, student) not related with person's main academic activity
p.hp	Personal homepages (researcher, student) e.g. research profile, CV
p.list	Personal pages with link lists as main content (e.g. bibliographies, research-related or teaching-related or leisure-related link lists; link-rich (text-sparse) resource guides)
p.proj	Personal research project pages (researcher, PhD student)
p.publ	Personal publications (e.g. working papers, reports, conference papers, posters)
p.soft	Personal software programs, download pages, demos, documentation, manuals, FAQs, tutorials
p.teach	Personal teaching-focused web pages (e.g. lecturers' pages, tutorials, students' assignments, notes)

<sup>a</sup> Full-text conference papers and posters were placed in the *p.publ* genre.

<sup>b</sup> Text-sparse resource guides resembling link lists were classified as *i.list* or *p.list*. Text-dense resource guides resembling papers or manuals were classified as *i.publ* or *p.publ*.

bundled after the functions and purposes of the examined pages. The genre classes reflect functions and purposes of web pages that would be relevant to look for when searching academic web sites for information on departments, researchers, research projects, publications, teaching, etc. The genre categorization was conducted by the author alone, with the limitations this circumstance implies.

The genre classes partially overlap genre classifications in other studies. For example, Rehm [14] describes a web genre hierarchy on an academic web site in his attempt to build an automatic genre classification tool. Other web genre classifications are made by, e.g. [7, 9, 10, 18]. However, none of these genre classifications had sufficient specificity required in the present investigation of an academic web space.

The purpose of dividing the examined web pages into two main categories, *personal* and *institutional* pages, was to get a picture of how the two categories appear

in academic link structures, especially as providers of links across genres and topics in an academic web space.

*Personal* web pages were defined as personally created pages used for personal academic or non-academic purposes. While primarily located in personal directories on web sites, some personal web pages are copied or moved into institutional directories. All 530 web pages, as well as their parent, sibling and child pages, were visited and manually examined, in order to identify such cross-locations. Page layout, text, links, and web page creator names were useful clues for such identification. Typical personal web pages are personal homepages, hobby pages, personal link lists, personal publications, software programs, and personal teaching pages including student's assignments. Among the examined pages were personal pages created by technical and administrative staff, researchers, PhD students and other students.

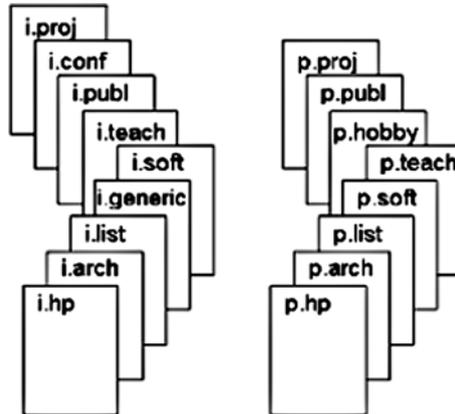
*Institutional* web pages are created for official, institutional and non-personal purposes, both academic and non-academic. The term *institutional* was used in the study as a generic term to cover web pages at any academic unit level in the investigated web space. Typical institutional web pages are thus homepages of schools, departments, centers, research groups, etc.; generic campus-wide service pages; institutional link lists pointing to research partners, institutions, research projects etc.; staff lists, institutional reports, newsletters and other publications; as well as conference and workshop pages.

The typology is not necessarily exhaustive for an academic web space. A larger sample of web pages would probably yield different and additional categories. Nor are the categories in the table mutually exclusive, but overlap with fluid borders. For example, institutional software documentation, manuals and tutorials were placed as *i.soft* and not as *i.publ* or *i.teach* because of the close relation to accompanying software programs.

In order to classify the pages in a consistent way, some rules of prioritized categorization order were used. Figure 12.4 shows the decided prioritized order among the 9 institutional and 8 personal genre classes with the highest prioritized genres in front. For example, if a personal homepage also included a link list as part of the page, this page was not classified as a *p.list* but as *p.hp*, because *p.hp* "overruled" *p.list* according to the prioritized categorization order shown in the figure.

As shown in Fig. 12.4, the prioritized order of institutional and personal web genres was identical apart from two specifically institutional genres (*i.generic* and *i.conf*) and one specifically personal genre (*p.hobby*), cf. Table 12.1.

The selection of the prioritized order was based both on "pre-coordinated" and "post-coordinated" opinions by the author. One pre-coordinated opinion was that institutional and personal homepages were the top hierarchical pages in the corresponding institutional or personal web territories and thus should be prioritized highest in the genre order. Another pre-coordinated opinion was that link lists, either institutional or personal, were of special interest in the study because of their possible cross-topic links. The remaining prioritized order emerged in a post-coordinated fashion when working with the page classifications and the definitions of the genres in Table 12.1. For example, the order of *i.soft* > *i.publ* > *i.proj* emerged when it turned out to be most relevant to collate all web pages related to a software in the same



**Fig. 12.4** Prioritized order of genre categorization of institutional and personal web pages (cf. Table 12.1). Highest priorities in front

category, because of the abovementioned close relation between software documentation and software programs.

Inevitably, higher prioritized genres received higher counts and lower prioritized genres got lower counts.

## 12.3 Results and Discussion

The findings in the study cannot be generalized but are indicative only due to the small, stratified sample of the 10 path nets. Instead, the results should be viewed as an exploratory identification and conceptualization of elements in relation to how web page genres may be interconnected in an academic web space. It would be interesting to investigate this kind of genre connectivity in a larger study of academic web spaces – and in other types of web spaces as well.

### 12.3.1 Source Genres, Target Genres and Genre Pairs

Table 12.2 shows the institutional and personal genre classes of the examined 281 source pages and 249 target pages in the 10 path nets.

As shown in Table 12.2, the institutional and personal genre classes made up about 50% each of the examined *source* pages in the 10 path nets. This result may indicate a relatively large importance of personal web pages for providing site outlinks, i.e., links to other sites, in an academic web space. Personal link creators like researchers and students can be thus important cohesion builders in academic web spaces.

About 60% of the *target* pages belonged to institutional genre classes, whereas 40% belonged to personal ones. This may indicate a higher authoritativeness

**Table 12.2** Institutional and personal genre classes of 281 source pages and 249 target pages

Genre class	No. of <i>source</i> pages	<i>n</i> = 281(%)	No. of <i>target</i> pages	<i>n</i> = 249(%)
i.archive	0	0.0	4	1.6
i.conf	26	9.3	12	4.8
i.generic	6	2.1	10	4.0
i.hp	0	0.0	42	16.9
i.list	73	26.0	23	9.2
i.proj	16	5.7	24	9.6
i.publ	8	2.8	7	2.8
i.soft	6	2.1	9	3.6
i.teach	4	1.4	18	7.2
<i>Sub total</i>	<i>139</i>	<i>49.5</i>	<i>149</i>	<i>59.8</i>
p.archive	0	0.0	2	0.8
p.hobby	2	0.7	12	4.8
p.hp	19	6.8	34	13.7
p.list	83	29.5	5	2.0
p.proj	0	0.0	4	1.6
p.publ	7	2.5	22	8.8
p.soft	14	5.0	14	5.6
p.teach	17	6.0	7	2.8
<i>Sub total</i>	<i>142</i>	<i>50.5</i>	<i>100</i>	<i>40.2</i>
	281	100.0	249	100.0

of institutional target pages compared to personal pages. However, the share of personal target pages may still indicate a relatively large importance by personal web pages for receiving site inlinks in an academic web space.

Four genre classes were not represented among the examined source pages. There were no institutional and personal archive pages, nor institutional homepages and personal project pages among the visited source pages in the path nets. A larger sample size would probably have yielded outlinks from these genre classes. However, all genre classes were represented among the visited target pages.

A horizontal reading of Table 12.2 shows some differences between genre classes that may be *site outlink-prone* or *site inlink-prone* as suggested by this small sample. For example, institutional homepages may receive more inlinks from other sites than they provide outlinks. This circumstance is comprehensible as the purpose of institutional homepages is to function as access points to web pages in the institution. Contrary to this, personal link lists are outlink-prone and less inlink-prone, again reflecting the purpose of the genre. However, more institutional link lists (9.2%) than personal link lists (2.0%) receive inlinks among the visited target pages, perhaps reflecting a larger authoritative quality of the institutional link lists.

Based on Table 12.2, Tables 12.3 and 12.4 list the most frequent source and target genre classes. The most frequent *source* genre classes were personal link lists (29.5%) and institutional link lists (26.0%). This result is not surprising as the purpose of link lists is to provide site outlinks to other web sites. The third most frequent source genre class was conference pages (9.3%) that typically had site outlinks to personal homepages of participators and to other conferences. The

**Table 12.3** Most frequent genre classes of source pages

Source genre classes	No. of <i>source</i> pages	Percent
p.list	83	29.5
i.list	73	26.0
i.conf	26	9.3
p.hp	19	6.8
p.teach	17	6.0
i.proj	16	5.7
p.soft	14	5.0
i.publ	8	2.8
p.publ	7	2.5
i.generic	6	2.1
i.soft	6	2.1
i.teach	4	1.4
p.hobby	2	0.7
i.archive	0	0.0
i.hp	0	0.0
p.archive	0	0.0
p.proj	0	0.0
	281	100.0

**Table 12.4** Most frequent genre classes of target pages

Target genre classes	No. of <i>target</i> pages	Percent
i.hp	42	16.9
p.hp	34	13.7
i.proj	24	9.6
i.list	23	9.2
p.publ	22	8.8
i.teach	18	7.2
p.soft	14	5.6
i.conf	12	4.8
p.hobby	12	4.8
i.generic	10	4.0
i.soft	9	3.6
i.publ	7	2.8
p.teach	7	2.8
p.list	5	2.0
i.archive	4	1.6
p.proj	4	1.6
p.archive	2	0.8
	249	100.0

most frequent *target* genre classes were institutional homepages (16.9%), personal homepages (13.7%) and institutional research project pages (9.6%).

Table 12.5 shows the distribution of target genre classes on institutional and personal source genre classes. As shown in the table, institutional homepages (*i.hp*) was the most frequent target genre class for four source genre classes, *i.list*, *i.proj*, *i.publ*, and *p.hp* in the sample. This could, for example, be a joint research project

**Table 12.5** Distribution of target genre classes on institutional and personal source genre classes

<i>Institutional</i> source genre class	Target genre class	No. of links	Sub total	<i>Personal</i> source genre class	Target genre class	No. of links	Sub total	
i.conf	p.hp	10	34	p.hobby	p.hp	3	3	
i.conf	i.conf	9		p.hp	i.hp	8	22	
i.conf	i.hp	4		p.hp	p.hp	4		
i.conf	i.proj	4		p.hp	i.publ	3		
i.conf	p.publ	3		p.hp	i.proj	2		
i.conf	i.generic	2		p.hp	i.teach	2		
i.conf	i.list	2		p.hp	p.soft	2		
i.generic	i.list	4	6	p.hp	i.soft	1		
i.generic	i.generic	2		p.list	p.publ	17	112	
i.list	i.hp	33	87	p.list	i.hp	15		
i.list	i.list	14		p.list	p.hobby	11		
i.list	i.teach	10		p.list	p.hp	10		
i.list	i.archive	7		p.list	i.list	8		
i.list	i.proj	5		p.list	i.proj	8		
i.list	i.publ	5		p.list	i.conf	7		
i.list	p.hp	5		p.list	i.generic	7		
i.list	p.hobby	2		p.list	i.teach	7		
i.list	p.proj	2		p.list	i.soft	4		
i.list	p.teach	2		p.list	p.list	4		
i.list	p.publ	1		p.list	i.archive	3		
i.list	p.soft	1		p.list	p.soft	3		
i.proj	i.hp	9		20	p.list	p.teach		3
i.proj	i.proj	3			p.list	i.publ		2
i.proj	i.conf	2			p.list	p.proj		2
i.proj	i.soft	2			p.list	p.archive		1
i.proj	p.hp	2			p.publ	p.hp		2
i.proj	i.list	1	p.publ		i.archive	1		
i.proj	i.publ	1	p.publ		i.list	1		
i.publ	i.hp	7	8	p.publ	p.hobby	1		
i.publ	i.proj	1		p.publ	p.publ	1		
i.soft	i.conf	2	6	p.publ	p.teach	1		
i.soft	p.soft	2		p.soft	p.hp	9	21	
i.soft	i.soft	1		p.soft	p.soft	5		
i.soft	p.hp	1	p.soft	i.hp	2			
i.teach	i.teach	4	4	p.soft	i.soft	2		
				p.soft	p.archive	2		
				p.soft	p.publ	1		
				p.teach	i.list	4	22	
				p.teach	p.soft	4		
				p.teach	p.teach	4		
				p.teach	i.hp	2		
				p.teach	i.teach	2		
				p.teach	i.proj	1		
				p.teach	p.hobby	1		
			p.teach	p.hp	1			
			p.teach	p.list	1			
			p.teach	p.proj	1			
			p.teach	p.publ	1			
			165				187	
							352	

page pointing to homepages of partner institutions, or a personal homepage pointing to the institutions of earlier jobs and studies.

Due to the small sample size, probably some pairs of genre classes are not represented in the table. For example, there are no links from *i.publ* to *i.publ*, or from *i.list* to *i.conf*. However, due to the design of the sampling, links are all between subsites located at different universities, thus excluding all links within the same university, for example, from one institutional publication to another publication at the same university.

Based on Table 12.5, Table 12.6 shows that there are outlinks from personal link lists to all 17 target genre classes, whereas institutional link list genre has 12 different *out-genres*, that is, target genre classes. It is a small sample, but these differences may indicate more diverse interests and purposes for creating personal link lists. The table also shows the percentage of outlinks targeted to institutional pages for each source genre. For example, 54.5% of the followed links in the data set from personal lists were targeted to institutional pages, whereas 85.1% of the followed links from the institutional lists had such targets. Again, this indicated difference is not surprising because of the different purposes for the two list genres.

Reversing the spectator's perspective, Table 12.7 shows the distribution of source genre classes on institutional and personal target genre classes.

Based on Table 12.7, Table 12.8 shows that personal homepages had the highest variety of inlinking source genres in this small study, having 10 different *in-genres*, whereas institutional homepages had 8 different *in-genres*. Not surprisingly, only 38.3% of the inlinks to personal homepages originated from institutional source pages, whereas 66.3% of the inlinks to institutional homepages had such origin.

As already indicated, the investigation revealed a rich diversity in interlinked genre pairs in the investigated academic web space. There were 83 different genre pairs connecting the examined web pages located at subsites at different

**Table 12.6** Source genre classes with outlinks to most different target genre classes

Source genre classes	No. of target genre classes	No. of outlinks	No. of outlinks to inst. target pages	<i>n</i> = 352 (%)
p.list	17	112	61	54.5
i.list	12	87	74	85.1
p.teach	11	22	9	40.9
i.conf	7	34	21	61.8
i.proj	7	20	18	90.0
p.hp	7	22	16	72.7
p.publ	6	7	2	28.6
p.soft	6	21	4	19.0
i.soft	4	6	3	50.0
i.generic	2	6	6	100.0
i.publ	2	8	8	100.0
i.teach	1	4	4	100.0
p.hobby	1	3	0	0.0
		352	226	64.2

**Table 12.7** Distribution of source genre classes on institutional and personal target genre classes

Source genre class	<i>Institutional</i>			Sub total	<i>Personal</i>			
	target genre class	No. of links			Source genre class	target genre class	No. of links	Sub total
i.list	i.archive	7		11	p.soft	p.archive	2	3
p.list	i.archive	3			p.list	p.archive	1	
p.publ	i.archive	1			p.list	p.hobby	11	15
i.conf	i.conf	9		20	i.list	p.hobby	2	
p.list	i.conf	7			p.publ	p.hobby	1	
i.proj	i.conf	2			p.teach	p.hobby	1	
i.soft	i.conf	2			i.conf	p.hp	10	47
p.list	i.generic	7		11	p.list	p.hp	10	
i.conf	i.generic	2			p.soft	p.hp	9	
i.generic	i.generic	2			i.list	p.hp	5	
i.list	i.hp	33		80	p.hp	p.hp	4	
p.list	i.hp	15			p.hobby	p.hp	3	
i.proj	i.hp	9			i.proj	p.hp	2	
p.hp	i.hp	8			p.publ	p.hp	2	
i.publ	i.hp	7			i.soft	p.hp	1	
i.conf	i.hp	4			p.teach	p.hp	1	
p.soft	i.hp	2			p.list	p.list	4	5
p.teach	i.hp	2			p.teach	p.list	1	
i.list	i.list	14		34	i.list	p.proj	2	5
p.list	i.list	8			p.list	p.proj	2	
i.generic	i.list	4			p.teach	p.proj	1	
p.teach	i.list	4			p.list	p.publ	17	24
i.conf	i.list	2			i.conf	p.publ	3	
i.proj	i.list	1			i.list	p.publ	1	
p.publ	i.list	1			p.publ	p.publ	1	
p.list	i.proj	8		24	p.soft	p.publ	1	
i.list	i.proj	5			p.teach	p.publ	1	
i.conf	i.proj	4			p.soft	p.soft	5	17
i.proj	i.proj	3			p.teach	p.soft	4	
p.hp	i.proj	2			p.list	p.soft	3	
i.publ	i.proj	1			i.soft	p.soft	2	
p.teach	i.proj	1			p.hp	p.soft	2	
i.list	i.publ	5		11	i.list	p.soft	1	
p.hp	i.publ	3			p.teach	p.teach	4	10
p.list	i.publ	2			p.list	p.teach	3	
i.proj	i.publ	1			i.list	p.teach	2	
p.list	i.soft	4		10	p.publ	p.teach	1	
i.proj	i.soft	2						
p.soft	i.soft	2						
i.soft	i.soft	1						
p.hp	i.soft	1						
i.list	i.teach	10		25				
p.list	i.teach	7						
i.teach	i.teach	4						
p.hp	i.teach	2						
p.teach	i.teach	2						
				226				126
								352

**Table 12.8** Target genre classes with inlinks from most different source genre classes

Target genre class	No. of source genre classes	No. of inlinks	No. of inlinks from inst. source pages	<i>n</i> = 352(%)
p.hp	10	47	18	38.3
i.hp	8	80	53	66.3
i.list	7	34	21	61.8
i.proj	7	24	13	54.2
p.publ	6	24	4	16.7
p.soft	6	17	3	17.6
i.teach	5	25	14	56.0
i.soft	5	10	3	30.0
i.conf	4	20	13	65.0
i.publ	4	11	6	54.5
p.teach	4	10	2	20.0
p.hobby	4	15	2	13.3
i.archive	3	11	7	63.6
i.generic	3	11	4	36.4
p.proj	3	5	2	40.0
p.list	2	5	0	0.0
p.archive	2	3	0	0.0
		352	165	46.9

universities. As shown in Table 12.9, based on Tables 12.5 and 12.7, the most frequent genre pairs were institutional link lists linking to institutional homepages (9.4%), personal link lists linking to personal publications (4.8%), personal link lists linking to institutional homepages (4.3%), and institutional link lists linking to other such lists (4.0%).

**Table 12.9** Most frequently interlinked genre pairs (cut-off < 7 links)

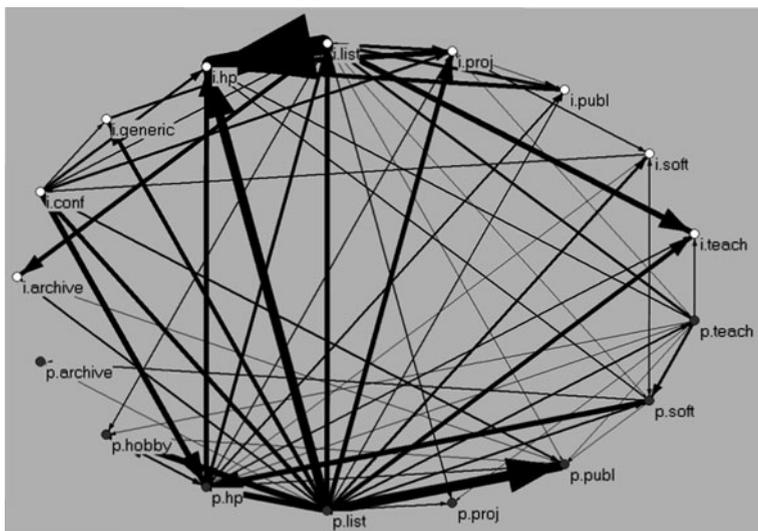
Source genre class	Target genre class	No. of links	<i>n</i> = 352(%)
i.list	i.hp	33	9.4
p.list	p.publ	17	4.8
p.list	i.hp	15	4.3
i.list	i.list	14	4.0
p.list	p.hobby	11	3.1
i.conf	p.hp	10	2.8
i.list	i.teach	10	2.8
p.list	p.hp	10	2.8
i.conf	i.conf	9	2.6
i.proj	i.hp	9	2.6
p.soft	p.hp	9	2.6
p.hp	i.hp	8	2.3
p.list	i.list	8	2.3
p.list	i.proj	8	2.3
i.list	i.archive	7	2.0
i.publ	i.hp	7	2.0
p.list	i.conf	7	2.0
p.list	i.generic	7	2.0
p.list	i.teach	7	2.0

### 12.3.2 Web of Genres

Converting data in Tables 12.5 and 12.7 to a genre adjacency matrix, the earlier mentioned network analysis software *Pajek* extracted the genre network graph in Fig. 12.5. The figure shows the genre pairs identified in the previous section revealing how links connect source and target genre classes at the investigated academic subsites. Dark nodes in the figure denote personal genre classes and white nodes denote institutional genre classes. Link width reflects frequencies of genre pairs (cf. Tables 12.5 and 12.7) and the figure thus clearly illustrates how institutional homepages (*i.hp*) and personal publications (*p.publ*) are inlink-prone genres, whereas personal link lists (*p.list*) and institutional link lists (*i.list*) are outlink-prone genres as also previously noted.

Figure 12.5 gives an impression of the diverse *genre connectivity* and some possible link paths between genres in an academic web space. The Web may thus be conceived as a *web of genres* with page genres linked to other genres. No other studies have been found discussing how web page genres are interconnected in large web spaces.

In large-scale academic web spaces, let alone the Web at large, there will inevitably be additional page genres as well as richer diversity of page genre combinations. The rich diversity of genre pairs may reflect a corresponding diversity of link motivations, including motivations related to teaching, research, leisure interests, social contacts, etc. A good overview of link motivations in an academic web space is given by [17].

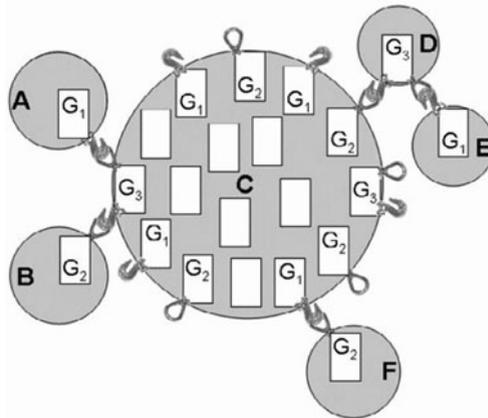


**Fig. 12.5** Genre connectivity in an academic web of genres with interlinked genre pairs among 17 genre classes (Table 12.1). *Dark nodes* denote personal genre classes and *white nodes* institutional genre classes. Link width reflects link frequencies (cf. Tables 12.5 and 12.7). *Pajek* network analysis software conceals thinner reciprocal links under thicker links. Genre selflinks are not shown

### 12.3.3 “Hook” Genres and “Lug” Genres

As shown in the previous sections, genres provide a diversity of link sources and link targets for each other. A metaphor of *hooks* and *lugs* can be used to conceive how genres “pull” web sites together across web spaces. In Fig. 12.6, hooks and lugs have been mounted symbolically on web pages. Pages with just hooks represent outlink-prone page genres ( $G_1$ ), like personal link lists providing many outlinks to other genres. Pages with just lugs represent inlink-prone page genres ( $G_2$ ) such as institutional homepages attracting many inlinks from other genres. Pages with both hooks and lugs represent out-/inlink-prone page genres ( $G_3$ ), like institutional link lists both providing and receiving many links across academic web domains. Naturally, real web pages can contain more outlinks and inlinks than represented by the single hooks and lugs included for sake of simplicity in the figure.

“Hook” genres and “lug” genres thus pull the Web together. Such genre connectivity affects web cohesion and reachability structures, that is, how web users and web crawlers may reach and retrieve web resources when following links from web page to web page – from genre to genre – as outlined in the introduction.

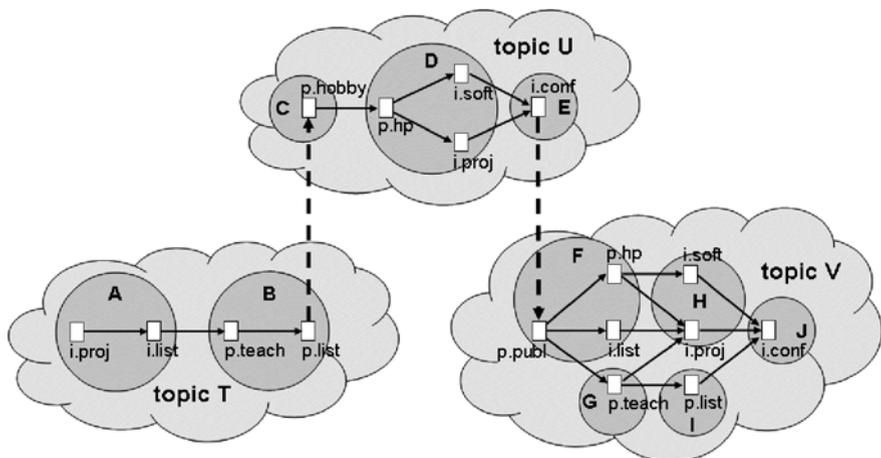


**Fig. 12.6** Some web page genres function as outlink-prone *hook* genres ( $G_1$ ), inlink-prone *lug* genres ( $G_2$ ), or combined *hook & lug* genres ( $G_3$ ) pulling web sites A–F together

### 12.3.4 Genre Drift, Topic Drift and Small-World Implications

As noted in the introduction, most links connect web pages and sites containing similar topics leading to topically clustered aggregations of interlinked web pages and sites on the Web. At the same time, some links connect different topical clusters.

As simplistically visualized in Fig. 12.7 with three topical clusters in different academic domains, web sites and topical clusters consist of a diversity of web page genres. Within a topical cluster it is possible to follow link paths from web page



**Fig. 12.7** Intra-cluster genre drift and inter-cluster topic drift along shortest link paths from web site A in topical cluster T to web site J in topical cluster V. Cross-topic links are denoted with *dashed bold links*

to web page – from genre to genre – thus creating a trail of interconnected pages belonging to different genres. Such a link path within a cluster may within just a few link steps connect a low-connected web page (with no outlinks pointing out of the web site) with a well-connected page with many site outlinks also to other topics. For example, in topical cluster T, a link path connects an institutional project page at web site A with an institutional link list at the same site linked to a personal teaching page and personal link list at web site B in the same topical cluster. The personal link list provides cross-topic links, one of which points to a personal hobby page at site C in topical cluster U.

In this context, the terms *genre drift* and *topic drift* are useful. *Genre drift* [3] deals with the change in page genres when following links from web page to web page, whereas *topic drift* (e.g., [2]) deals with the change in page topics when following links from web page to web page.

As shown in the figure and in the above example, genre drift within web sites and within a topical cluster containing many web sites may thus enable short link paths from a web page – with no cross-topic links – to a web page containing cross-topic links. Such a cross-topic link in turn creates topic drift by reaching out to other topical clusters hence causing short link distances between the clusters.

Topic drift negatively affects possibilities for topically focused web crawls [2]. However, in the context of the overall study outlined in this chapter, it is also important to understand how topic drift affect small-world properties in the shape of short distances across topical domains on the Web. The 10 path nets in the study all constituted deliberately induced topic drift in order to construct investigable “mini” small-world link structures [4].

As simplistically illustrated in Fig. 12.7, genre diversity within a topical cluster leads to genre drift along link paths whereas web page genres linking to a diversity

of topics lead to topic drift. Topical clusters with genre diversity thus entail genre drift along *intra*-cluster link paths. At the same time, some web page genres like institutional or personal link lists are more diversity-prone often containing cross-topic links. Such genres with topical diversity thus entail topic drift along *inter*-cluster link paths.

The combination of genre drift *within* topical clusters and topic drift *between* clusters may lead to short link distances across a web space, the small-world phenomena investigated in the study.

## 12.4 Conclusion

Using a network analytic approach, the chapter has presented an empirical investigation and exploratory identification and conceptualization of elements in relation to genre connectivity, that is, how web page genres may be interconnected, in this case in an academic web space.

The investigation of genre connectivity presented in the chapter formed part of a study with special focus on small-world properties in the shape of short link distances along link paths between web sites in an academic web space. Of course, network analytic approaches to genre connectivity and genre drift could also be based on data from non-academic web spaces and without focus on small-world aspects of these web spaces.

In the chapter, a pragmatic categorization into personal and institutional genre classes, i.e. bundled genre categories, was used in order to conceptualize, analyze and discuss genre connectivity in an academic web space. Other genre categorizations are possible. However, the main aim of the chapter has not been to identify what specific genres are connected but to develop an overall approach and framework to conceptualize how genres are connected in link structures on the Web. The chapter has showed how an academic web space comprises a *web of genres* containing a rich diversity of interconnected genres. Personal web page creators like researchers and students may be important cohesion builders in academic web spaces.

The data set in the study was collected in 2001. In a new study it would be interesting to investigate how emerging web 2.0 genres for collaboration and resource sharing, for instance blogs, wikis, social tagging and social networking, are integrated in academic web spaces and how they affect genre connectivity and the role of personal link creators.

Site outlink-prone “hook” genres like personal link lists were the most frequent source genre classes in the outlined study. Site inlink-prone “lug” genres like institutional homepages and personal homepages received most site inlinks. On this background, the chapter has discussed how “hook” genres and “lug” genres affect web cohesion and reachability structures on the Web by pulling web sites together.

Further, the chapter has discussed how topical clusters with genre diversity lead to genre drift, i.e. changes in page genres along link paths, and how web page genres prone of topical diversity lead to topic drift, i.e. changes in page topics along

link paths. Combinations of genre drift and topic drift may thus lead to short link distances, that is, small-world web spaces. In this context, the study may contribute to a better understanding of how dynamic link structures of the Web are threading trails across web genres and topics for web users and web crawlers to reach and discover.

As genre connectivity and genre drift affect cohesion and reachability structures on the Web it would be interesting in a large-scale study to investigate possible patterns of genre connectivity and genre drift in academic as well as non-academic web spaces incorporating some of the methodologies and the conceptual framework developed in the present study.

## References

1. Agatucci, C. 2000. Cyber rhetoric (3): web genres & purposes. Available in Internet Archive: <http://web.archive.org/web/20000920061159/http://www.cocc.edu/hum299/lessons/rhet3.html>
2. Bharat, K., and M.R. Henzinger. 1998. Improved algorithms for topic distillation in a hyper-linked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 104–111. ACM Press.
3. Björneborn, L. 2004. Small-world link structures across an academic web space: A library and information science approach. PhD thesis, Royal School of Library and Information Science, Copenhagen, Denmark. <http://www.iva.dk/lb>
4. Björneborn, L. 2006. “Mini small worlds” of shortest link paths crossing domain boundaries in an academic Web space. *Scientometrics* 68(3):395–414.
5. Björneborn, L., and P. Ingwersen. 2004. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology* 55(14):1216–1227.
6. Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, J. 2000. Graph structure in the Web. *Computer Networks* 33(1–6):309–320.
7. Crowston, K., and M. Williams. 2000. Reproduced and emergent genres of communication on the World Wide Web. *The Information Society* 16:201–215.
8. Davison, B.D. 2000. Topical locality in the Web. In *Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 272–279. ACM Press.
9. Haas, S.W., and E.S. Grams. 1998. Page and link classifications: connecting diverse resources. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, 99–107.
10. Haas, S.W., and E.S. Grams. 2000. Readers, authors, and page structure: a discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science* 51(2):181–192.
11. Koehler, W. 1999. Classifying Web sites and Web pages: The use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science* 31(1):297–307.
12. Milgram, S. 1967. The small-world problem. *Psychology Today* 1(1):60–67.
13. Newman, M.E.J. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.
14. Rehm, G. 2002. Towards automatic web genre identification: A corpus-based approach in the domain of academia by example of the academic’s personal homepage. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, IEEE Computer Society, vol. 4, 101–110.
15. Steyvers, M., and J.B. Tenenbaum. 2005. The large-scale structure of semantic networks: statistical analyses and a model for semantic growth. *Cognitive Science* 29(1):41–78.
16. Thelwall, M. 2002/2003. A free database of university web links: data collection issues. *Cybermetrics* 6/7(1): paper 2. <http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html>

17. Thelwall, M. 2003. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research* 8(3): paper 151. <http://informationr.net/ir/8-3/paper151.html>
18. Thelwall, M., and G. Harries. 2003. The connection between the research of a university and counts of links to its web pages: an investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and Technology* 54(7):594–602.
19. Watts, D.J., and S.H. Strogatz. 1998. Collective dynamics of “small-world” networks. *Nature* 393(June 4): 440–442.
20. Weare, C., and W.-Y. Lin. 2000. Content analysis of the World Wide Web: opportunities and challenges. *Social Science Computer Review* 18(3):272–292.



**Part V**  
**Case Studies of Web Genres**

# Chapter 13

## Genre Emergence in Amateur Flash

John C. Paolillo, Jonathan Warren, and Breanne Kunz

### 13.1 Genres, Multimedia and the Web

While the core functionality of the World-Wide Web is the exchange of textual communication in the form of formatted text documents over the Internet, this functionality predated the advent of the web, and arguably other features, specifically graphics, were responsible for the attractiveness and rapid adoption of the web and its technologies. Today, one rarely sees a news story on a newspaper web site without embedded graphics of some kind, generally in the form of advertising using Flash multimedia. Multimedia and graphics on the web are notorious for their technological instability, with multiple proprietary and open formats in competition, requiring browser add-ins and software updates for sites to remain usable or even legible. Similarly, the forms of multimedia and graphics are constantly in flux, as old forms are refined and new forms arise on a continuing basis. Advertising makes extensive use of animated formats, especially Flash, although video, sometimes streamed and sometimes delivered in Flash or other formats, is becoming more prevalent. Video weblogs and user-contributed video sites such as YouTube have extended the opportunities available for users to create and manipulate multimedia forms.

For these reasons, a characterization of web genres needs to address multimedia in some form. At the same time, the volatility of multimedia forms draws attention to issues of genre evolution, development, or emergence, which also characterize communication forms on the web more generally, including text. This chapter aims to contribute to understanding multimedia genres on the web. An attempt to comprehensively characterize web-based multimedia genres, however, would presuppose some established or well-defined contextual boundaries, comparable to those employed by Biber [2, 3], with adequate procedures for sampling multimedia types and observing features of interest. While we have some hope that this can eventually be done, our view is that such an attempt at this time would likely be premature.

---

J.C. Paolillo (✉)

School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA  
e-mail: paolillo@indiana.edu

Consequently, we instead focus on the mechanisms of genre emergence and the social processes involved. Specifically, we examine the emergence of multimedia genres through processes of social positioning within a particular context, the amateur Flash exchange site [www.newgrounds.com](http://www.newgrounds.com) (henceforth, Newgrounds). We find that social positioning influences both the conventional forms and meanings of the messages communicated on the site. Furthermore, the social processes involved are similar to those reported for other computer-mediated communication modes, such as Usenet newsgroups [15, 17, 18], Web-based discussion boards [22], Internet Relay Chat [13], Listservs [1], and email [24], among others. Hence, our observations point to a potential source of genre emergence shared across many types of Internet communication.

The emergence of new genres of communication is a recurrent interest in research on digital media. The near-constant state of change in digital media technologies requires users to adapt their communication patterns as the newer technologies are adopted. These may eventually stabilize into socially recognized categories, reflecting expected kinds of content form and purposes of communication. We follow Hymes [14] in identifying stable, culturally recognized categories of communication as genres. Similar definitions are invoked in research on explicitly digital genres. For example, Erickson [11], working to synthesize definitions of genre taken from research on digital media [6, 10, 26–28], defines genres as follows:

A genre is a patterning of communication created by a combination of the individual, social and technical forces implicit in a recurring communicative situation. A genre structures communication by creating shared expectations about the form and content of the interaction, thus easing the burden of production and interpretation. [11]

Hymes' ethnographic approach foregrounds the notion of the community ("speech community") as the locus of genre conventions; again, similar notions surface in research on digital genres, such as "discourse community" from the field of rhetoric [6, 7, 9, 11], "organization or professional community" [25], and "community of practice" [11]. These notions are sufficiently congruent as to be treated equivalently, though we take Hymes' notion to be more inclusive.

Studies of digital genres have been largely taxonomic in orientation. However, a major reason for referencing the communities in which genres are used is to frame the processes and mechanisms of genre emergence. While the introduction and development of new technologies is important in these processes, particularly in the case of digital genres, it is widely recognized that social processes are critical in shaping the resulting genres. In part, these processes concern whatever messages need to be communicated among the community members; hence, the community, however defined, determines which people are involved in processes of genre emergence and what the likely communicative needs are, as well as what existing genres might be reproduced in the digital environment [6, 11, 26]. At the same time, there may also be social dynamics particular to a community that shape its communicative needs and the genres that may emerge from it.

While recognition of genre emergence places theoretical emphasis on social process, studies of genre emergence tend not to directly address its mechanisms.

Partly this has to do with methodological difficulties: genre emergence is a situated, organic process that unfolds in real-life communication. Such circumstances are difficult to explore experimentally, and it is hard to establish the causal chains implied by the intent to identify mechanisms. Consequently, existing accounts of genre emergence tend to adopt ethnographic and historical approaches. Yet digital media offer opportunities to study genre emergence in a new way. First, digital media tend to provide numerous instances of communication, which are readily stored and archived, and can be examined in detail, post-hoc. Hence, we can create large corpora of exemplars from which to draw inferences about what genres exist, what communicative purposes they serve, and what their typical characteristics are. Empirical methods used for the quantitative analysis of genres (as evidenced by other chapters in this volume) are facilitated by the documents' existence in digital form. More importantly, from the perspective of emergence, is that the systems in which the digital documents reside often carries other traces of social interaction in the form of metadata and user profiles; these can be analyzed using similar methods to reveal patterns of social process that are potentially relevant to genre emergence.

Our approach to the genre analysis of amateur Flash multimedia draws from that of Biber [2, 3], in which a feature matrix for a sample of texts is analyzed using a latent structure model. In this approach, the texts are sampled from traditional corpora such as the London-Lund and Lancaster-Oslo-Bergen corpora, supplemented with additional texts for otherwise un-represented text types. The features are various linguistic variables, lexical classes, syntactic constructions, etc. that are counted in each of the texts, and the latent structure model is a Common Factor Analysis model with non-orthogonal rotation of the factors. Shared variation among the texts is then interpreted in terms of genre (or "register") conventions pertaining to the texts.<sup>1</sup> Our approach has analogs for each of these characteristics: representative samples of Flash animations from a corpus, a feature matrix counting structural features in each animation, a latent structure model for analysis and interpretation in terms of genre. The manifestations of these characteristics differ in certain specific details necessary to the material being studied, but the overall architecture is the same.

To the genre analysis we add a social network analysis based on user profiles. Through social network analysis, we can identify the organic structure among the participants in a community [8, 25], in this case, the amateur Flash multimedia authoring community of Newgrounds. This analysis is expressed in terms of social positions and relations among them, which are characteristically plotted in a reduced social network diagram. Centrality, prestige and power are easily read from such a plot, and the outcomes of other social processes, e.g. competition, can sometimes be read as well.

---

<sup>1</sup> Biber uses the notion "text type" as a methodological intermediary in his approach, where the final interpretation is in terms of "register" or "genre" (which he treats as equivalent [2, 3], but cf. [12, 20] for a distinct sense of "register"). See also Chapter 14 by Grieve et al. (this volume) for these terms, as well as Biber et al. [5]; Biber and Kurjian [4]. There is also a distinct use of "text type" elsewhere [19].

In this chapter, we propose that a more compelling account of genre emergence can be constructed by coupling the social network analysis with the genre analysis, and cross-correlating them. In this way, social groups are identified and associated with independently identified emergent genres of Flash content. The associations observed, when taken together with our combined several hundred hours worth of observing communication on the site, help illuminate the processes by which the emergent genres arose. Hence, social network analysis, we argue, can contribute important understandings to the processes of genre emergence in digital media.

The organization of our chapter is as follows. In the next section, we orient the reader to our research site, Newgrounds, highlighting characteristics salient to its users that play a role in genre emergence on the site. We then outline our data and analytical methods, of which there are three major components: identifying social relations on Newgrounds, identifying candidate emergent genres, and longitudinally correlating the relationship between genre and social relations. We then discuss the results, and conclude with wider implications for the study of genres on the web and genre emergence more generally.

## 13.2 Flash and Newgrounds in Amateur Multimedia

Multimedia on the Internet takes many forms, but by any measure Flash is a very important one. Unlike alternative formats such as SVG, Flash enjoys broad support, with plugins or native support in the majority of web browsers. Flash is also widely deployed in web-based advertising, and a some types of websites (e.g. many official fan websites for celebrities, models and popular music artists) make extensive use of Flash for content that is otherwise easily delivered as HTML.

Flash has a long history of use in amateur Internet animations. While originally developed as a program for handwriting recognition, its authors quickly re-purposed its vector-based graphics rendering engine for delivering animated images on the web, once it became clear that it could provide resolution and bandwidth advantages over rasterized video. As a nominally open yet proprietary format, Flash was primarily developed for authoring tools marketed by Macromedia (now part of Adobe), although third-party authoring tools, such as FlashMaker and the open-source Ming library also exist. It has been adopted as a development platform by many higher-education programs in animation and multimedia. Numerous websites specialize in collecting and archiving completed animations, as well as works in progress, artwork, ActionScript programming, and other source material needed to make Flash animations. For this set of reasons, Flash has a broad and international following among amateur animators.

Newgrounds has a central position in the Internet ecology of Flash. Originally a personal website of Internet game author Tom Fulp, Newgrounds has evolved into a major hosting service for amateur Flash. In April 2000, under Fulp's entrepreneurship, Newgrounds added a "flash portal" where users could upload their own creations to be hosted on the site. Pre-figuring many of the social networking and discussion features of YouTube and other media sites, the portal provided a

rating system for judging submissions, a mechanism for communicating reviews and critiques to authors, discussion fora, and author profiles with social networking features (“favorite flash”, and “favorite authors”). Since the Flash portal opened, it has attracted over one million distinct users and over 300,000 flash animations. Most users are teenage or young adult males living in English-speaking countries, but Flash authors from a wide range of international backgrounds exhibit their work on the site.

Informal observation suggests that Newgrounds is a site of genre emergence. Several recognized forms of Flash originated on Newgrounds and are still mainly practiced there. One such type is a narrative constructed around characters from video games, in particular, older console games such as the Super Mario Brothers, Sonic the Hedgehog, Megaman, Final Fantasy, The Legend of Zelda and others. Stick figure animations are also common, and like the video game animations, they tend to be focused on elaborately choreographed fight scenes. Other types are more subtle to identify.

One such example is “animutation” a form based largely on animation of bitmaps synchronized to music (especially Japanese popular music), thematically exploring paradoxes around masculine “geek” identity [16]. Similarly, there are “clock movies”, featuring avatars made with inanimate objects that have clocks or other objects for faces). Both animutations and clock movies make heavy use of inter-textual references, especially to other animations on Newgrounds; one needs quite a bit of background knowledge about the authors and their animations to fully appreciate them.

A number of popular Internet animations originated on Newgrounds, and these often have a canonical status, being widely emulated or parodied. Curators on the site often arrange them into “collections”, semi-official listings of related animations. Examples are “All Your Base are Belong to Us”, a fast-paced montage of still frames which circulated during an Internet craze of the same name, and the “Numa Numa Dance”, in which a 19-year old Gary Brolsma from New Jersey lip-syncs to a Romanian pop song. These different types of animations are potential sites of genre formation for amateur multimedia.

The cultivation of distinctively Newgrounds-based styles such as these, and the attendant competition for viewers’ attention on the site, also leads to the expression of specific kinds of messages within the movies themselves. Typically, these involve an author’s social alignment with or against some other author or authors. One type of message we see quite frequently is one we call “ownage”, in which an animator appropriates another animator’s character to show it in a compromised (often mutilated) condition. Often times, this is accompanied by text declaring that the author “owns” (or “pwns”), in video gaming parlance, the other animator. The intent of this message is to assert a superior position over another (usually popular) animator.

Hence, Newgrounds is a site central to the distribution, critique and support of amateur Flash animation. Flash hosted on the site exhibits formal structures that are potentially distinctive to Newgrounds Flash content, and its content involves at least some distinctive kinds of messages, expressing social positioning of animators on

the site. It is thus a site of potential genre emergence. We turn now to the methods that we use to investigate the emergence of Flash genres on Newgrounds.

### 13.3 Method

Our method has two main parts: a social network analysis of user profiles from Newgrounds, and an analysis of selected movies for genre features and cultural references. The primary methods in both analyses are quantitative, relying on Principal Components Analysis and hierarchical clustering to identify relevant groupings. For the social network analysis, we obtained a sample of user profiles by randomly sampling 10,000 initial profiles and iteratively crawling social network links from those. These data were then arranged in a user-by-user sociomatrix, to which we applied a minimum threshold, ultimately keeping actors with at least eight “favorite author” nominations. This sociomatrix was then reduced to a set of structurally equivalent social positions using Principal Components Analysis followed by Hierarchical Cluster Analysis, which we then interpreted by inspecting the cluster members and viewing the corresponding profiles and movies. We then aggregated the sociomatrix according to the clusters and plotted the reduced social network diagram to get a sense of the structural relations among the different social positions.

For the genre analysis, we constructed a stratified random sample of movies, sampling movies from each of the social positions at equal probability levels. In this way, we ensured that movies produced by each group of authors would be represented comparably. We then viewed these movies and classified them according to a set of variables hypothesized to be related to the potential genres of Flash movies. In addition, we coded the same random sample for cultural reference variables. Both sets of variables were constructed by the research team using an iterative (hermeneutic) grounded theory approach [23]. Genre features and cultural references were arranged into document-by-feature and document-by-keyword matrices, which we again analyzed by Principal Components and Hierarchical Cluster Analysis. These three analyses were then cross-correlated and compared with the dates that the movies were first posted, so that an interpretation of genre emergence over time could be developed.

#### 13.3.1 Sampling

The sample data were collected from two types of pages on the Newgrounds site: movie description pages<sup>2</sup> and user profile pages.<sup>3</sup> Both types of pages are created by the site’s users and are open to the public web. Movie description pages contain several important pieces of information. An “author” field lists names and links

---

<sup>2</sup> For example <http://newgrounds.com/portal/view.php?id=206373>

<sup>3</sup> For example <http://newgrounds.com/gold/profile/template.php?id=318335>

to the profiles of as many as five authors, along with a submission date and time. Beneath the author information, a six-level rating scale (0–5) permits site visitors to rate the submission; these ratings are aggregated to provide an indication of a submission’s overall popularity. If the aggregate rating (a weighted average) falls below 1, the submission is removed from the site, or “blammed”. Several more pieces of information are specific to the submission itself, including a link to the actual Flash file, a space for author comments, the most recent review, and a link where registered users provide their own review.

Users’ profile pages provide some useful personal and social information, including an image identifying the user (possibly an avatar image), age, gender, location, occupation, and a personal message, some of which are optional. In addition, user profiles provide four menus providing links to Flash files that they have authored, audio files they have authored, their favorite Flash artists and their favorite Flash content hosted on the site. User activity on the site is reported in aggregate in terms of “levels” and “ranks”, based on the amount of reviewing, they have done and their voting patterns. Each user profile also has an integer serial number, making random sampling straightforward.

Using a seed of 10,000 random integers between 1 and 856,613 (the user count as of November 11, 2005), the corresponding user profiles were collected. Of the initial 10,000 pages, 1,115 contained menus with the necessary social network information. This information was extracted using DOM parsing and each link found was recorded in a database table with the following four columns:

- (i) the crawling iteration number
- (ii) the profile number containing the link
- (iii) the destination profile or movie number linked to, and
- (iv) the type of link (i.e. which of the four drop-down menus)

In each iteration, new profile pages (identified through the “Favorite Flash Authors” from the previous iteration) were collected and parsed, resulting in a “snowball” sample. The process was repeated until no more new profiles were found. When the favorite author sampling no longer returned results, the “Favorite Flash Content” relationships were used to find Flash movie description pages. Those movie pages were then visited and the authors from their collaborator lists were extracted for use as the seeds of subsequent snowball samples. In all, over 38 sampling iterations, 17,479 pages were visited for this study. Out of the more than 900,000 profile pages available at the time of this sample, 8,314 were visited and 158,723 unique author-author relationships were identified. This suggests a fairly densely interlinked set of core Newgrounds users.

### ***13.3.2 Identifying Potential Emergent Genres***

The genre feature analysis was conducted in two phases. In the first phase, members of the research team independently viewed a common sample of 50 Flash files that

had been randomly selected from the corpus of Flash authored by any of the users in our sample. Each team member made notes which were shared and discussed in joint meetings when the Flash was viewed again. From the discussion, the following 67 potential genre features were identified for further examination.<sup>4</sup> These genre features fall into six major structural categories:

1. Production elements:
  - a. Preloader: Newgrounds, Flashportal, Games of Gondor, Armor Games, or other branded preloader
  - b. Credits: opening, closing, during play
2. Authorship: single author, animation + sound collaboration, other collaborations
3. Narrative:
  - a. Main point: action, fight, dance, drama, collage, participatory narrative, game
  - b. Characters: video game, anime, celebrities, avatars, other characters
  - c. Narrative advancement: by scenes, camera effects, special effects
  - d. Pacing: slow, dense
4. Technique:
  - a. Graphic composition: vector animations, bitmaps, 3D animations, video, stop-action, slideshow
  - b. Artistic technique: carefully drawn, rapidly drawn,
  - c. Animation technique: vector-based transformations, frame-by-frame
  - d. Backgrounds: fill, gradient, bitmap pictures
  - e. Camera effects: zoom, pan, tilt/rotate
  - f. Characters: stick figures, clocks, locks, glocks, stars, other abstract
5. Audio:
  - a. Music: J-pop, movie, rock/metal, pop, classical, rap/hip-hop/R&B, video game music, other musical genres
  - b. Voices: computerized voices (Speakonia), voice acting, subtitles
  - c. Misc. audio: sound effects, subtitles, distortion
6. Interactivity: buttons, scene menus, keyboard/mouse actions, play/pause controls, subtitles on/off.

The original sample of 50 files was recoded according to the 67 features collectively by the entire research team, with any variant codings discussed in subsequent meetings until there was full agreement on the coding definitions and the codes for the first 50 files. An additional 850 files were then identified as a second random sample. These were split into three groups and analyzed by the individual members of the research team, resulting in a combined set of 900 Flash files. Of these,

---

<sup>4</sup> The resemblance to Biber [2] in the number of genre features is entirely accidental.

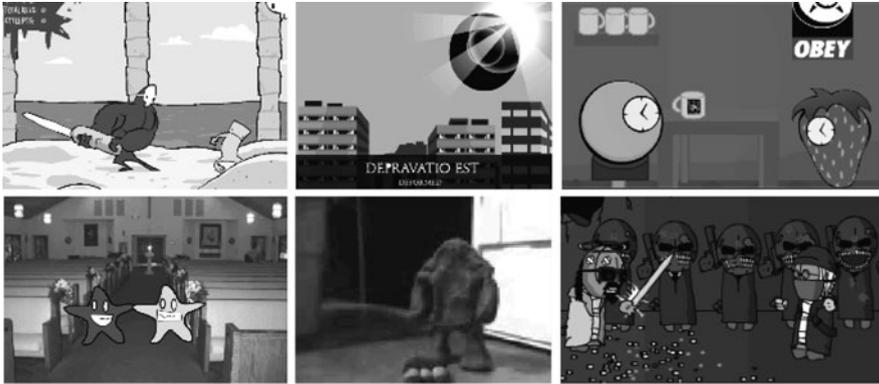
29 had been blammed, deleted or were otherwise unavailable for analysis, leaving 871 Flash files in our sample. Each of the 67 genre features was coded as a binary attribute, present or absent. Some of the features coded are in entailment relations, e.g. use of any of the camera techniques (zoom, pan, rotate/tilt) implies that camera effects are used for narrative advancement, and single authorship implies that a submission is not a collaboration. Other superficially similar features may not be in an entailment relation, e.g. a submission may use bitmap backgrounds without needing to use a bitmap composition technique such as editing in Photoshop.

Some of the features selected are specific to Newgrounds Flash, and hence require further explanation. For example, many submissions use abstract characters of different sorts, by which we mean participants in the narrative whose features (arms, legs, faces, etc.) are abstracted in some way. The most obvious of these is the stick figure, but Newgrounds exhibits many other types. One abstract character type is found in the “Madness” series, where a character’s head and torso are simple geometric shapes, the face is simply a crossed pair of lines (representing nose and eyes), and hands and feet are not attached to the character’s body by visible limbs. Another salient type of abstract character is the “Clock” character: canonically a fruit with a clock face in place of facial features, animated to give the impression of facial expressions. Clock characters typically function as avatars: the animators who post them tend to use nicknames with “clock” (Orange Clock, Pineapple Clock, Raspberry Clock, etc.) and typically identify with the “Clockcrew”, an informal association of animators. Other authors using abstract characters follow the clock style, replacing the eyes, nose and mouth with other objects such as a keyhole (the “Lock Legion”) or pistol (the “Glock Group”). These latter types of abstract characters also tend to represent avatars of their animators.

Another important set of features concerns the means by which the submissions are composed. While all of the files on Newgrounds are delivered as Flash (.swf) files, a variety of tools are used to compose them. Most are composed using Flash (and Flash tutorials are a common type of submission), although it is common to use bitmap graphics, or a combination of vector graphics and bitmaps. A small proportion of Newgrounds submissions are composed as slideshows (timed sequences of still shots), videos (imported from another application) or stop-action movies (usually with clay animation, legos or action figures) (Fig. 13.1).

### ***13.3.3 Cultural References and Message Content***

Genres are only partially defined according to their formal features. Another important component is in the purpose or communicative intent of the communicative event. In order to identify categories of message content, cultural references are defined as any normative social or cultural expression invoked in an animation, independent of any stylistic or technical aspects of that expression. Generally, cultural references are taken here to be socially semantic or ideational expressions (people,



**Fig. 13.1** 1 A set of frames from various Flash files on Newgrounds. *Top row*: characters in a game by Tom Fulp and Dan Paladin, an animation employing sophisticated Flash vector animation techniques (gradients, blur, panning, etc.) and a Clockcrew movie with clock avatars. *Bottom row*: Star Syndicate avatars on a bitmap background, a claymation by Knox, and a scene from a movie in the Madness series by Krinkels

things, and ideas), in contrast to the forms and techniques used as genre features above. We did not attempt to interpret personal, non-social meanings of individual authors.

Cultural reference features were identified inductively from the animations in much the same way as were the 67 genre features. Fifty movies from the sample of 890 were viewed by the entire research team, and all cultural references that any of the team members could identify were given labels for use in coding the rest of the sample. Unlike with the genre features, the remaining movies were coded by a single member of the research team (the second author), who accumulated and employed a controlled vocabulary of 3,016 distinct cultural references as he worked. Although the coder is not an author or reviewer on Newgrounds, he has a similar demographic background to a typical older Newgrounds author (28 year-old US male).<sup>5</sup> Signs and symbols that were unfamiliar, and which may have been cultural references, were searched-for on the Newgrounds website and on the Web as a whole until the sign had been adequately understood. At the end, a codebook (available upon request) containing code-to-meaning correspondences was manually constructed from the complete vocabulary of codes. Intracoder reliability was confirmed informally by revisiting a subset of 20 of the movies 3 months later and, using the codebook, confirming that the same codes still seemed appropriate.

As with the six categories of genre features, one can identify categories of cultural reference features. The following six categories summarize cultural reference codes that were each applied to more than 10 movies:

<sup>5</sup> The age and gender distributions of Newgrounds users are known from forthcoming studies not reported here.

1. Sex and violence (most common codes; each line is from most to least frequent)
  - pistol, fire, machinegun, sword, knife, katana, ninja, decapitation, xeyes, money, cigarette, zombie, redeyes, robot, pirate, shotgun, dismemberment, aliens, bomb, nuclear, marijuana, suicide, skull, beer, axe, rifle
  - sex (depicted/suggested), penis, boobs, trickythec clown, masturbation, prostitute
  - gay (as insult), homosexuality (depicted/suggested), fag (as insult)
2. Newgrounds author groups
  - SBC, speakonia, clockcrew, B, orangeClock, pineappleClock, anticlockclock, raspberryclock, king, pwn, bananaClock, truffleclock, pepsiclock
  - starsyndicate, dailytoon, tehedn, paly
  - locks
  - sticks
3. Mass media and video games
  - matrix, starwars, lotr, startrek
  - tmnt, simpsons, powerrangers, pokemon, dbz, transformers,
  - game, mario, finalfantasy, legendofzelda, sonic, nintendo, megaman, xbox
  - tv, tvstatic
  - mswindows, mcdonalds
4. Celebrities and famous people (including prominent Newgrounds personalities)
  - gwbush, michaeljackson, hitler, binladen, americanflag, nazi
  - devil, jesus, christmas, god, crucifix, santaclause, christianity
  - legendaryfrog, wadefulp, tomorrowsnobody, foamy, superflashbros, piconjo, knox, perfectkirby, tomfulp
5. Nature
  - earth, moon, outspace, starrynight, sun, spaceship, rain, flowers
6. Slang
  - omg, lol, :(, wtf, <3, :, blam, rotfl

These categories go some distance towards defining the communicative intent of these Flash movies. The first category describes types of weapons (pistol, fire, machine guns), characters (ninjas, zombies, robots), and actions (decapitation, suicide, dismemberment) that often appear in violent sequences of the movies, which are extremely common. Also common are themes of sexual function, sexual violence, and homophobia as well as the use of gateway drugs (cigarettes, marijuana, beer). A second category includes references to the “crews” mentioned earlier (Clock Crew, Star Syndicate, Lock Legion). Much like street gangs in real life, each crew has its own intricate system of signs and authority figures, which its members use to distinguish themselves and to mark their (in this case, artistic) territory. For example, the Clock Crew’s founder, Strawberry Clock, often declares himself “King

of the Portal”, which is a designation given to authors of high-ranking movies by the site’s proprietors. Category three describes mass media influences invoked in the movies, including large science fiction and action films (Matrix, Star Wars, Lord of the Rings), television cartoons and children’s shows (Simpsons, Teenage Mutant Ninja Turtles, Pokemon), and popular console video games (Mario, Final Fantasy, Legend of Zelda). In the fourth category, several types of famous people emerge, namely politicians and pop culture figures who are often ridiculed in the animations (GW Bush, Michael Jackson, Bin Laden, Hitler), figures associated with Christianity (devil, Jesus, God), and popular authors on Newgrounds (Legendary Frog, Tomorrow’s Nobody, Super Flash Brothers). The fifth category shows how the natural world is typically depicted in the movies. Science fiction and action movies often have space, Earth, or the moon as their background. Movies set outside on Earth often include the sun, rain, and flowers. Finally, a number of popular Internet slang terms have found their way into Newgrounds movies, having their usual meanings, including omg, lol, :, :(, wtf, <3 (a heart), and rotfl. The local term “blam” also appears often, because users fear having their movies rejected, and use the word to threaten others with public rejection. The visual nature of all six types of cultural references makes their communicative intents particularly salient in the animations, perhaps even more salient than if they were to appear in textual modes of group computer-mediated communication.

## 13.4 Results

Our results are presented in four parts. In Section 13.4.1, we describe the social network analysis followed by the genre analysis in Section 13.4.2, and the analysis of cultural references in Section 13.4.3. Finally, in Section 13.4.4, we describe the mapping across these three analyses, using a log-linear modeling framework, to answer our questions about the relation of genre emergence to social structure and process.

### 13.4.1 Network Analysis

Our network analysis was conducted in two steps. We first constructed a sociomatrix from the favorite flash author information in the user profiles, considering only those authors identified eight or more times as “favorite”. This was reduced to a set of seven social positions, using Principal Components Analysis and Hierarchical Cluster Analysis (using Euclidean distance and Ward’s clustering method). This method identifies socially equivalent actors, in terms of their ties to other actors [8, 25]. A dendrogram for this cluster analysis is given in Fig. 13.2.

The cut taken of the dendrogram, leaving seven clusters was arbitrary, and was intended to break apart some of the larger groups in case interesting structure within them might be missed at a higher level of cut. Inspection of the principal components plots for these clusters showed that each was separated along a distinct principal

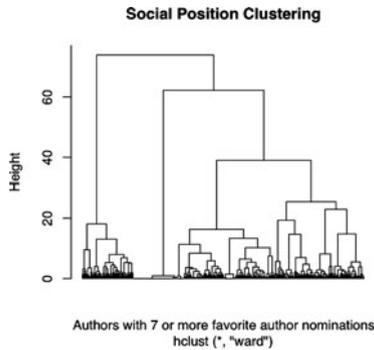
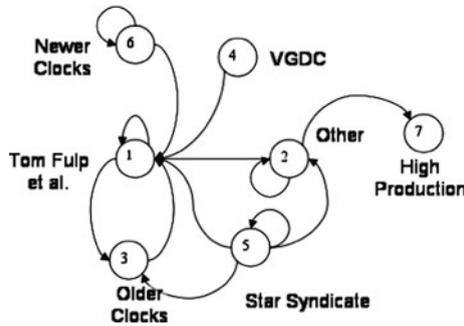


Fig. 13.2 Cluster dendrogram of social positions by favorite author nominations

component dimension, hence, the separation of these clusters is also justified on empirical grounds.

Moreover, membership in the different clusters is readily interpretable, at least from the perspective of the experienced user of Newgrounds. One cluster includes both Newgrounds entrepreneur Tom Fulp and his principal artistic collaborator, Dan Paladin. This group of authors represents the “mainstream” of Newgrounds Flash, and we label it “Tom Fulp et al.”, on account of his naturally central position on the site. A second cluster, resolved as distinct from the first cluster on dimension 3, is a group of highly popular authors including Legendary Frog and the Super Flash Brothers (called here “High Production”). These authors concentrate on creating movies with high production values, generally based around video game themes (though not exclusively). Three more clusters consist of creators of avatar movies, principally older and newer members of the Clock Crew (“Older Clocks” and “Newer Clocks”, respectively) and members of the Star Syndicate, whose principal product has a series of daily Flash collaborations called “daily toons”. The last two clusters are a group of authors making movies based on video game themes (“VGDC”, for “Video Game Digital Character”), and a residual group of authors with no clearly characterizable style (“Other”).

We re-counted the favorite author links as ties among the seven social positions identified, and visualized them in a reduced sociogram, presented in Fig. 13.2, from which it is clear that Fulp’s group and the Older Clocks are the most central participants on Newgrounds. These two groups simultaneously lend each other social capital and receive ties from members of many of the other groups, especially Other, which represents the undifferentiated residual group. Four groups show significant self-ties, meaning that their members provide support to other members of the same group. Among these only Fulp’s group is central; the other three show no significant ties of support from other groups. Hence, the popularity of these three groups, and the success of their Flash content on Newgrounds, depends largely on their internal social cohesion. The High Production group differs from the others in its peripherality and lack of internal cohesion.



**Fig. 13.3** Reduced sociogram indicating ties among the seven social positions of Flash authors according to favorite Flash author nominations

The sociogram in Fig. 13.3 reflects a state of competition among the different groups, which we noticed in observing movie reviews and references made in the movies themselves. In particular, the Star Syndicate and the Newer Clocks, which do not have a tie in Fig. 13.3, appear to be in fairly direct competition, with members voting unfavorably on the animations produced by the other group, and ownage messages being expressed, especially in the animations of the Star Syndicate. Both reviews (positive and negative) and favorite author ties represent attempts by site members to operate on their own social positions. Positive reviews and “favorite” nominations strengthen one’s alliance with others, while negative reviews diminish the popularity of competitors; both are strategies for increasing the popularity of one’s own work.

### 13.4.2 Genre Features

The seven social positions of favorite Flash authors were used to construct a stratified random sample having approximately equal numbers of selections from each of the seven positions. Movies from each of the author groups were pooled, and randomly selected within each group. The codings for the 67 genre features were arranged into an 871 by 67 element matrix, with the genre features as columns, the individual Flash files as rows, and 1 or 0 values in each cell, indicating presence or absence of a feature for a given movie. The matrix was column-wise transformed into z-scores, and a Principal Components Analysis was conducted on the result. A scree plot of the variances of the Principal Components indicates between five and six dimensions of shared variation among the genre features. We adopted a conservative solution having five dimensions. Hierarchical cluster analysis (using Euclidean distance and Ward’s method) was performed on the component scores for the Flash files, yielding the cluster dendrogram in Fig. 13.4. These clusters represent features with shared variation, i.e. features that tend to correlate in some set of movies. A number of different cluster cuts were tried on this analysis; for the subsequent analyses, a cut of 5 clusters was found to be most suitable; larger numbers of clusters caused the data to be too sparsely distributed.

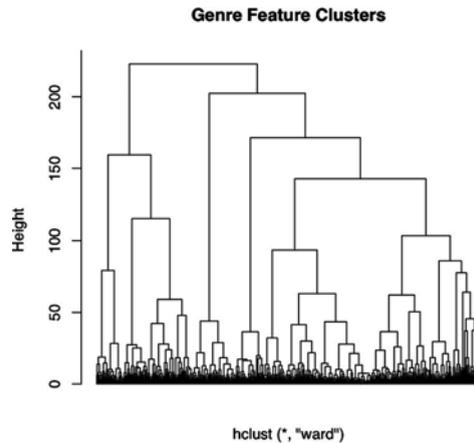


Fig. 13.4 Cluster dendrogram for genre features

For interpretation of the clusters, principal components scores and loadings of the 67 features for the first four PCs are plotted in Figs. 13.5 (PCs 1 and 2) and 13.6 (PCs 3 and 4). The features loading highest on PC 1 (in Fig. 13.5, left frame) are mass-collaborative authorship (A-mass), raster composition techniques (raster-comp), bitmap images, rapid drawing techniques, clock and star avatars (and avatars generally), speakonia voices, and flashing backgrounds. Hence, PC 1 appears to characterize avatar-movies, especially from members of the Clock Crew and the Star Syndicate. The low end of PC 1 appears to be characterized primarily by single authorship (A-sing). Features loading highest on PC 2 are video composition, stop-action animation and composition, and frame-by-frame animation, while those loading lowest are vector composition and animation, use of background color fills and gradients, careful drawing style and camera zooming. PC 2, therefore, appears to differentiate animations according to different animation techniques.

In the scores in Fig. 13.5 (right panel), three clusters stretch in the positive direction on PC 1: Clusters 1, 3 and 5; of these, proportionately more points from Cluster 3 tend in this direction. Otherwise, Clusters 1, 2 and 5 overlap near the origin but are shifted slightly to the negative direction for both PCs, while Clusters 3 and 4 are spread over a diagonal band from the upper left to the lower right. Cluster 3 is highest on PC 1 and neutral on PC 2, while Cluster 4 is highest on PC 2 and neutral on PC 1.

The feature loading the highest on PC 3 is the dramatic story type, along with a number of features also loaded negatively on PC 4: clock and lock avatars, speakonia voices, Newgrounds preloader, and single authorship. Those loading negatively on PC 3 are sound effects and buttons, along with two features also loaded negatively on PC 4: games, and keyboard/mouse interactivity. Features loaded positively on PC 4 are abstract characters, stop-action and frame-by-frame animation, and other preloaders. Hence, PC 3 appears to contrast dramatic avatar movies with interactive

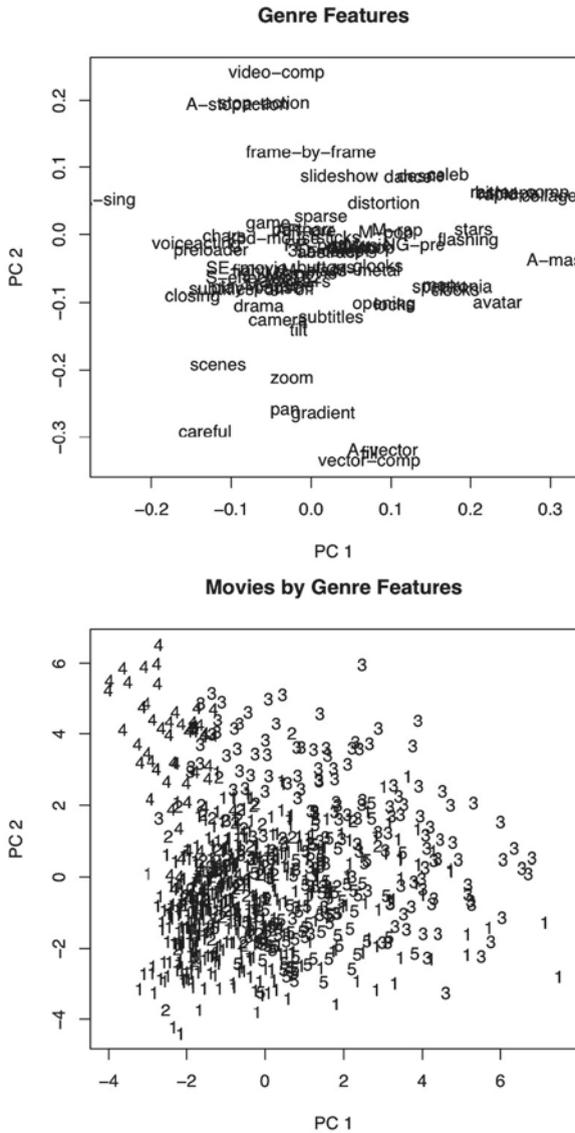


Fig. 13.5 Principal components scores and loadings for genre features on 871 Flash files, PC 1 and 2

games, and PC 4 distinguishes avatar movies from movies with abstract characters and frame-by-frame techniques (e.g. clay animation movies by artist Knox).

Among the scores for PCs 3 and 4, the most striking pattern is the separation of Clusters 2 and 5 from the remaining clusters in the negative direction on PC4, with Cluster 5 in the positive direction of PC 3, and Cluster 2 in the negative direction. Hence it appears that Cluster 3 is characterized by mass authorship and

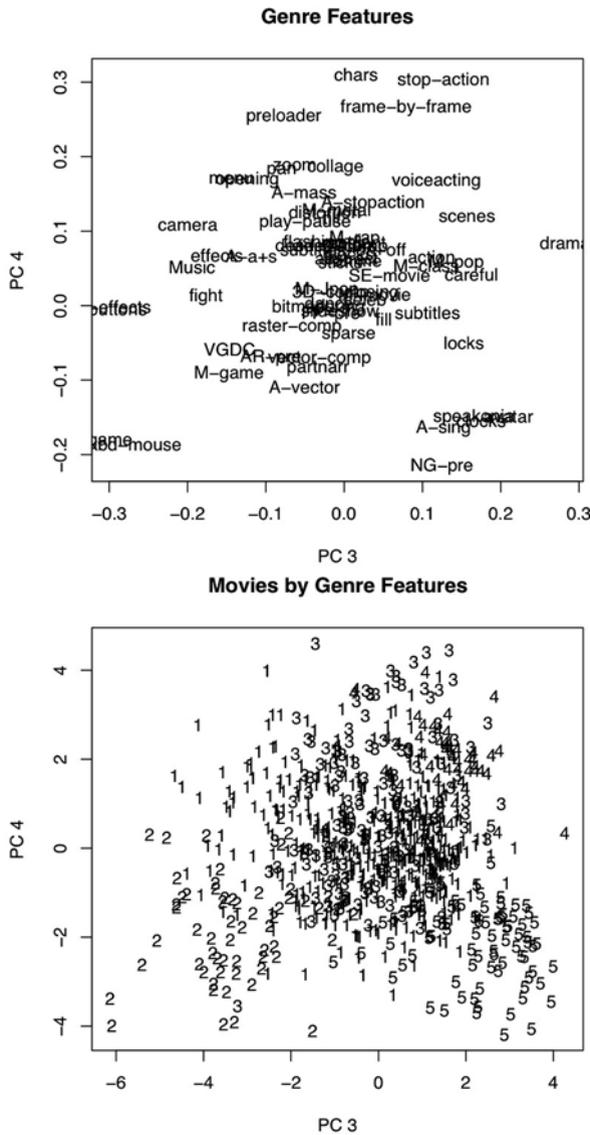


Fig. 13.6 Principal components loadings and scores for genre features of 871 Flash files, PC 3 and 4

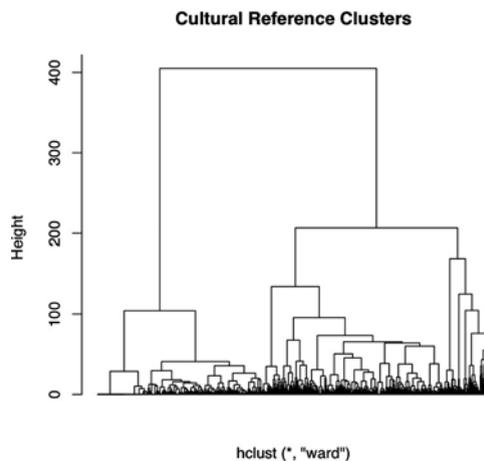
avatar features, Cluster 5 is characterized by single authorship and avatar features, while Cluster 2 is characterized by features of interactive games. Finally, Cluster 4 is characterized by non-vector composition types (video, frame-by-frame, stop-action) whereas Cluster 1 is characterized by relatively neutral values of all these features.

### 13.4.3 Cultural References

Cultural references were treated similarly to the genre features, with the exception that the coding generated many more cultural references than genre features. In our first pass, we constructed an 871 by 3,016 matrix, which was too sparse for successful analysis. Consequently, we needed to select a more restricted set of cultural references. Restricting membership to cultural references occurring 10 times or more resulted in a smaller set of 149 cultural references; the resulting 871 by 149 matrix was found to be suitable for our analysis. Principal components analysis and hierarchical clustering resulted in the dendrogram in Fig. 13.7. Again a cut of 5 clusters was investigated in the subsequent analysis.

The cultural references loading highest on PC 1 (Fig. 13.8) are the yellow “frowny face” icon, the spelling “teh” for “the”, “lol” (an acronym for “laugh out loud”), money, “pwn” (for “own”, or dominate), head, gay, and “omg” (for “oh my god”). Those loaded negatively on PC 1 are games, the Newgrounds Flash artist *Tomorrow’s Nobody*, Flash tutorials, and clay animation. Therefore, PC 1 contrasts abstract symbolic references with references to specific authors and their work. On PC 2, cultural references are only pulled out in the positive direction, and all of these represent specific Newgrounds Flash characters from Clock animations and the letter B. The letter B is important in Clock animations because their inspirational leader, who originally went by the name “Strawberry Clock” notoriously posted a Flash file containing only a still letter “B”, and managed to get people to vote for it so it could pass judgment and not be blammed from the site.

Among the scores on the first two principal components, Cluster 2 is clearly separated out on PC 1, and Cluster 4 is separated out on PC 2. Hence Cluster 4 appears to be associated with Clockcrew animations, while Cluster 2 invokes strong



**Fig. 13.7** Cluster dendrogram of 871 Flash files based on principal component scores of cultural references found in 10 or more movies

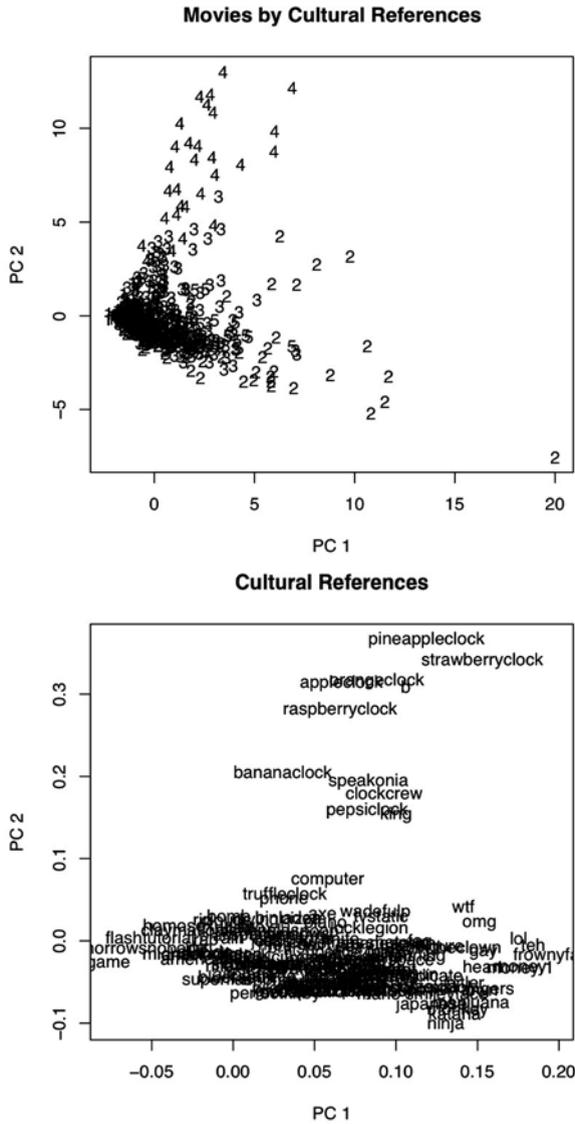


Fig. 13.8 Principal components loadings and scores for cultural references in 871 Flash files, first two principal components

evaluations commonly found in Internet-based youth culture. The remaining clusters are concentrated on the origin in these two principal components.

The cultural reference loadings on PC 3 and 4 (Fig. 13.9) show a more complicated, three-spoke pattern, where the spokes are oriented negatively on PC 3 and neutrally on PC 4, positively on both PC3 and PC 4, and positively on PC 3 but negatively on PC 4. The first spoke concerns themes found in Star Syndicate

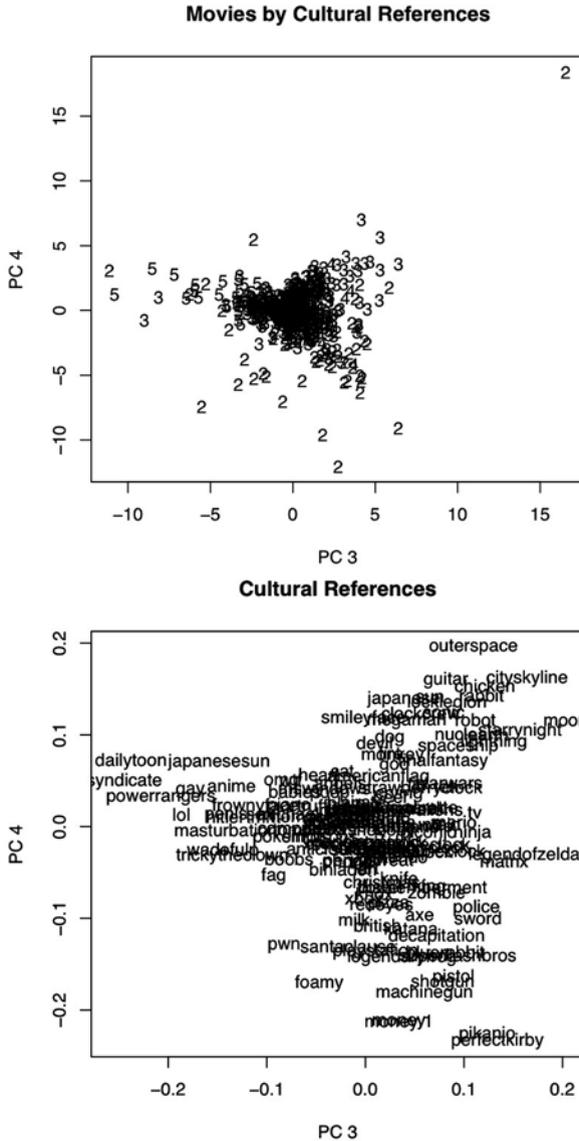


Fig. 13.9 Principal components loadings and scores for cultural references in 871 Flash files, PC 3 and 4

animations. The second involves references to outer space, spaceships, robots, the moon, as well as chickens, rabbits and Clockcrew and Lock Legion characters.

The third spoke involves references to Newgrounds Flash authors Pikanjo, Perfect Kirby, Legendary Frog and Super Flash Brothers, as well as various types of guns. This spoke appears to reference a common plotline in movies by these authors,

where a character raids a military compound and kills everyone he/she encounters using a variety of weapons. These authors' animations also feature frequent references to stick-figure movies (which tend to have similar plotlines) as well as to the Matrix movie series. The same plot line figures in narratives, in which the fight in a military setting provides a context in which two or more rival authors "duke it out". The author composing the animation generally wins, hence, the message of these animations is similar to *ownage*.

Among the scores, these spokes separate Cluster 5, Cluster 3 and Cluster 2 respectively, with the exception that Cluster 2 has members overlapping with Cluster 5 as well as the one member extremely positively loaded on both PC 3 and PC 4. This outlying member is also positively loaded on PC 1, so it probably represents an animation densely packed with cultural references. In fact, all of Cluster 2 loads highly on PC 1, which accounts for its separation from the other clusters, which is otherwise not apparent on PC 3 and 4.

Like the genre features, the cultural references clearly differentiate different clusters of movies. Clockcrew, Star Syndicate and High Production Flash authors are strongly represented in these references, and hence interaction of Newgrounds Flash authors on Newgrounds provides a major resource for subject matter in Newgrounds Flash. In other words, members of the community make reference to their own and others' experience on the site in their movies, but the movies fall into distinct types, with regard to whose experiences they index; these groups partly reflect the articulation of social groups of Flash authors on Newgrounds.

#### ***13.4.4 Genre, Emergence and Social Network***

So far we have evidence of three different patterns among Flash authors on Newgrounds. First, from the favorite author network of Newgrounds authors, we can identify clear structural positions, which appear to differentiate users both by self-identification and in regard to preferred types of content. Second, in the analysis of genre features, there appear to be distinct forms of animation produced by Newgrounds Flash authors. Some of these are not in evidence elsewhere, as they are specialized to the social context on Newgrounds (e.g. features of the avatar animation type). Hence, this and possibly other types may represent emerging genres of Flash animation. Finally, a parallel trend can be observed in the cultural references, in which distinct groups of references to avatars were observed.

In order to fully address the question of genre emergence, we need to ask if these three patterns are related to one another, and whether we can observe change over time among them. To address this, we classified each of the 871 Flash files in our sample according to the social position it was sampled from, and the genre feature and cultural reference clusters identified from them. In addition, we computed a time period based on the Julian date that each animation was originally posted. Six time periods were recognized, based on these dates, with each one being a full year. We then aggregated across each of these variables, counting the number of movies in each combination of categories. The resulting table comprised 270 cells, with

a maximum of 16 in the largest cell. This table was submitted to log-linear (Poisson regression) modeling, to identify which combinations of independent variables were more prevalent, and which changed significantly over time. All independent variables were modeled as categorical variables, except for time, which was treated as continuous.

Log-linear modeling was conducted first from a full main-effects only model, and subsequently excluding non-significant main effects. Main effects for the category variables in this model represent the overall cluster sizes, in terms of their representation in numbers of Flash animations. They are not in themselves interesting to interpret, but they need to be in the model so that the category–category combinations that are tested take into account the relative sizes of the different categories. A main effect for time, as a continuous variable, represents overall growth of submissions on Newgrounds over time. Again, from the perspective of emergent genre, this would not be too interesting, but it is needed to compare against the different category–category combinations over time, so that genre emergence can be properly tracked.

Once a satisfactory main effects model was identified, interactions were systematically tested, to identify category combinations that were more or less common than the relative sizes of the individual categories alone would predict. When significant interactions were found, efforts were made to simplify the categories, using the cluster dendrograms as a guide, so as to arrive at an analysis that accounted for as much significant variation with as few parameters as possible. In this way, we arrived at the model in Table 13.1, which has three significant main effects (including time), six two-way interactions, and one significant three-way interaction with time. In Table 13.1, parameter estimates in our best model are shown along with their Wald tests for significance, and the corresponding significance level. All non-significant effects are excluded from this model. The residual deviance for this model is 319.09, on 259 df, compared to a value of 632.95 on 169 df for the null (intercept-only) model, suggesting that the 10 parameters identified account for about 50% of the observed variation in the counts of cluster combinations.

**Table 13.1** Log-linear model for Newgrounds Flash movies categorized by network position, genre feature clusters and cultural reference clusters, over time

	Estimate	Std. error	z value	Pr(>  z )	
(Intercept)	−0.20700	0.13191	−1.569	0.116594	
Time	0.20463	0.02507	8.163	3.26e-16	*
GF 1	1.09593	0.07982	13.731	< 2e-16	*
CR 2,4	−0.93375	0.13130	−7.112	1.15e-12	*
Clockcrew**GF 3	0.73001	0.12472	5.853	4.82e-09	*
Clockcrew**GF 1,2	−0.34420	0.11204	−3.072	0.002125	***
Fulp et al.**CR 1	0.27749	0.12877	2.155	0.031167	**
Fulp et al.**CR 4	1.31709	0.52106	2.528	0.011480	**
VGDC**CR 5	−1.71193	0.70988	−2.412	0.015883	**
High Production**CR 2	2.79049	0.80070	3.485	0.000492	*
High Production**CR 2:Time	−0.51139	0.19650	−2.603	0.009254	***

Significance codes: \* 0.001, \*\* 0.05, \*\*\* 0.01.

The three main effects in this model are Time, which indicates that the number of animations increases slightly over each time step, the Genre Feature Cluster 1 (GF 1), which is a bit larger than the other clusters, and the Cultural Reference Clusters 2 and 4 (CR 2,4), which are a bit smaller than the other clusters. Two interaction effects suggest that members of the Clockcrew tend to preferentially exhibit Genre Feature Cluster 3 (GF 3), while avoiding Genre Feature Clusters 1 and 2 (GF 1,2). Similarly, two more interactions suggest that Fulp et al. tend to exhibit Cultural References associated with Clusters 1 and 4 (CR 1, CR 4). The remaining two indicate that Cultural Reference Cluster 5 is less frequent among the VGDC group, while Cultural Reference Cluster 2 is more frequent among High Production authors, although this effect is declining somewhat over time.

Our strongest evidence of genre emergence is among members of the Clockcrew, who preferentially employ features associated with avatar movies (principally clock avatars and speakonia voices), while avoiding genre features associated with GF 1 and 2; Cluster 1 is the closest to the origin in the principal components analysis of the genre features, and so not strongly characterized by any specific features; hence, Clockcrew animations tend to be marked by one or more of the genre features we identified. Cluster 2 represents animations with game interactivity, so it appears that Clockcrew animators tend not to make Flash games.

The remaining significant interaction effects indicate that there is a tendency for certain groups to craft messages around certain specific cultural references: Fulp et al. favor messages involving general themes (CR Cluster 1, again at the origin) as well as members of the original Clockcrew, whereas VGDC animators tend to avoid references to the Star Syndicate, and High Production authors initially favored references to strong Internet-culture based evaluative language (“omg”, “wtf”, etc.), but tend to avoid it, even as strongly as they once preferred it, in the later time periods. We do not find strong evidence of association between the cultural reference clusters and the genre feature clusters, suggesting that at least at the level we have observed it, genres are not characterized by specific message content.

## 13.5 Discussion and Conclusions

Our investigation of genre emergence in the amateur Flash of Newgrounds authors has identified the importance of the genre features of avatar animations. These animations utilize highly abstracted characters to represent individuals in the Newgrounds community. Beyond this, they share other characteristics as well, such as their use in acting on the social positions of the individuals represented by their avatars. Social positioning is a pervasive aspect of interaction on Newgrounds, in part encouraged by the site’s design (there are weekly “awards” in several categories). Participants who can successfully elicit support from others stand a better chance of seeing their work pass judgment, and hence users obtain strength by organizing into mutually supportive cliques, which are potentially more successful at surviving in the competitive environment of Newgrounds. Moreover, these

cliques become loci of innovations, such as the avatar animation type, which are then cultivated as emerging genres by members of the clique.

These mechanisms, of competition, clique formation, mutual support, innovation and cultivation, represent potentially important yet previously unobserved processes in genre emergence. Other computer-mediated communication contexts, such as email discussion lists, Usenet newsgroups, weblog networks, wikis, social media sites like YouTube, etc. share many of the same circumstances that led to these processes on Newgrounds. Competition is common in online communication, as are the mutually supporting cliques that often follow it. Consequently it is reasonable to ask whether genre emergence in these contexts, as is arguably happening on YouTube [21], for example, is shaped by similar associations with social network position. To the extent that new circumstances beyond social network position might be observed to be relevant in additional contexts, such investigations can potentially go beyond the present study in further illuminating genre emergence.

Our observations of emergent genres on Newgrounds did not extend to being able to observe change in the structural features of the genres over time. On the one hand, this is a consequence of the size and the complexity of the study: we first had to demonstrate that characteristic message forms existed, and that these were associated with social positions, before we could examine their distribution over time. On the other hand, the fluidity of the genres and the rapid pace of change is also a major factor. For example, the core members of the original Clockcrew organized almost overnight, creating a small number of simple avatar animations in a short period in 2001, early in the history of Newgrounds as a Flash portal. Since then, we have observed several distinct avatar animation types, distinguished by subtle formal features in their avatars: Clockcrew, Lock Legion, Glock Group, Star Syndicate, Block Band, etc. Animators move among these groups as schisms develop and heal, although the Clockcrew and its forms have remained relatively stable over the 6 year period of observation.

Future studies of genre emergence may build on this work in a number of ways. First, we have only used a fraction of the social network information available on Newgrounds: users' reviews of others' Flash are available with rich commentary and numerical ratings on the various characteristics of the contributions. These furthermore have timestamps, so it would be possible to observe more closely the effects of social positioning as well as changing artistic tastes on genre emergence. Such data could substantially enrich the account presented here. In addition, members enter and leave the community, and may traverse it by passing through one or more social positions. Observation of the network dynamics, whether by logging changes in users' profiles, or as enacted in communication via reviews and forum posts, could also enrich the understanding of genre emergence. In connection with this, it may be useful to focus on specific events, whether external to the site (e.g. geopolitical events that influence the topics discussed), or internal ones (e.g. the formation of a schism in an author group). Whatever specific approach is taken, the social network approach coupled with an empirical analysis of message structure and content offers a powerful means to examine community and social process, and thereby illuminate how the adoption of new technologies leads to the development of new communicative forms.

**Acknowledgments** The authors wish to thank four anonymous reviewers and numerous colleagues in SLIS and Informatics at Indiana University for constructive criticism and comments regarding this work. An earlier version of this work was presented in January 2007 at the 40th Hawaii International Conference on System Sciences, Waikaloa, HI.

## References

1. Alonzo, M., and M. Aiken. 2004. Flaming in electronic communication. *Decision Support Systems* 36(3):205–213.
2. Biber, D. 1988. *Variation in written and spoken language*. Cambridge, UK: Cambridge University Press.
3. Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.
4. Biber, D., and J. Kuriyan. 2007. Towards a taxonomy of web registers and text types: A multi-dimensional approach. In *Corpus linguistics and the web*, eds. M. Hundt, N. Nesselhauf, and C. Biewer. Amsterdam and New York: Rodopi.
5. Biber, D., E. Cosmay, K. Jones, and C. Keck. 2007. Introduction to the identification and analysis of vocabulary-based discourse units. In *Discourse on the move*, eds. D. Biber, U. Connor, and T. Upton. Amsterdam/Philadelphia: John Benjamin.
6. Crowston, K., and M. Williams. 1997. Reproduced and emergent genres of communication on the world-wide web. In *Proceedings of the Thirtieth Annual Hawaii International Conference on System Sciences*, IEEE Computer Society. Los Alamitos, CA.
7. Crowston, K., and B. Kwasnik. 2003. Can document-genre metadata improve information access to large digital collections? *Library Trends* 52(2):345–361.
8. Degenne, A., and M. Forse. 1999. *Introducing social networks*. London: Sage Publications.
9. Dillon, A., and B.A. Gushrowski. 2000. Genres and the WEB: Is the personal home page the first uniquely digital genre? *Journal of the American Society of Information Science* 51(2):202–205.
10. Erickson, T. 1997. Social interaction on the net: Virtual community as participatory genre. In *Proceedings of the 30th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society. Los Alamitos, CA.
11. Erickson, T. 2000. Making sense of Computer-Mediated Communication (CMC): Conversations as genres, CMC systems as genre ecologies. In *Proceedings of the 33rd Hawaii International Conference on Systems Science*, IEEE Computer Society. Los Alamitos, CA.
12. Ferguson, C.A. 1959. Diglossia. *Word* 15:325–340.
13. Herring, S.C. 1999. The rhetorical dynamics of gender harassment on-line. *The Information Society* 15(3):151–167.
14. Hymes, D. 1972. *Foundations of sociolinguistics*. Philadelphia, PA: University of Pennsylvania Press.
15. Kayany, J. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the American Society for Information Science* 49(12):1135–1141.
16. Kendall, L. 2007. Colin Mochrie vs. Jesus H. Christ: Messages about masculinities and fame in online video conversations. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society. Los Alamitos, CA.
17. Kim, M.-S., and Raja, N. 1991. Verbal aggression and self-disclosure on computer bulletin boards. Paper presented at the 41st Annual Meeting of the International Communication Association. Chicago, IL.
18. Lee, H. 2005. Behavioral strategies for dealing with flaming in an online forum. *Sociological Quarterly* 46(2):385–403.
19. Longacre, R.E. 1983. *The grammar of discourse*. New York, NY: Plenum Press.
20. Paolillo, J. 2000b. Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics* 36:215–259.

21. Paolillo, J. 2008. Structure and network in the YouTube core. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE Computer Society. Los Alamitos, CA.
22. Spertus, E. 1997. Smokey: automatic recognition of hostile messages. In *Proceedings of the 14th National Conference on Artificial Intelligence*, 1058–1065. Menlo Park, CA: AAAI Press.
23. Strauss, A.L., and J.M. Corbin. 1998. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage Publications.
24. Turnage, A.K. 2007. Email flaming behaviors and organizational conflict. *Journal of Computer-Mediated Communication* 13(1), article 3.
25. Wasserman, S., and K. Faust. 1994. *Social networks analysis: Application and methods*. Cambridge, UK: Cambridge University Press.
26. Yates, J., and W.J. Orlikowski. 1992. Genres of organizational communication: A structural approach to studying communication and media. *Academy of Management Review* 17(2):299–326.
27. Yates, J., W.J. Orlikowski, and J. Rennecker. 1997. Collaborative genres for collaboration: Genre systems in digital media. In *Proceedings of the 30th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society. Los Alamitos, CA.
28. Yates, J., W.J. Orlikowski, and K. Okamura. 1999. Explicit and implicit structuring of genres in electronic communication: Reinforcement and change of social interaction. *Organization Science* 10(1):80–103.

# Chapter 14

## Variation Among Blogs: A Multi-Dimensional Analysis

Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova

### 14.1 Introduction

A blog, short for a weblog, is a website containing an archive of regularly updated online postings. The postings are generally made by one person and presented in reverse chronological order. The archive is generally made freely available to the public. The postings tend to consist primarily of raw text, but may also contain hyperlinks and other media, including picture, video and sound files. Often blogs allow for readers to post comments as well. In terms of content, blogs appear to fall into one of two major types: personal blogs in which an author discusses their own life and thematic blogs in which an author discusses a topic other than themselves. Popular subjects for thematic blogs include current events, politics, arts, entertainment, sports and technology, though in principle any topic is permissible.

Excluding newsgroups, online archives of an individual's postings began to appear in the mid-1990s with two different types of websites: online diaries and commentary pages. Two of the earliest online-diaries are Claudio Pinhanez's diary, which was hosted by the MIT Media Lab website from 1994 until 1996, and Justin Hall's diary, which was updated regularly for 11 years, starting in 1994 [13]. Online diaries naturally conform to many of the key features now associated with blogs: they are regularly-updated and chronologically-ordered online archives of postings written by a single person. These same features also characterize portions of some mid-1990s websites that posted commentary on news and gaming, such as the website *Blue's News*, which featured Stephen Heaslip's commentary on the video game *Quake* starting in 1995. In these websites the postings of experts were logged and archived in a similar manner to online diaries. Over time it appears that these two forms of online writing merged to form the blog variety.

In the late 1990s, numerous factors contributed to a rise in the popularity of blogs. Free easy-to-use blog publishing software was made available, allowing people unfamiliar with web-design to create blogs. *Opendiary*, first offered in 1998, was

---

J. Grieve (✉)  
QLVL Research Unit, University of Leuven, Leuven, Belgium  
e-mail: Jack.Grieve@arts.kuleuven.be

the first of these programs. Similar programs, such as *Live Journal* and *Diaryland*, became available in 1999. Blog hosting services were also introduced in 1999, most notably *blogger.com*, which would eventually be purchased by Google in 2003. These services became popular because they made it easy for potential bloggers to post their blogs and for blogs to be found by potential readers.

At the beginning of the millennium, blogging became an even more important influence on modern culture when numerous political blogs became major web destinations. Two of the earliest political blogs were Bob Somerby's *Daily Howler* and Mickey Kaus' *Kausfiles*, which were launched in 1998 and 1999. But it was not until the early 2000s that major news stories were actually being broken on the blogs. One of the first these of stories was the Trent Lott scandal in 2002. At a birthday party for Strom Thurmond, Lott implied that the United States would be better country today if Thurmond had won the presidency in 1948 on his platform of racial segregation. Mainstream media ignored this speech, but the political bloggers attacked Lott vigorously, bringing the issue to the attention of the public.

It is also during this time that the term *blog* first came to be used. The term is a shortened form of *weblog*, which was coined by Jorn Barger in 1997 to describe the list of web links he had assembled on his website *Robot Wisdom* [27]. The term *blog* was formed in 1999, when Peter Merholz split *weblog* into *we blog* in a play on words [18]. According to the Oxford English Dictionary, the first time this word appeared online was in May of 1999 on Brad Graham's weblog ([www.bradlands.com](http://www.bradlands.com)), in a posting referencing Merholz: "Cam points out [lemonyellow.com](http://lemonyellow.com) and PeterMe decides the proper way to say 'weblog' is 'wee'-'blog' (Tee-hee!)." Five days later, Peter Merholz used the word in context on his weblog ([www.peterme.com](http://www.peterme.com)): "For those keeping score on blog commentary from outside the blog community." By 2002, *blog* was voted as the new word most likely to succeed by the American Dialect Society.

While blogging has thus become an important, common and widely recognized variety of online language (the blog search engine Technorati.com currently searches over 100 million blogs), relatively little is known about its linguistic properties. There have been numerous studies of blogs that have taken a non-empirical and non-linguistic approach, focusing on such topics as the social import of blogging (e.g. [10–12, 20, 21, 24]), the content analysis of blogs [15–17, 23], and the rhetorical analysis of blogs [22]. Very few studies, however, have analyzed the linguistic properties of this variety of language. The major exceptions are Herring and Paolillo [14], which attempts to use selected linguistic features (the relative frequency of pronouns, determiners, and certain function words; cf. [1]) to analyze gender differences in blogs, and Pushmann [25], which examines variation in expressions of futurity in blogs.

There has also been research devoted to identifying the different sub-types of blogs. The main distinction that has been made in past research (e.g. [15, 19]) is between personal blogs and thematic blogs, as defined at the beginning of this section. Krishnamurthy [19], for example, identifies the personal vs. thematic dimension as the most important classifier of blog types. Krishnamurthy [19] also claims that whether a blog is a product of an individual person or a group of people is the

second major determinant of blog type; however, as Herring et al. [15] note, group blogs are relatively rare and are probably best analyzed as distinct individual blogs. In addition, some researchers [14, 15] further divide thematic blogs into two categories: *filter blogs*, which contain a blogger's comments on newspaper articles and postings from other web sites that are either linked to or reproduced in the blog itself; and *k-logs*, short for *knowledge-logs*, which are essentially informational web-sites on a particular topic written by an expert. Pushmann [26] also identifies a *corporate blog* type, where corporations "blog" about their products, though corporate blogs seem to be best classified as a type of thematic blogs, given their narrow focus. While insightful, none of these classifications are based on linguistic data.

While blogs have thus been the subject of previous research, very little is currently known about the overall linguistic characteristics of blogs or the linguistically defined sub-types of blogs. Given the importance of this variety, a linguistic analysis of blogs is clearly of descriptive value. But because the blog is a new variety of language that is still under development the results of this study could be the basis for future diachronic linguistic investigations as well. In addition, the study of linguistic variation in blogs has clear applications in the domain of web-searching: blog search engines could be improved by linguistic-based searches.

The goals of the present study are to identify the principal dimensions of linguistic variation in the blog variety of the English language and to use these dimensions to identify the primary sub-types or *text types* of the blog variety. In order to answer these research questions a 2 million word corpus of blogs was compiled. The values of numerous functional linguistic features were measured for each blog in the corpus and subjected to a factor analysis in order to identify the principal dimensions of linguistic variation for the blog variety. The dimensions produced by the factor analysis were then used as predictors in a cluster analysis, in order to identify the primary blog text types. We hypothesize that these text types will correspond to the categories of personal and thematic blogs.

## 14.2 Corpus Compilation and Analysis

The corpus compiled for this study represents the modern (2003–2005) American blog variety of the English language. The corpus contains 500 sub-corpora each containing text from a single blog, usually spanning numerous postings. These 500 sub-corpora are the basic unit of analysis in this study. Each sub-corpus contains between 9,864 and 1,099 words. On average, each sub-corpus contains 4,500 words. In total, the corpus contains 2,261,520 words.

The corpus was compiled using [globeofblogs.com](http://globeofblogs.com) as an index to locate blogs. Blogs were selected for inclusion in the corpus based on three criteria. First, 10 blogs were selected from each state, using information which was provided by [globeofblogs.com](http://globeofblogs.com). Second, blogs were selected so as to obtain as even a distribution as possible across age and gender; as such, many blogs were excluded when their authors did not provide demographic information. These first two criteria were put in

place because originally this corpus was used to investigate social dialect differences within blogs, but demographic differences are not considered here. Third, blogs were selected that contained as much text as possible. Topic was not controlled.

Once a blog was identified, its contents were then manually downloaded by copying and pasting from the internet browser into text files. In particular, blog posts were copied, while all other text included on the blog websites – including navigation panels, headers and boilerplate text – was excluded from the corpus. In some cases the extraneous text was avoided simply by not copying the extraneous text; in other cases, the extraneous text was deleted from the text files by searching and replacing.

Once compiled, the blog corpus was automatically tagged using the Biber grammatical tagger [4].<sup>1</sup> The current version of this tagger incorporates the corpus-based research carried out for the *Longman Grammar of Spoken and Written English* [9]. The tagger identifies a wide range of grammatical features, including word classes (e.g. nouns, modal verbs, prepositions), syntactic constructions (e.g. WH relative clauses, conditional adverbial clauses, *that*-complement clauses controlled by nouns), semantic classes (e.g. activity verbs, likelihood adverbs), and lexical-grammatical classes (e.g. *that*-complement clauses controlled by mental verbs, *to*-complement clauses controlled by possibility adjectives).

Overall, the tagger performed successfully over the blog corpus, at approximately 95% accuracy. Nonetheless, the tagger was not quite as successful as it would have been if more standardized texts had been analyzed, due primarily to the presence of non-standard spellings in some of the blogs. The tagger, however, is designed to assign tags to words based on grammatical context when those words (or spellings) are not included in its dictionary. The presence of creative spellings therefore caused minimal error in tagging. It is therefore assumed that the results of the study are valid and would be the same if the corpus had been manually tagged without error.

### 14.3 Factor Analysis

In order to identify the main dimensions of linguistic variation in the blog variety of the English language, the values of numerous linguistic features were computed for each blog sub-corpus and subjected to a factor analysis. This procedure identified the major patterns of linguistic co-occurrence across the data set. The resultant factors were then interpreted as underlying dimensions of functional linguistic variation. This section describes the factor analysis technique, presents the results of the factor analysis, and interprets these factors so that the major patterns of functional linguistic variation in the blogs register can be determined.

---

<sup>1</sup> The Biber Tagger is available at the Corpus Linguistics Laboratory at Northern Arizona University.

### ***14.3.1 Method***

A factor analysis is a multivariate statistical procedure that identifies systematic co-occurrence patterns in a set of variables. Essentially, a factor analysis is a method of data reduction: it reduces a large set of variables into a smaller set of aggregated factors by determining which of these variables pattern similarly across the dataset.

The application of factor analysis here is based on the theoretical assumption that register differences involve underlying linguistic co-occurrence patterns. When a speaker or writer shifts from one register to another, they naturally shift from one set of co-occurring linguistic features to a different set of co-occurring features. When applied to linguistic data, factor analysis can therefore be used to identify sets of linguistic features that tend to co-occur across the texts of a corpus. This approach was originally developed to analyze the range of spoken and written registers in English [2, 4], and is applied here to uncover the underlying parameters of linguistic variation among blogs. Similar approaches are adopted in Chapter 1 by Santini (this volume) and Chapter 13 by Paolillo et al. (this volume).

For the present study, rates of occurrence were computed for 131 functional linguistic features (see Appendix) across the 500 sub-corpora, based on tag counts for each sub-corpus. The values of these variables were then subjected to a factor analysis (in SAS, using Varimax rotation). Only 54 linguistic features were retained in the final analysis. Several features were dropped because they were redundant or overlapped to a large extent with other features. For example, the counts for common verbs, nouns, and adjectives overlapped extensively with the semantic categories for those word classes, even though the counts were derived independently. In other cases, features were dropped because they were rare in these web documents or because they did not co-occur significantly with other features in these texts (e.g. semantic classes of phrasal verbs). Finally, some features were combined into more general classes. For example, three features that incorporate a passive verb phrase were originally distinguished: agentless passives, passives with a *by*-phrase, and non-finite passives as nominal post-modifiers. These features were combined into the general category of passive verbs in the final analysis.

### ***14.3.2 Results***

The solution for four factors was selected as optimal. Taken together, these factors account for 40% of the shared variance and are readily interpretable. Subsequent factors accounted for relatively little additional variance. The features with loadings over 0.30 (positive or negative) are listed in Table 14.1; these are the features used to compute factor scores for the individual blog sub-corpora, where factor scores are computed for a text by summing the frequency of the features loading on that factor.

**Table 14.1** Factor loadings

Factor	Loading	Features
1	Positive	Prepositions, attributive adjectives, nominalizations, passives, WH relative clauses, <i>that</i> relative clauses, post nominal <i>to</i> clauses, post nominal <i>that</i> clauses
	Negative	Emphatics, first person pronouns, discourse particles, hedges, past tense, time adverbials, place adverbials, progressive verbs, <i>to</i> clauses with desire/intent/decision verbs, quantity nouns, activity verbs
2	Positive	Present tense, second person pronouns, <i>do</i> as PRO-verb, demonstrative pronouns, <i>be</i> as main verb, indefinite pronouns, WH questions, possibility modals, predictive modals, conditional subordination, necessity modals, mental verbs
	Negative	Prepositions, past tense
3	Positive	Demonstrative pronouns, emphatics, pronoun <i>it</i> , hedges, clausal coordination, adverbs, conjuncts, predicative adjectives, factive adverbs, likelihood adverbs
	Negative	Second person pronouns, nouns
4	Positive	<i>That</i> deletion, past tense, third person pronouns, adverbial subordination (other), <i>that</i> clauses with factive verbs, <i>to</i> clauses with speech act verbs, <i>to</i> clauses with modality/cause/effort verbs, communication verbs
	Negative	Nouns, attributive adjectives

### 14.3.3 Interpretation of Factors

Once the factors have been extracted, it is necessary to interpret each factor in order to explain why particular features co-occur. This is accomplished by considering the function of the linguistic features loading on each factor and the style of blogs with high positive and negative factor scores.

*Factor 1: Informational vs. Personal Focus.* To interpret Factor 1, it is first necessary to consider the functional significance of the combination of features with large positive and negative loadings. Positive features on Factor 1 are prepositions, attributive adjectives, nominalizations, passives, and various post-nominal modifying clauses. The negative features on Factor 1 are emphatics, first person pronouns, discourse particles, hedges, past tense, quantity nouns, progressive verbs, activity verbs, *to* clauses with desire/intent/decision verbs, and time and place adverbials.

Most of the positive features on Factor 1 are associated with nouns and noun modification: nominalizations are derived nouns, attributive adjective and post-nominal modifying clauses are noun modifiers, and prepositions often function as the head of post-nominal preposition phrases. Even passives are associated with a highly nominal style: the passive emphasizes the patient noun phrase. The features on the positive end of Factor 1 are therefore quite homogeneous in terms of their linguistic function: all are associated with a highly nominal style. The functional significance of this grouping of features in actual texts tends to be an information-focused style: in particular, high frequency of nouns and noun modifiers in a text

indicates a high informational density, because noun modification is a strategy for packing as much information into a text as possible.

This highly informational language is exemplified by texts from the corpus that score high on Factor 1. Samples of texts from three of the highest scoring blogs are provided in Table 14.2, where pre- and post-nominal modifiers have been highlighted in italics. All three text samples exhibit a highly informational style: clearly, the goal of each text is to inform the reader about a particular topic. Notice that while the texts all have a similar informational focus, they all discuss different topics: economics, technology and world affairs. This informational style can also be characterized as being quite formal. This is because informational style is usually associated with relatively formal registers like newspaper reportage and academic writing.

While the features with high positive loadings on Factor 1 are nominal, many of the features with high negative loadings are verbal, most notably various subtypes of verbs and adverbs. Several negatively loaded features on Factor 1 are also associated with a high degree of personal involvement, including first person pronouns, *to* clauses with desire/intent/decision verbs, emphatics, discourse particles, and hedges. Many of the negative features on Factor 1 are therefore associated with an involved and informal style. Overall, this finding is similar to most multi-dimensional register studies (e.g. [3, 4, 6, 7]), where the first factor reflects an opposition between nominal/informational discourse and verbal/involved discourse.

However, there is an added consideration with Factor 1 in the present analysis: the negative features loading on Factor 1 do not correspond exactly with previous multi-dimensional studies. Some features normally associated with involvement,

**Table 14.2** Text samples from blogs with highly positive Factor 1 scores

File	Factor 1 score	Text sample
TX02	38.43	Even before <i>bankruptcy</i> filings began rising this spring, an <i>American Bankers Association</i> survey of 350 member institutions found that <i>credit card loan</i> delinquencies had been increasing when measured by the number of <i>accounts past due</i> . When measured by <i>dollars</i> lost, it has declined. In September, it reported that the rate rose to a record of <i>4.81%</i> during the <i>second</i> quarter, driven in large part by the <i>higher price of gasoline</i>
NM06	38.04	It's already come up for discussion with the <i>PeopleFinder</i> project and will probably continue to be an area of <i>discussion</i> in the various <i>Recovery 2.0</i> efforts. The <i>cryptorighs.org</i> project called <i>Highfire</i> is dealing with it from a <i>different</i> approach as well
ND05	36.63	In any event, it is impossible to talk about the <i>CIA leak</i> investigation without getting into a <i>discussion about the justification for the War in Iraq</i> . Libby, Cheney, Rove, and others in the <i>Administration</i> were outspoken about their support of the statements made by the president in his <i>State of the Union Address</i> prior to the <i>Iraq</i> invasion which claimed <i>Saddam had attempted to purchase uranium yellowcake from Niger</i>

such as present tense and 2nd person pronouns, are missing here. In their place we find several features that can be associated with a personal narrative style: first person pronouns, past tense, activity verbs, progressive aspect. The frequency of first person pronouns is the most obvious indication that Factor 1 is associated with a personal style. Texts with high negative scores on Factor 1 have one major topic: their author. The remaining features offer evidence of the nature of the personal discussion in which these texts are engaged. In particular, numerous features are associated with a narrative style (past tense, activity verbs, progressive aspect) and one feature is associated with expressing personal plans or desires (*to* clauses with desire/intent/decision verbs). In other words, these personal texts seem to both recount and comment on the lives of their authors.

Table 14.3 presents blog excerpts that illustrate the function of the negative features on Factor 1, where discourse particles and first person pronouns have been italicized.

The personal nature of these texts is immediately noticeable; further analysis reveals that these texts are both personal narratives and personal commentaries. Compared to the texts presented in Table 14.2, the difference is clear: texts with highly positive Factor 1 scores are far more impersonal, informative and formal than texts with highly negative Factor 1 scores, which tend to be personal, narrative/reflective and highly involved. Overall, Factor 1 therefore seems to reflect a contrast between impersonal/informative writing and personal narrative/commentary writing, and is therefore labeled the *Informational vs. Personal Focus Dimension*.

*Factor 2: Addressee Focus.* The positive features on Factor 2 are present tense, *do* as pro-verb, *be* as main verb, mental verbs, conditional subordination, WH questions, second person, demonstrative and indefinite pronouns, and possibility,

**Table 14.3** Text samples from blogs with highly negative Factor 1 scores

File	Factor 1 score	Text sample
NB09	-30.28	<i>So</i> in high school, ur a loser when you dont have a boyfriend right? <i>well</i> if one of my friends asked me this <i>i</i> would tell them no your not a loser you dont need a guy to complete you. But when <i>i</i> think about myself there is always a question in my mine.. why cant <i>i</i> get a guy, why does he like her more than me.. if <i>i</i> was skinnier maybe he'd go out with me. <i>Well</i> <i>i</i> didnt used to feel that way
AL02	-28.26	<i>I</i> am sick and tired of the girl who does nothing at work. The other day she left without out doing half of the things we are supposed to do everyday. <i>I</i> got fed up of her actions and <i>I</i> sent an e-mail to the boss about it. <i>I</i> put it under the guise of concern for her, <i>I</i> really truly am. Kinda, sorta, maybe . . . <i>Anywho</i> , the next day she still hadn't done her work or put away her stuff!!!! <i>Now</i> she has two full buggy's full of things to put away
MI03	-27.35	No, <i>I</i> didn't quite kill anyone (yet :P) my little sister came in my room without knocking and . . . saw . . . <i>well</i> . <i>I</i> 'm not trying to sound gross to anyone, but the fact was my boyfriend was over. It was horrible . . . <i>I</i> think <i>I</i> 'm scarred for life! <i>I</i> havn't even spoken to her since the scenario. <i>I</i> think <i>I</i> walked in on my older sister and her boyfriend once, but <i>I</i> was only five . . .

predictive and necessity modals. The negative features on Factor 2 are prepositions and past tense.

The positive features on Factor 2 are associated with interactive discourse and addressee focus. The high frequency of second person pronouns is most obviously associated with an interactive style: *you* is a direct reference to the audience of the text. Additionally, the high frequency of other pronouns indicates an interactive style because the frequent use of pronouns implies that the reader is in the same basic frame of reference as the author. Similarly, WH-questions are directly related to interaction between interlocutors, and other features, such as modals, mental verbs, and conditional subordination are used in blogs for giving advice or instruction to the reader. Finally, the high frequency of present tense verbs reflects a focus on current events, in contrast to the past tense verbs that have a negative loading on Factor 2.

Variation in interaction across Factor 2 can easily be seen if some of the most highly positively and negatively scoring blogs are considered. Table 14.4 presents texts samples with highly positive Factor 2 scores and Table 14.5 presents texts samples with highly negative Factor 2 scores. Second person pronouns, WH questions and modals are highlighted in italics in each table.

The blogs in Table 14.4 are clearly far more interactive and addressee-focused than the blogs in Table 14.5: the first blog is basically an advertisement, the second is written in a conversational style, and the third discusses the relocation of a blog. The blogs in Table 14.5, on the other hand, discuss a variety of topics in a variety of styles, but none refer directly to the readership. These differences result in clear differences in the highlighted features: there are 12 highlighted features in Table 14.4 but only 1 highlighted feature in Table 14.5.

Overall, Factor 2 therefore seems to reflect functional variation in blogs associated with the degree of interaction and the focus on the addressee: blogs scoring highly positively on Factor 2 directly refer to and interact with their readership, whereas blogs scoring highly negatively on Factor 2 essentially ignore their audience, choosing instead to simply express information. This factor is therefore labeled as the *Addressee Focus Dimension*.

**Table 14.4** Text samples from blogs with highly positive Factor 2 scores

File	Factor 2 score	Text sample
MN07	19.76	Anybody interested in any custom sewing needs (a friend of mine <i>will</i> be hitting Sharon up for a kilt, I know) <i>should</i> get ahold of her . . . . Even for clothing for the Mundane . . . :) She does that stuff, too. Thank <i>you</i> , Sharon, it's beautiful!
MT08	28.35	Hopefully when I'm on a new shift and not dog-ass tired, I'll be angry/opinionated again. I can't foresee this apathy being permanent. Maybe in the mean time I'll post poetry (fortunately for <i>you</i> , not mine!), or other meaningless drivel. I'm also open to suggestions . . .
ND07	22.56	I hope <i>you</i> find the new place to your liking. And if not . . . . it's my blog and the fact that I like it is all that matters . . . . And those of <i>you</i> who <i>may</i> have me on your blogroll, make sure <i>you</i> change it to the new URL or <i>you'll</i> never know

**Table 14.5** Text samples from blogs with highly negative Factor 2 scores

File	Factor 2 score	Text sample
AL09	-22.87	The trip downstream was beautiful as the sun rose to burn off the early morning misty haze hanging over the water. Once I reached the three mile mark (right at 30:03), the water opened up a little and the breeze started to stir up the surface. It actually helped push me along and my speed improved by 0.1 mph, as I passed Buzbee's
MI08	-21.68	For many reasons, the Caesars made their mark on professional sports in Detroit. It introduced the town to Mike Illitch, who <i>would</i> become a player in Detroit with his ownership of the Tigers, Red Wings, Joe Louis Area, and with ventures like the Fox Theater. It brought what amounted to, in baseball terms, a minor league team to the suburbs of Detroit with crowds of 5,000-plus . . .
WI01	-19.05	Reading on the internet indicates well water with minerals is best for plants, instead of treated city water, or bottled water. Perhaps this switch in water is what caused the African Violet to bloom, perhaps it is the time of year, perhaps because I had started to rotate the plant so different areas receive light from the North window

*Factor 3: Thematic Variation.* The positive features on Factor 3 are pronouns (demonstratives, *it*), emphatics, hedges, predicative adjectives, various adverbs (including conjuncts or linking adverbials), and clausal coordination. The negative features on Factor 3 are second person pronouns and nouns.

At first glance, many of the features loading positively on Factor 3 appear to be associated with a spoken and conversational style: clausal coordination and predicative adjectives are frequent in spoken discourse because they are often the product of unplanned language production; demonstrative pronouns and *it* are associated with generalized and inexplicit references and reduced lexical content, indicative of a shared context of communication; and hedges, emphatics, and factive and likelihood adverbs are all used to express stance, a common function of spoken language. Of course, no blog is spoken, but blogs that have large positive scores on Factor 3 (see Table 14.6) do impart a conversational tone nonetheless. In particular, these blogs shift rapidly from one topic to the next, as is common in spoken discourse. For example, in the first sample, the blogger discusses myspace, golfing and his job in the space of a single one page posting. In contrast, the text samples with highly negative Factor 3 scores (presented in Table 14.7) focus on one topic – not just across these samples, but across all the postings contained in the sub-corpus for that blog. Associating Factor 3 with thematic variation also offers further explanation for the high loading of clausal coordination and conjuncts: if a text discusses numerous disparate ideas and topics, perhaps it is particularly necessary to explicitly connect these parts through grammatical links.

Factor 3 is therefore labeled as the *Thematic Variation Dimension*, where blogs with highly positive scores present information on a variety of topics, while blogs with highly negative scores are focused on a single issue.

**Table 14.6** Text samples from blogs with highly positive Factor 3 scores

File	Factor 3 score	Text sample
PA04	21.45	I just spent the entire evening re-updating my Myspace. What a waste of my life. Oh well, it doesn't matter Today I had my first ever golf outing. You see, it was the first practice for the golf team [. . .] Anyways . . . the days before that, on Saturday and Sunday, I worked. I made \$140 this weekend, \$70 each day
MO01	19.23	I signed a lease last wednesday for my firstest apartment, which is pretty scary. Basically, I agreed to part with a lot of my very own hard earned dollars a month for a year [. . .] I can't remember what my last entry says, but basically Mark is an idiot, and I haven't spoken to him [. . .] At any rate, it looks like they're going to finally sentence jody
ND02	18.90	Student Congress was fine, I think that speaking infront of those people about my opionons on an issue was possibly one of the scariest things I've ever done though [. . .] I purchased the new Broken Social Scene CD [. . .] So, today, Lucy and I went to Wal-Mart

**Table 14.7** Text samples from blogs with highly negative Factor 3 scores

File	Factor 3 score	Text sample
MI08	-18.22	For many reasons, the Caesars made their mark on professional sports in Detroit. It introduced the town to Mike Illitch, who would become a player in Detroit with his ownership of the Tigers, Red Wings, Joe Louis Area, and with ventures like the Fox Theater
NM05	-16.70	The Braves had little chance of retaining his services as Mazzone has had a long-standing agreement with Sam Perlozzo, his best friend since childhood, that he would serve as his pitching coach if Perlozzo got a permanent job
MO09	-14.32	On the various New Testament lists of the Twelve Apostles (Matthew 10:2-4; Mark 3:16-19; Luke 6:14-16; Acts 1:13), the tenth and eleventh places are occupied by Simon the Zealot (also called Simon the "Cananean," the Aramaic word meaning "Zealot") and by Judas of James

*Factor 4: Narrative Style.* The positive features on Factor 4 are past tense, third person pronouns, *that* deletion, factive verbs with *that* clauses, certain forms of adverbial subordination (e.g. since, while), communication verbs, *to* clauses with speech acts verbs, and *to* clauses with modality/cause/effort verbs (e.g. *allow*, *leave*, *order*). The negative features on Factor 4 are nouns and attributive adjectives.

Most of the features that load positively on Factor 4 are associated with a narrative style. The two clearest markers of narration are past tense verbs and third person pronouns. In addition, communication verbs are used to report the speech of others (common in narrative), and the adverbial subordinators that load on Factor 4 (e.g. since, while) are used to make temporal reference – a way to mark time in a

narrative. In contrast, the negatively loaded features on Factor 4 are associated with high information density, similar to the positive features on Factor 1.

The co-occurrence of these features in blogs is illustrated in Tables 14.8 and 14.9. The texts scoring highly positive on Factor 4 are all clearly narratives, whereas the texts scoring highly negative on Factor 4, while accomplishing various communicative goals, do not attempt to tell a story. The differences in pronoun usage, use of communication verbs, and tense are particularly clear (third person pronouns and communication verbs are highlighted in the two tables; neither of these features are found in the second table). Factor 4 has thus been labeled as the *Narrative Style Dimension*.

**Table 14.8** Text samples from blogs with highly positive Factor 4 scores

File	Factor 4 score	Text sample
GA06	19.50	Well folks I stopped at the natural health store and picked up something to help me sleep. I'm hoping it works and I won't have to go the prescription drug route. Wish me luck Spoke to PD today. <i>He called</i> while I was at work and kept me on the phone for hours. At one point <i>he asked</i> if I was still attracted to <i>him</i> . I <i>told him</i> unfortunately, yes
MO06	17.06	There wasn't much we could do, in fact <i>he</i> was highly agitated when we were there and being that my mom was exhausted we decided to go home and give the nurses a rest. Later that night my mom and I went back to see <i>him</i> he was sitting up without restraints but he still didn't know anyone. <i>He was talking</i> more <i>he was saying</i> "I need . . . , I"
VA08	16.95	After the meeting I wanted to <i>speak</i> to the 2 teachers, to see how my boy was doing in his class, but <i>everyone</i> swarmed around <i>them</i> and some of the mothers really got chit chatty with the two teachers and it made it very difficult to <i>speak</i> with <i>them</i>

**Table 14.9** Text samples from blogs with highly negative Factor 4 scores

File	Factor 4 score	Text sample
NM06	-18.21	I've begun to seriously question the viability of the World Future Society. Most of the interesting news I read comes from specialty websites, blogs, and press releases. And with all the material being produced on a daily basis, a monthly publication that devotes the majority of its' article space to a quick review of "Future Jobs" isn't meeting much of a need
TX02	-15.45	Even before bankruptcy filings began rising this spring, an American Bankers Association survey of 350 member institutions found that credit card loan delinquencies had been increasing when measured by the number of accounts past due. When measured by dollars lost, it has declined
NC01	-14.62	According to Heise Online, during a London demonstration of the shipping version of the Xbox 360 hardware, Microsoft showed the console's ability to play nice with Apple's venerable iPod devices. iPods are able to connect with standard Apple cables using the USB ports of the 360

## 14.4 Text Type Analysis

The previous section discussed the major dimensions of linguistic variation in blogs, which were extracted using a factor analysis. These dimensions will now be used to identify the main text types of the blog variety.

A text type is a variety of language that is defined exclusively by linguistic properties [5, 6]. In other words, a text type is a variety of language that is composed of texts that are maximally similar in terms of their linguistic characteristics. Our definition of a text type therefore differs from our definition of a register, which is a variety of language defined by situational (i.e. non-linguistic) characteristics. Text types therefore do not necessarily correspond to registers: text types can be composed of texts from different registers if these texts are linguistically similar. However, while registers are defined based on situational characteristics, empirical analysis has demonstrated that registers are usually characterized by pervasive linguistic features as well [4]. This is because situational context tends to exert functional pressures on linguistic output. Text types and registers thus represent complementary ways to dissect the textual space of a language.<sup>2</sup>

In order to identify the major text types of the blog register, the dimensions of linguistic variation produced by the factor analysis will be used as predictor variables in a cluster analysis, which will group blogs into clusters based on the similarity of their scores across the four factors. These clusters will then be interpreted as text types and interpreted functionally by considering the thematic domains and communicative purposes of the blogs grouped into each type.

### 14.4.1 Method

In the multi-dimensional approach, text types are identified quantitatively using a cluster analysis. A cluster analysis is a multivariate statistical technique used to classify objects into groups based on the values of numerous predictor variables. To identify text types, a cluster analysis is used to group sub-corpora into clusters based on shared linguistic characteristics: the texts grouped together in a cluster are maximally similar linguistically, while the different clusters are maximally different linguistically. This approach has been used to identify the general text types in English and Somali [5, 6], and has also been used to identify text types among web sites [8].

---

<sup>2</sup> In addition, while the concept of a *genre* is not as important in our system as the concepts of *text type* and *register*, we define a *genre* in a very similar manner to how we define *register* – i.e. as a variety of language defined by the external situation in which it is produced. However, while a register is characterized by pervasive linguistic features, a genre is characterized by conventionalized linguistic features.

In the present study, a cluster analysis was used to determine the main text types of the blog register, with the four dimensions of variation produced by the factor analysis used as predictor variables. The resultant groups identified by the cluster analysis are interpreted as representing blog text types.

The FASTCLUS procedure from SAS was used for the cluster analysis. Disjoint clusters were analyzed because there was no theoretical reason to expect a hierarchical structure. Peaks in the cubic clustering criterion and the pseudo-F statistic (produced by FASTCLUS) were used to determine the optimal number of clusters. These measures are heuristic devices that reflect goodness-of-fit – the extent to which the texts within the clusters are similar and the extent to which the clusters are maximally distinguished.

In order to define the text types produced by the cluster analysis, two types of information are considered: the four dimension scores for each cluster, and detailed consideration of prototypical blogs from each cluster, where blogs closest to a cluster centroid are deemed to be prototypical members of that cluster.

#### 14.4.2 Results

Based on an analysis of peaks in the clustering criterion and the pseudo-F statistic, three clusters were identified as the optimal solution. Table 14.10 presents the basic descriptive statistics for the three clusters identified by the cluster analysis: the number of blogs in each cluster, the dispersion within each cluster (maximum distance to the cluster centroid), the nearest cluster, and distance between neighboring cluster centroids.

The vast majority of the blogs (94.6%) were classified into either Cluster 1 or 3, which are also the two clusters that are closest together. Cluster 2, which contains comparatively few blogs, is therefore the most distinctive.

Table 14.11 presents the values for each of the four factors across the three clusters, averaged across the blog sub-corpora in each cluster.

Cluster 1 is characterized by a negative score on Factor 1 and positive scores on Factors 2, 3, and 4, whereas Clusters 2 and 3 are characterized by positive scores on Factor 1, and negative scores on Factors 2, 3, and 4. Clusters 2 and 3 are distinguished by the relative strength of the scores across the factors, especially on Factor 1: blogs in Cluster 2 are characterized by far larger positive values on Factor 1 than blogs in Cluster 3.

**Table 14.10** Cluster analysis results

Cluster	Frequency	Max distance	Nearest cluster	Cluster distance
1	236	29.30	3	21.75
2	27	27.79	3	25.67
3	237	25.65	1	21.75

**Table 14.11** Average factor scores across clusters

Cluster	Factor 1 informational vs. personal	Factor 2 addressee focus	Factor 3 thematic variation	Factor 4 narrative style
1	-11.45	2.15	4.14	3.49
2	32.01	-5.24	-7.55	-8.69
3	7.75	-1.54	-3.26	-2.48

### 14.4.3 Interpretation of Clusters

Based on the factor averages, blogs in Cluster 1 should be more highly personal, both in terms of topic and voice, and relatively more narrative, addressee focused, and thematically variable than blogs in the other two clusters. This interpretation of Cluster 1 can be validated by considering samples from blogs that are prototypical of Cluster 1 (i.e. which fall closest to the centroids of Cluster 1). Samples from the three most prototypical blogs are presented in Table 14.12.

In all cases, these blogs have a clear personal focus – not only in terms of the style of the blog, but also in terms of the thematic focus of the blog. In the first blog sample, the author narrates an event from his or her life. In the second blog sample, the author opines on his or her favorite scary movies and books. In the third blog sample, the author discusses balancing his or her family’s budget. These blogs also appear to be both fairly informal and narrative in their style. This first text type therefore seems to correspond roughly to the personal blog type or the online diary blog type usually identified in past research as one of the major types of blogs.

**Table 14.12** Text samples from prototypical Cluster 1 blogs

File	Dist. to centroid	Text sample
VA06	2.26	As always, Red Lobster was gorgeous and tasted even better knowing that I wasn’t going to have to pay for any of it. I’m terrible like that, I know. It was good conversation, good food, and a good time. K entertained us with his random knowledge of how people can be buried and Minnie told us that she was pulled over by a cop for speeding again on Wednesday, but after some hardcore flirting only got the police officer’s cell phone number
WV07	2.65	Scary movies are so hard to make good through the end. Usually I find them lame, or too gory to deal with. I’d have to say, The Ring came close. As far as books go, it’s about the same deal, so I’m just going to pick the one that freaked me out the most: Whispers, by Dean Koontz freaked me out in ways I never thought possible
MT09	3.52	Balancing the budget here at home is really easy – there is a lot of toys that we want and when there are funds available to meet that end we do it. We never give up our savings right off the top of the payroll. Hopefully after a long time doing it this way – payday can come and we won’t even really notice – nor look forward to. I soooo look forward to that day

Based on cluster averages, blogs in Cluster 2 should be highly impersonal and informational, non-interactive, thematically focused and non-narrative. This assumption can be validated by considering samples from blogs that are prototypical of Cluster 2. Samples from the three most prototypical blogs are presented in Table 14.13. As expected, the three blog samples provided in Table 14.13 are all characterized by a highly informational/impersonal style, clearly presenting the author's strong opinion on a particular subject in an formal style reminiscent of newspaper articles and academic writing.

Similar to Cluster 2, blogs in Cluster 3 are characterized by an informational style as well – although the average Cluster 3 scores for Factor 1 are far weaker than the average Cluster 2 scores for Factor 1. The same is true of the other three dimensions, which are relatively unmarked for Cluster 3, though Cluster 3 is somewhat more addressee focused than Cluster 2. Cluster 3 therefore seems to be an intermediate type between Cluster 1 and Cluster 2. This pattern can be better understood by considering samples from blogs that are prototypical of Cluster 3, which are presented in Table 14.14. These examples reveal the difference between the blogs in Cluster 3 and Cluster 2: while blogs in Cluster 2 are completely impersonal and informational, blogs in Cluster 3 often use a personal voice to discuss and offer opinions on impersonal topics – in all of these most prototypical cases, surprisingly, storms and hurricanes. These blogs appear to be more informal and conversational than blogs in Cluster 2 as well. Cluster 3 therefore falls in between Cluster 1 and Cluster 2, by using a personal voice (like those blogs in Cluster 1) to discuss impersonal topics (like those blogs in Cluster 2).

To summarize, blogs in Cluster 1 are written in a very personal voice and are concerned primarily with the blogger's own life. Given this subject matter it is not surprising that these blogs also tend to be narratives and tend to vary thematically.

**Table 14.13** Text samples from prototypical Cluster 2 blogs

File	Dist. to centroid	Text sample
ND05	3.08	The bottom line is the Administration could have simply corrected Wilson by pointing out that the Vice President's office did not recommend his trip to Niger, and that the CIA was solely responsible for the recommendation. There was no reason to mention that Wilson had a wife who worked in the CIA, whether or not she was in fact a covert agent
KY09	5.12	Supporters of the war in Iraq, and many of President Bush's neighbors in Crawford, Texas, must surely be casting about for some present-day equivalent to the then ratings-challenged Alphanet. Cindy Sheehan and Camp Casey apparently have frayed the nerves of both groups
WV06	7.33	Politics and religion, in their most extreme forms, have taken insight and inspiration and systemized them into imperious and unyielding masters. These masters are ideology and dogma. They appeal to the simple-minded, who take comfort in dualistic absolutes. There is no place for subtlety or nuance. Neither is there openness to change or growth

**Table 14.14** Text samples from prototypical Cluster 3 blogs

File	Dist. to centroid	Text sample
VA04	2.04	Stories I am hearing form the locals are almost unimaginable. One woman was telling me as she stood in her house at the start of the storm she had a few feet of water flood into her house. Then the eye came through and there was quiet. Then the other side of the storm came over and her house flooded so high she, her mom, and daughter had to stand on stools to keep their heads above water
MA04	2.63	Of course, I also saw signs on Topsail that it isn't just storms that can damage beach areas. I was astounded to find one morning that the normal surf was eroding the beach as the tide came in. There had been no overnight storm that I knew of, and the waves didn't seem particularly intense at the time. Still, the beach that had looked like this the day before
VA07	3.73	Don't get me wrong. I am delighted to see what a good job FEMA and all the other agencies seem to have done with Hurricane Rita. It's nice to know that these agencies can do what they are supposed to be doing – providing, of course that you get the political appointees the hell out of the way and let trained people do their jobs!

These blogs also tend to be relatively addressee focused – a functional pattern that seems to reflect the conversational style of these blogs and their author's desire to have their blogs read, enjoyed and commented on by their readers, and perhaps in particular by their friends. Cluster 1 is therefore labeled the *personal diary blog type*. This blog type is very common: out of the 500 of the blogs contained in the corpus, 236 blogs fall under Cluster 1.

Blogs in Cluster 2, on the other hand, are written in very formal and impersonal style and are used by their authors to convey information on a particular topic. These blogs read like newspaper and academic articles because of their similar communicative goals. Cluster 2 is therefore labeled as the *expert blog type*. This blog type, however, is very uncommon: out of the 500 blogs contained in the corpus, only 27 blogs fall under Cluster 2. Cluster 2 is also the most distinct of the three clusters.

The remaining 237 blogs fall under Cluster 3. These blogs, like those in Cluster 2, are informational; however, these blogs are characterized by a relatively personal and addressee focused tone, like those in Cluster 1. In other words, in terms of style, they read much like the personal diary type blogs found under Cluster 1. These blogs are therefore labeled as the *commentary blog type*: they are used by their authors to convey their opinions on one or more topics, but unlike expert blogs they are written in a more personal tone.

Because most of the blogs in the corpus are written in a personal and conversational style that characterizes Clusters 1 and 3 (which contain 473 of the 500 blogs and are also the two closest clusters), it appears that this style is the standard blog voice. Given the nature of the medium – a personal web site which is read and commented on by others – the fact that this tone is generally adopted is not surprising. The major division in blog writing is therefore based on topic: blogs that

focus on their authors' lives (Cluster 1) are distinguished from blogs that focus on impersonal topics (Cluster 3). The expert blog type (Cluster 2) is then seen as a marginal third blog type which is informational, like those blogs in Cluster 2, but which is written in a distinctly non-blog-like style, more consistent with standard forms of informational writing. The basic division between personal blogs and thematic blogs, posited in the introduction and past research, therefore seems to be correct.

## 14.5 Summary of Findings

In conclusion, based on a factor analysis of functional linguistic variation across a 2 million word corpus of blogs, four principal dimensions of linguistic variation were identified, which represent significant patterns of functional linguistic variation for this variety of language. These four principal dimensions are the informational vs. personal focus dimension, the addressee focus dimension, the thematic variation dimension, and the narrative style dimension. The four dimensions produced by the factor analysis were also used as predictors in a cluster analysis in order to identify the major linguistically-defined categories of blogs. Two major blog text-types were identified by the cluster analysis: personal blogs and thematic blogs. Both of these blog types are characterized by a highly personal and conversational style, which appears to be the standard blog voice. The difference between these two types involves the content of the blogs: blogs that focus on their authors' lives are distinguished from blogs that focus on impersonal or informational topics. In addition, a marginal third blog type – labeled the expert blog – was identified, although this blog type appears to be quite rare. The expert blog is informational, like the thematic blog, but is written in a distinctly non-blog-like style, similar to standard informational writing. It was therefore concluded that there are two basic types of blogs: personal blogs and thematic blogs. This finding confirms common assumptions about blog registers.

## Appendix

### *List of 131 Initial Features*

Verbs, private verbs, public verbs, mental verbs, activity verbs, persuasive verbs, communication verbs, occurrence verbs, causative verbs, existence verbs, aspectual verbs, common verbs, pro-verb do, auxiliary have, be as main verb, transitive phrasal verbs, intransitive phrasal verb, mental phrasal verb, communication phrasal verb, occurrence phrasal verb, copular phrasal verb, aspectual phrasal verb, activity phrasal verb, modals, predictive modals, possibility modals, necessity modal, present tense, past tense, perfect aspect, progressive aspect, infinitives, passives, agentless passives, by passives, post-nominal passive, prepositions, pronouns, first

person pronouns, second person pronouns, third person pronouns, demonstrative pronouns, pronoun it, nouns, nominalizations, group nouns, place nouns, quantity nouns, technical/concrete nouns, abstract nouns, cognitive nouns, process nouns, human nouns, adjectives, attributive adjectives, predicative adjectives, color attributive adjectives, evaluative attributive adjectives, time attributive adjectives, size attributive adjectives, topical attributive adjectives, relational attributive adjectives, particles, adverbs, non-factive adverbs, factive adverbs, likelihood adverbs, attitudinal adverbs, adverbials, time adverbials, place adverbials, conjuncts, clausal coordination, phrasal coordination, downtoners, amplifiers, general emphatics, general hedges, conjunctions, conditional subordinators, causative subordinators, concession subordinators, other subordinators, wh words, wh questions, wh clauses, wh relative clauses, object relatives, subject relatives, that relatives clauses, all that clauses, all that clauses with verbs, all that clauses with nouns, all that clauses with adjectives, that clause with non-factive verbs, that clause with factive verbs, that clause with attitudinal verbs, that clause with likelihood verbs, that clauses with factive adjectives, that clauses with attitudinal adjectives, that clauses with likelihood adjectives, that clauses with non-factive nouns, that clauses with factive nouns, that clauses with attitudinal nouns, that clauses with likelihood nouns, all to clauses, to clauses with all adjectives, to clauses with all verbs, to clauses with mental verbs, to clause with desire/intent/decision verbs, to clause with effort verbs, to clause with probability verbs, to clause with speech act verbs, to clause with modality/cause/effort verbs, to clause with all nouns, to clauses with certainty adjectives, to clauses with ability/will adjectives, to clauses with personal affect adjectives, to clauses with ease/difficulty adjectives, to clauses with evaluative adjectives, verb complements, adjective complements, type token ratio, average word length, text length, that deletion, contraction, stranded prepositions, split auxiliaries, post nominal to clauses, post nominal that clauses.

## References

1. Argamon, S., M. Koppel, J. Fine, and A.R. Shimoni. 2006. Gender, genre, and writing style in formal written texts. *Text* 23(3):321–346.
2. Biber, D. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62:384–414.
3. Biber, D. 1987. A textual comparison of British and American writing. *American Speech* 62:99–119.
4. Biber, D. 1988. *Variation across speech and writing*. Cambridge, MA: Cambridge University Press.
5. Biber, D. 1989. A typology of English texts. *Language* 27:3–43.
6. Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, MA: Cambridge University Press.
7. Biber, D. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
8. Biber, D., and J. Kurjian. 2007. Towards a taxonomy of web registers and text types: A multi-dimensional analysis. In *Corpus linguistics and the web*, eds. M. Hundt, N. Nesselhauf, and C. Biewer, 109–132. Amsterdam: Rodopi.

9. Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.
10. Blood, R. 2002. Introduction. In *We've got blog*, ed. J. Rodzwillz. Cambridge, MA: Perseus.
11. Delwiche, A. 2004. Agenda setting, opinion leadership, and the world web logs. Presented at the Annual Conference of the International Communication Association. New Orleans, LA.
12. Gilmore, D. 2003. Moving toward participatory journalism. *Nieman Reports* 57(3):79–80.
13. Harmanci, R. 2005. Time to get a life—pioneer blogger Justin Hall bows out at 31. *San Francisco Chronicle*, 20 Feb 2005.
14. Herring, S.C., and J.C. Paolillo. 2006. Gender and genre variation in weblogs. *Sociolinguistics* 10(4):439–459.
15. Herring, S.C., L.A. Scheidt, S. Bonus, and E. Wright. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society Press.
16. Herring, S.C., L.A. Scheidt, S. Bonus, and E. Wright. 2005. Weblogs as a bridging genre. *Information, Technology and People* 18(2):142–171.
17. Herring, S.C., L.A. Scheidt, I. Kouper, and E. Wright. 2007. A longitudinal content analysis of weblogs: 2003–2004. In *Blogging, citizenship and the future of media*, ed. M. Tremayne. London: Routledge.
18. Kluth, A. 2006. It's the links stupid. *The Economist* 379 (April 22, 2006):5.
19. Krishnamurthy, S. 2002. The multidimensionality of blog conversations: The virtual enactment of September 11. Presented at Internet Research 3.0. Massricht, The Netherlands.
20. Lasica, J.D. 2001. Blogging as a form of journalism. *USC Annenberg Online Journalism Review*. Retrieved 14 Jan 2008 from [www.ojr.org/ojr/workplace/1017958873.php](http://www.ojr.org/ojr/workplace/1017958873.php)
21. Lasica, J.D. 2003. Blogs and journalism need each other. *Nieman Reports* 57(3):70–73.
22. Miller, C.R., and D. Shepherd. 2004. Blogging as social action: a genre analysis of the weblog. In *Into the Blogosphere: Rhetoric, Community and the Culture of Weblogs*, eds. L. Gurak, S. Antonijevec, L. Johnson, C. Ratliff, and J. Reyman. Minneapolis, MN: University of Minnesota Press.
23. Papcharissi, Z. 2004. The blogger revolution? Audience and media as producers. Presented at the Annual conference of the International Communication Association. New Orleans, LA.
24. Park, D. 2003. Bloggers and warbloggers as public intellectuals: Charging the authoritative space of the weblog. Presented at Internet Research 4.0. Toronto, OJN, Canada.
25. Pushmann, C. 2007a. Corpora, blogs and linguistic variation – arguments for using structured web data in corpus development. Presented 8 Nov 2007, at the University of Paderborn, Germany. <http://www.slideshare.net/coffee001/corpora-blogs-and-linguistic-variation-paderborn>
26. Pushmann, C. 2007b. Blogs or flogs? Genre conventions and linguistic practices I coporate web logs. Presented 31 Aug 2007, at the Telematica Instituut, Enschede, The Netherlands. <http://www.slideshare.net/coffee001/blogs-or-flogs-genre-conventions-and-linguistic-practices-in-corporate-web-logs>
27. Wortham, J. 2007. After 10 years of blogs, the future's brighter than ever. *Wired Magazine*. Retrieved 14 Jan 2008 from [http://www.wired.com/entertainment/theweb/news/2007/12/blog\\_anniversary](http://www.wired.com/entertainment/theweb/news/2007/12/blog_anniversary)

# Chapter 15

## Evolving Genres in Online Domains: The Hybrid Genre of the Participatory News Article

Ian Bruce

### 15.1 Introduction

The genre modelling and research presented in this chapter originate from the sub-branch of Applied Linguistics concerned with theorising and designing courses for the teaching of academic literacy skills. Specifically, a model previously employed in the analysis of academic genres [13–18] is used here as a basis for examining the particular web genre of an online news article followed by postings of reader comments – termed here a participatory news article. The chapter first provides an overview of approaches to the categorisation of texts in terms of genres, referring to a number of landmark studies and publications and considering some of the key difficulties in establishing systematic and comprehensive models that are able to account for all of the types of knowledge that writers and readers draw upon in order to identify and ratify a text as belonging to a particular genre category. Following this overview of genre theories and related issues, a dual approach to genre is presented as a way of resolving the difficulties in establishing an appropriate (and comprehensive) theory of genre – that of social genre and cognitive genre [17]. Models for the types of constituent knowledge of social genre and cognitive genre are then presented followed by an explanation of the methodology used to examine the target genre of this chapter, the participatory news article. The findings of the analysis of the sample of participatory news texts are then presented, leading to a discussion of how these findings may relate to the wider issues of recognising and categorising web genres.

First, in order to undertake any meaningful discussion of genre as a categoriser of texts, it is important to define the object of classification and, in particular, the underlying constructs of text and discourse. “Text” is, in effect, a written document or the written record of spoken event (such as the transcription of a dialogue). Widdowson says that “text” is “the overt linguistic trace of a discourse process. As such, it is available for analysis. But interpretation is a matter of deriving a discourse

---

I. Bruce (✉)  
University of Waikato, Hamilton, New Zealand  
e-mail: [ibruce@waikato.ac.nz](mailto:ibruce@waikato.ac.nz)

from the text, and this inevitably brings context and pretext into play” [85, p. 169]. “Text”, therefore, is the written record as it appears on the page (or screen), while discourse includes the written record as well as the social and cognitive operations that surround it, in both its creation and processing. Deriving a discourse from a text may involve context or background information, the expectations and presuppositions of writers and readers, connections with other texts (intertextuality) and the communicative conventions common to a particular discourse community. Therefore, identifying and describing genres (categories of texts) requires comprehensive models that account for both rhetorical organisation and linguistic elements (characteristic of a variety of text) as well as the socially-constructed, contextual and presuppositional knowledge that enables a discourse to be derived from the text. Creating (and classifying) recognisable and ratifiable genres (paper or web-based), therefore, involves elements of knowledge that involve both “text” and “discourse”.

However, two major problems arise when considering existing approaches to categorising texts into genre categories. The first is that few existing approaches to text classification have attempted to provide a comprehensive theory that is able to account for the complexity of intermeshing types of knowledge that relate to both text and discourse. The second is that there is a wide range of terminologies and approaches that are used to classify texts. This multiplicity of approaches to text classification is illustrated by the two lists presented in Table 15.1 following, lists which are merely separated into terms used to classify whole texts and those to classify parts of texts.

It is not the purpose of this chapter (nor is it possible within its scope) to attempt to explain the individual terminologies and unravel the differences that exist among all of the approaches to classification listed here. The purpose in presenting Table 15.1 is merely to emphasise that approaches to text classification are not standardised; there is a multiplicity of terminologies relating to divergent, theoretical approaches. Therefore, any review of theory and research relating to the classification of texts in terms of such categories as “genre” and “text type” needs to acknowledge the fact that terminology is used in very different ways by different researchers. This is not simply a terminological problem of naming or designation. It is also a problem that arises out of fundamental disagreement about the very

**Table 15.1** Diversity of approaches to classifying texts

Whole texts	Parts of texts
Discourse types [81]	Discourse patterns [40–43]
Genres [37, 4, 32]	Genres [76]
Macro-genres [56–58]	Generic values [6]
Text genres [67]	Language styles [11]
	Macro-functions [23]
	Macro-genres [34]
	Macro-structures [79]
	Rhetorical functions [49]
	Rhetorical modes [73]
	Sequences [1, 2]
	Text types [9, 67, 83]

nature of the object of enquiry – what it is that is being investigated and classified. For some, classification of texts in terms of genres is largely a social phenomenon (termed here social genre), something that is directly reflected in texts in terms of their socially recognised purposes and recurrent patterns used in the organisation of their content, for example: editorials, postcards, research articles. For others, genre categories are seen as a more general, rhetorically motivated phenomenon (termed here cognitive genre), such as: argument, explanation, recount or description. In this case, the phenomenon is reflected only indirectly, if at all, in the overall content structuring of whole texts.

Two approaches to genre classification and analysis have been influential in educational contexts: one is the approach of linguists influenced by Systemic Functional Linguistics (the Sydney School) [30, 31, 37, 54, 55, 58, 59, 80] and the other is the English for Specific Purposes approach [4, 5, 7, 26–29, 46, 47, 74–78]. In both cases, genre is seen primarily as a social genre construct, existing in order to achieve some kind of conventionalised social purpose or function within a particular cultural context or discourse community. In both approaches, genres are identified in terms conventionally recognised, organisational patterns for the staging of content, which are related, in turn, to specific linguistic features of exemplar texts. Both approaches are reviewed in some detail here for the reason that they represent more comprehensive attempts to theorise the classification of texts in terms of genre.

### *15.1.1 The Systemic Functional Approach to Genre*

In Systemic Functional Linguistics, emphasis is placed on the social use of language in context. Language is seen as a social-semiotic, a system capable of realising and expressing the entire range of potential meaning employed by a society. Systemic Functional linguistics proposes comprehensive theories for analysing the relationships between the meaning-making that occurs within a society and the manifestation of this meaning-making in language. For example, in order to analyse the operation of language within types of social situation (context of situation), Systemic Functional linguists propose the concept of register. Martin [55] notes that the “organisation of context has to be considered from a number of angles if it is to give a comprehensive account of the ways in which meanings configure texts” (p. 494). Halliday proposes that the different “angles” from which to analyse a social situation (register) are:

[t]he Field of Discourse [which] refers to what is happening, the nature of the social action that is taking place . . . [t]he Tenor of Discourse [which] refers to who is taking part, to the nature of the participants, their statuses and roles . . . [t]he Mode of the Discourse [which] refers to what part language is playing . . . its function in the context, including the channel (is it spoken or written or some combination of the two?) [36, p. 12]

Halliday proposes that each aspect of a context (field, tenor, mode) may be correlated with particular linguistic features of a text. Thus, register is meaning-making within a particular type of social situation or, as Halliday [35] says, a register is “the semantic variety of which a text may be regarded as an instance [and] can

be defined as the configuration of semantic resources that the member of a culture typically associates with a situation type” pp. 110–111.

Some Systemic Functional theorists relate register, language use in a particular type of social situation, to the notion of genre. For example, Eggins [30] says that “[a] genre comes about as particular values for field, tenor and mode regularly co-occur and eventually become stabilized in the culture as ‘typical’ situations” (p. 58). Similarly, Martin [53] defines genre as “a staged, goal oriented, purposeful activity in which speakers engage as members of our culture” (p. 25). The stages or steps that are conventionally followed in the typical organisation of the content of a genre are called the schematic structure. As an example of a schematic structure, Hasan [37, p. 64] describes the essential functional stages of the everyday genre of a sales encounter in a shop as: Greeting, Sales Initiation, Sales Inquiry, Sales Request, Sales Compliance, Sale, Purchase, Purchase Closure, Finis

Thus, the construct of genre proposed by Systemic Functional linguists, therefore, refers to a regularised pattern of purposive language use in a certain social situations that is typical of a cultural group. Genres as categories of texts are classified in terms of their overall social purpose(s) and are able to be described in terms of:

- schematic (or generic) structure, a regularly occurring pattern for the organisation of content, consisting of functionally-related stages that Hasan [37] claims can be reduced to a group of genre-defining obligatory elements drawn from the generic structure potential (GSP) – the range of elements that can potentially occur in the staging of a particular genre; and,
- lexico-grammatical features which systematically correlate with the genre-defining functional elements of the schematic structure or GSP (features that are realised through the related register variables of field, tenor and mode).

### ***15.1.2 The English for Specific Purposes Approach to Genre***

Researchers and writers involved in the field of teaching academic literacies termed English for Specific Purposes (hereafter ESP) sometimes use genre as a classification device to identify types of text that have a common purpose or goal within a certain field of academic activity. Examples of such genres that have been analysed for ESP purposes are: introductions to research articles [74, 76]; science dissertations [26, 27, 44]; popularised medical texts [62]; job application, sales promotion letters and legal case studies [4], and grant proposals for European Union research grants [22].

Among ESP researchers and theorists, Swales [76] provides the most detailed proposal for a theory of genre, a construct that he describes as “a class of communicative events, the members of which share the same communicative or rhetorical purpose” [76, p. 58]. In providing a working definition of genre, Swales [76, pp. 45–57] includes the following defining features:

- A genre is a class of communicative events.
- The principal criterial feature that turns a collection of communicative events into a genre is some shared set of communicative purposes.

- Exemplars or instances of genres vary in their prototypicality
- The rationale behind a genre establishes constraints on allowable contributions in terms of their content, position and form.
- A discourse community's nomenclature for genre is an important source of insight.

Below the genre level at which categorisation is determined by a set of communicative purposes, the text-internal elements of content organisation and linguistic encoding are analysed in relation to: (a) moves and steps; and, (b) linguistic structures which systematically relate to these moves and steps. Moves and steps, like the "schematic structure" of the Systemic Functional approach to genre, are regular patterns for organising content within a certain category of text (genre). Dudley-Evans [29] suggests that "decisions about the classification of the moves are made on the basis of linguistic evidence, comprehension of the text and understanding of the expectations that both the general academic community and the particular discourse community have of the text" p. 226. For example, Swales [76, p. 141] proposes a three move structure for the introductory section of research articles, consisting of:

- Establishing a territory;
- Establishing a niche;
- Occupying the niche.

This structural pattern is then related to the linguistic elements which may occur within the move framework.

Swales, [75, pp. 212–213; 76, pp. 24–27] proposes that genres exist within discourse communities. A discourse community is a socio-rhetorical network that exists to achieve certain goals. To achieve these goals, it has certain commonly used and understood configurations of language, which may involve some specialised vocabulary. However, Swales' [76] proposal for discourse communities has been subsequently challenged in a number of areas (for a summary, see Borg [12]). Issues that have been raised include: how large a discourse community might be; whether spoken language should also be a necessary defining element; the role of purpose as a defining element and the degree of stability a discourse community ought to have. However, in relation to genre studies, the notion of discourse community remains an important concept partly because it is more inclusive than Lave and Wenger's [51] competing construct of 'community of practice' which has more limiting requirements of "mutual engagement" and "joint enterprise" [82, p. 78], elements that preclude the existence of the potentially more disparate 'discourse communities' that Swales proposes. Later Swales [77, p. 204] distinguishes between the broader concept of a discourse community, the members of which may not be physically connected, and which communicates with itself through written communication and place discourse communities, which use both written and spoken communication. For example, in relation to the genre of the participatory news article (analysed in this chapter) is located within the extremely large and disparate discourse community of the news-reading public.

### *15.1.3 Problems with these Existing Approaches to Genre*

Both of the theories of genre reviewed here attempt to relate the conventionally-recognised, organisational staging of the content of texts (“schematic structure” or “moves and steps”) to associated sets of linguistic features, this combination of characteristics then being used as the means for identifying and classifying texts as examples of a particular genre. However, there are important issues that arise with this approach to genre in relation both to the role of linguistic knowledge in identifying genres and to categorisation theory.

First, the notion of identifying genres in terms of characteristic linguistic features has been challenged as the result of an extensive corpus-based study by Biber [9], who concludes that “[g]enres are defined and distinguished on the basis of systematic non-linguistic criteria, and they are valid in those terms” [9, p. 39]. On the other hand, in his corpus study, by examining linguistic features in relation to a number of dimensions, Biber found systematic clusterings of linguistic patterns that related to more general, non genre-specific text types. (“Text types” are usually smaller sections of text relating to a single, more general, rhetorical purpose, such as to retell a sequence of events, present an argument or provide an explanation.) Thus, Biber sees conventionally-recognised genres as distinct from text types, which are more general, non genre-specific, categories of text. More recently, in a variation of this approach, text types have also been identified in terms of clusters of vocabulary-based discourse patterns (VBDUs) [38], referring to “a block of discourse defined by its reliance on a particular set of words” [10].

Similarly, Brian Paltridge’s [64, 65] research, which employed the approach to genre influenced by Systemic Functional Linguistics, also challenges claims of deterministic relationships between recurrent content-organising patterns and linguistic features. As a result, Paltridge [66] argues that courses that teach academic writing need to focus on both genre and text type knowledge in order to account for the wide range of types of knowledge involved in the creation of extended texts. This dual approach to genre knowledge is also supported by Pilegaard and Frandsen [67], who make a similar distinction of “text genres, (e.g. novels, instructions, newspaper editorials, legal text or business letters); [and] . . . text types . . . (e.g. narrative, expository, descriptive, argumentative or instruction text types)” p. 3. Furthermore, a dual approach to examining genre knowledge is proposed by Bhatia [7], who proposes that genre knowledge needs to be investigated from two perspectives: an ethnographic perspective and a textual perspective p. 163.

Secondly, in proposing genres as complex categories (of text), neither the Systemic Functional nor the ESP approaches to theorising genre comprehensively incorporates the theories and research findings from cognitive science relating to human categorisation, such as, the roles of prototypes (or exemplars), levels of category knowledge (higher level general to lower level specific) and the types and roles of schematic knowledge see [19, 20, 48, 50, 63, 70].

In relation to category membership, the ESP approach to genre acknowledges prototype theory [69]. Prototype theory proposes that a category, such as a genre category may include a range of members from highly prototypical texts that closely

reflect the features of the genre category through to others that less closely reflect the category features. However, prototype theory is not accommodated within the Systemic Functional approach to the identification of a genre on the basis of obligatory elements of a GSP (if, in fact, a GSP can be clearly identified). Given the wide range of discourse elements that Paltridge [64] found possible when examining a set of texts introducing research in one discipline and the difficulty of setting up a GSP, the usefulness of this concept for identifying the higher level organisational elements of a genre category may be questionable.

Thus given the need for flexibility and inclusiveness when identifying the higher level, internal organisation of texts within a genre category, it is important to interrogate the adequacy of “generic/schematic structures” or “move and step” analyses to account for all higher-level, text-organising structures. For example, more general, rhetorical structures, which Carrell [20] terms “formal schemata”, are not included within Systemic Functional approaches to genre. On the other hand, while the ESP approach acknowledges the roles of two types of schema, content (move and step structure) and formal (rhetorical structure), Swales [76] suggests that it may be difficult to maintain a distinction between the two when examining a genre in that: “the nature of genres is that they coalesce what is sayable with when and how it is sayable” p. 88. However, in failing to consider the more general, rhetorical, organisational dimension as an organising influence on text and discourse, it appears that the both the Systemic Functional and ESP approaches to genre rely solely on matching patterns of content staging (schematic or move and step structures) to linguistic features.

Therefore, as a response to these concerns, it is suggested that an adequate approach the classification of texts in terms of a genre category must involve three elements:

- the social motivation, which relates to the recognised, conscious level of a whole text, its socially recognised function and conscious organisation;
- the cognitive organisation of intermediate-level, general rhetorical structures that involve a less conscious, more automatic use of language-organising structures often described in more abstract terms, such as exposition, argument and narrative; and,
- the actual linguistic realisations of the social and cognitive knowledge.

#### ***15.1.4 A Solution: Social Genre and Cognitive Genre***

Thus, in my own research and work on developing materials for academic writing courses, I have attempted to account for these different areas of knowledge by proposing a dual approach to genre: that of social genre and cognitive genre.

Social genre refers to socially recognised constructs according to which whole texts are classified in terms of their overall social purpose. Purpose here is taken to mean the intention to consciously communicate a body of knowledge related to a certain context to a certain target

audience Cognitive genre . . . the overall cognitive orientation and internal organisation of a segment of writing that realises a single, more general rhetorical purpose to represent one type of information within discourse. Examples of types of general rhetorical purpose relating to cognitive genres are: to recount sequenced events, to explain a process, to argue a point of view, each of which will employ a different cognitive genre. [16, p. 39]

Social genres and cognitive genres are not mutually exclusive categories, but, in effect, two sides of the same coin, or two complementary approaches to examining the discursual and textual elements of a genre. This dual approach to genre is similar to that proposed by Santini in her chapter of this book, with social genres accounting for what she refers to as “web genres” and cognitive genres accounting for the textual entities that she refers to as “rhetorical genres”. The proposal for social genre also accords with the key idea in Sharoff’s chapter that the identification of a genre relates closely to its socially driven purpose and social function. It also overlaps with elements of the “social network analysis” dimension of web genres proposed in the chapter by Paolilo, Warren and Kunz (who examine the communications of digital animators), but differs in that it focuses more on the language consequences of a discourse community’s knowledge and practices rather than the actual sociological features of the community that uses the target genre. This is because of the lack of organisational structure and the more disparate interests of the discourse community of the news reading public. Thus, discourse theory [21] rather than sociological theory provides the basis for the social genre framework proposed in this chapter.

The social/cognitive genre model as articulated in Bruce [17, p. 131] proposes that understanding the nature and operation of a social genre (such as a category of written texts within an academic or professional setting), involves knowledge relating to:

- context, which Widdowson [85] suggests involves specialist knowledge of a field and its particular language (technical lexis);
- epistemology, which Lea and Street define as “disciplinary assumptions about the nature of knowledge” [52, p. 162];
- writer stance, involving issues of addressivity and audience, such as Hyland [45] describes in terms of the use of metadiscourse; and,
- content schemata, the conventionalised, conscious staging of content in texts, such as schematic structure [37] or systems of moves and steps [76].

In relation to the categorisation of extended, written discourse at the level of cognitive genre, involving procedural or organisational knowledge, I propose that:

- certain types of general, rhetorical purpose instantiate a small number of prototypical textual patterns (cognitive genres), which are, in effect, a type of highly complex category;
- as complex categories, cognitive genres may be described in terms of different systems of intermeshing procedural (organising) knowledge, which relate hierarchically (higher level general and more specific lower level structures);
- this procedural knowledge is fundamental to a cognitive genre, and this has a considerable influence on linguistic choice.

Cognitive genres are, therefore, segments of text sharing common characteristics, sometimes referred to as “text types” (see Chapter 14 by Grieve et al., this book). In the cognitive genre model these are segments of text that relate to a single rhetorical purpose. Table 15.2 following is a summary of the proposed cognitive genre model.

The cognitive genre model draws three important ideas from theories of categorisation in cognitive science: the relationship between purpose or intentionality and category formation, the hierarchical organisation of complex knowledge and the role of metaphor in category organisation.

First, based on the idea from cognitive science that categories are formed in relation to intentionality and purpose (see [3, 61] drawing on the two existing taxonomies of text types [9, 68], the types of purpose (see Table 15.2, “rhetorical focus”) of the four cognitive genres that occur most commonly in academic prose are:

**Table 15.2** Summary of the cognitive genre model

Report: static descriptive presentation	
Rhetorical focus	Presentation of data that is essentially non-sequential
Gestalt structure	WHOLE PART structure of which PART has an UP DOWN structure
Discourse pattern	Preview-details
Interpropositional relations	Means-purpose, means-result, simple contrast, simple comparison, concession-contrarexpectation
Explanation: means-focused presentation	
Rhetorical focus	The presentation of information with a focus on means
Gestalt structure	SOURCE PATH GOAL schema, LINK schema
Discourse pattern	Preview-details
Interpropositional relations	Means-purpose, means-result, amplification, concession-contrarexpectation
Discussion: choice, outcome-focused presentation	
Rhetorical focus	Focus on the organisation of data in relation to (possible) outcomes
Gestalt structure	CONTAINER schemata
Discourse pattern	Generalisation-examples, matching
Interpropositional relations	Grounds-conclusion, reason-result, means-result, concession-contrarexpectation
Recount: sequential presentation	
Rhetorical focus	Presentation of data or information that is essentially sequential or chronological
Gestalt structure	SOURCE PATH GOAL schema
Discourse pattern	General-particular, problem-solution
Interpropositional relations	Means-purpose, means-result, amplification, chronological sequence, grounds-conclusion, reason-result

- the presentation of data or information that is essentially non-sequential (termed Report);
- the presentation of information with the orientation on means (termed Explanation);
- a focus on the organisation of data in relation to (possible) outcomes, conclusions, or choices (termed Discussion);
- presentation of data or information that is essentially sequential or chronological (termed Recount).

Secondly, drawing on the idea from categorisation theory that units of complex knowledge are hierarchically organised (higher level general structures to lower level, more specific structures), the cognitive genre model employs the following, top-down, cognitive systems of classification:

- **gestalt structure** At the upper level of the model, the rhetorical purpose will engage high-order, gestalt patterns termed image schemata [48] that broadly structure the content knowledge to be represented within the particular segment of text. This is based on the idea that gestalts provide a metaphorical basis for upper level category organisation in the way proposed by Lakoff [50, p. 283] in his “spatialization of form hypothesis”.
- **discourse patterns** While gestalts (image schemata) refer to the organisation of concepts or ideas, in relation to the overall organisation of the actual written text, they lead to the engagement of non-genre-specific discourse patterns (e.g. General-Particular, Problem-Solution) which have typical patterns of co-occurrence [41–43].
- **interpropositional relations** Rhetorical purpose also influences selection from a specific set of lower-order, cognitive categories termed interpropositional relations, e.g. Reason-Result, Chronological Sequence, Condition Consequence (see [24]). These are binary relations between propositions, which have a direct effect on linguistic organisation and selection central to the cohesion and coherence of a text.

The four types of rhetorical purpose of the cognitive genre model have been developed in relation to the more general descriptions of the four text types that Biber [9, p. 39], in an extensive corpus study, found to occur most frequently in academic prose. Biber’s typology has been critiqued more recently in relation to the types of texts included in the corpus and the opaqueness of some of his terminologies [71]. However, since the four text types relating to academic written prose largely mirror those of the typology of Quinn [68], which are based on educational needs analysis, they are selected to form the basis for the cognitive genre model. In terms of their structure and internal organization, they are conceptualized here in terms of top-down, cognitive structure rather than by their linguistic and stylistic features, which is the approach used Chapter 14 by Grieve et al., this volume.

Consideration was given to the applicability cognitive genre model to the web genres of the participatory article in an online news domain. After examination of the features of the other four text types in Biber’s [9] typology (intimate

interpersonal communication, informational interaction, imaginative narrative or situated reportage), it was decided that the categories previously applied to the analysis of academic written texts were sufficient to account for the range of types of communicative purpose that relate to the participatory news genre that is the focus of this study.

### ***15.1.5 A Web Genre: The Participatory News Article***

The target genre in this study is the web genre of the “participatory news article. This is an online news article immediately followed by readers’ comments. In the present stage of its evolution, it seems that the participatory news article as a web genre is a concatenation of the established genre of the written newspaper article from the arena of what Young [88] terms public discourse, a conventional, rhetorically organised text belonging to a recognised genre category, while reader comments belong to the area of interactive discourse, and tend to share the characteristics of informal, written interactions typical of email communication. Another way of describing the two parts of the concatenation may be in terms of the continuum used by Herring et al., [39, p. 10] as a way of categorising weblogs. At one end of their continuum Herring et al. [39] place standard web pages, which have the characteristics of “asymmetrical broadcast” and “multi-media”, which, in effect, describes the form of the news article part of the genre. At the other end of their continuum, Herring et al. place “asynchronous CMC” which they describe as having the features of being “constantly updated, [a] symmetrical exchange [and] text based”, characteristics that describe the readers’ comments part of the genre concatenation. Thus, the web genre of the participatory news article appears to bring together texts from the two polar ends of the continuum. The news article part of the concatenation usually, but not always, appears online using the same text that appears in the print edition of a newspaper, falling within the web genre category that Shepperd and Watters [72] term extant – “genres as they appear in their source media” p. 98, while the reader comments section could be described as novel — “genres wholly dependent on their new medium” p. 99.

A number of studies concerned with the identification and classification of web genres appear to draw upon North American approaches to genre theory (termed new rhetoric) deriving from the ideas of Miller [60] and exemplified in the seminal web-genre study of Yates and Orlikowski [86], which examined types of electronic communication within an organisation. In this approach to genre, the central focus tends to be on the social actions that surround genres, including their institutional functions or roles. To provide a framework within which to perform this type of genre analysis, Yates and Orlikowski (and others) employ structuration theory [33] as a way of examining the social relations and interactions surrounding a genre. They propose that each genre has a socially recognised communicative purpose and a common characteristic of form, in terms of the physical appearance and presentation of web genre documents (such as online forms, reports and memos). However, while this approach may account for the social function and physical appearance of

a web genre, it may have less capacity to provide a nuanced analysis of the characteristics of the textual manifestations of a web genre as “the linguistic trace of a discourse process” [85, p. 163].

Rather than identify web genres primarily in terms of social purpose and related (physical) forms, other web genre theorists have focused on the identification of the textual features of web genres. Santini [71], for example, drawing on existing linguistic theory of genre and text type [83] and corpus research [8, 9] focuses on the intermediate level of organisation of text type – what I have termed here “cognitive genre”. It seems that this approach may have more potential as a systematic identifier of the textual resources employed by different genres from different domains. This is because texts, subjected to analysis at this level, appear to be amenable to systematic description and identification, carried out in the present study in terms of clusters of relations between propositions, discourse and gestalt patterns. However, it must be noted that the study reported here, the systematic identification of text types (cognitive genres) is achieved by inferential rater analysis (analytical judgments made by a human rater), and does not rely, in the first instance, on corpus searching although the creation of wordlists and concordance searches can be used to confirm rater-identified features see [16, 18].

In the study reported here, it is proposed to apply the social genre/cognitive genre model to examining the participatory news article in order to achieve a more finely “nuanced” analysis of the web genre, an analysis that accounts for its socially constructed, rhetorical-organisational and linguistic features, that is, an examination of the particular genre as both text and discourse.

## 15.2 Methodology

This study applies the social genre/cognitive genre model to the analysis of the web genre of the participatory news article (an article from the online edition of a newspaper with a series of reader comments attached). The study is small-scale and exploratory, involving the analysis of a sample of ten participatory news articles from the online editions of seven different newspapers. The visual composition of the web page, the use of graphics and the inclusion of links are not part of the focus of this study, but rather the genre model is applied to an analysis of the textual and discursal elements of the participatory news article genre.

The sample of participatory news articles was selected from English language newspapers from the United Kingdom, the United States, Canada, Australia, New Zealand and South Africa. Inclusion of texts in the sample was based on two broad criteria: the first is that the article deals with issues or events that are of interest to a wide (and not necessarily national) readership; and the second criterion is that it includes a body of comments from readers about the article or related issues. The total sample (of both the news articles and reader comments) comprises 108,693 words. The total number of words in the news articles alone is 14,064, giving an average length of 1,172 words per article. The total number of reader comments is

Table 15.3 Sample of participatory news articles

Date	Title	Subject	Source
28/06/08	Christians challenge teaching of evolution	A Christian group supplies creationist materials to New Zealand schools	The Dominion Post (Wellington, New Zealand)
16/07/08	Hezbollah hands back Israeli soldiers in coffins	Israel exchanges Palestinian prisoners for the bodies of Israeli soldiers	The Times of London
25/07/08	Barack Obama – the world can expect better of America	Barak Obama gives a speech at the Berlin Victory Memorial	The Times of London
26/07/08	California is the first state to ban trans fats	The state of California passes a law banning trans fats in restaurant products and retail baked goods	The New York Times
30/07/08	Goodbye, Starbucks. Hello, coffee	Starbucks announces the closure of 61 of its 84 stores in Australia	The Sydney Morning Herald
1/08/08	Radovan Karadzic stands defiant before Hague war crimes tribunal	The Bosnian Serb leader's first indictment before The Hague tribunal	The Times of London
4/08/08	Aleksandr Solzhenitsyn dies at 89	The death of the Russian writer Aleksandr Solzhenitsyn	The New York Times
14/08/08	Russia: Georgia can “forget” regaining provinces	The war between Russia and Georgia over South Ossetia	The Globe and Mail, Toronto
22/08/08	US, Iraqi negotiators agree on 2011 withdrawal	Negotiations between the US and Iraqi Governments concerning the eventual withdrawal of US troops from Iraq	The Washington Post
26/08/08	Food riots as Zimbabwe aid ban continues	The continued ban by the Zimbabwean government on the distribution of food by aid organisations	The Mail and Guardian, South Africa

996, producing an overall total of 94,629 words, with an average length of 95 words per reader comment over the ten texts of the sample. The titles, content and source of the participatory news articles are summarised in the Table 15.3.

Each text (news article and comments) was downloaded and one copy was printed for rater analysis in relation to the elements of social genre and cognitive genre knowledge that they employ. Following a bottom-up approach, the paper copy of each news article text was marked-up with the cognitive genre elements of interpropositional relations, discourse patterns and gestalt structures. The social genre elements of metadiscourse devices (relating to author stance) were then marked on the text. Judgements in relation to social genre elements of context and epistemology and content schema were then made after further close reading. Figure 15.1 shows an example of analysis of part of a text to illustrate the type of mark-up that was applied to the news articles.

The reader comments sections of each of the ten texts of the sample were analysed in two ways. First each comment was read and assigned to one of three broad categories: a positive evaluation of the content of the new article; a negative evaluation of the content of the news article or a comment that related to the statements of another reader or that added some other point of information either about the article or another comment. (It must be emphasised that these three categories for reader comments are not hermetic; they are merely an indication of the main focus of a comment.) Secondly, the reader comments were examined for elements of the message that relate to the social genre/cognitive genre model in terms of both their organisation and the types of language resources that they employed. Because of their brevity (average length of 95 words), individual comments are not able to display the same complexity of organisation as the longer news article section of the web genre.

Gestalt Structure	Discourse Pattern	Text (Part of Move 1 Event – Recount)	Interpropositional Relations
	<i>Preview</i>	<b>Radovan Karadzic stands defiant before Hague war crimes tribunal</b>	
		David Charter in The Hague	
SOURCE ↓ PATH		Shorn of the disguise that had helped to keep him at liberty for 13 years as a fugitive, a defiant Radovan Karadzic appeared before a war crimes tribunal yesterday, complaining about his arrest and talking darkly of a plot to assassinate him.	Amplification } Bonding } Chronological Sequence
	<i>Details</i>	The former Bosnian Serb leader invoked the ghost of Slobodan Milosevic, once a defendant in the same courtroom in The Hague, by saying that he was receiving guidance from "an invisible advisor" and declaring that he would represent himself. The latter tactic was used effectively by Milosevic to stretch out his own trial before he died in custody.	Result } Means } Bonding } Chronological } Sequence } Means Purpose }
		Looking gaunt at his first appearance in Court One at the International Criminal Tribunal for the Former Yugoslavia, Dr Karadzic refused to enter a plea but used his right to take 30 days to study the 11 charges against him. In an initial hearing lasting just over an hour, Dr Karadzic declared that he remained a citizen of all three countries that he once hoped to unite as Greater Serbia.	Contrastive } Alternation }
		Retrieved August 1, 2008 from <a href="http://www.timesonline.co.uk/tol/news/world/europe/article4438630.ecc">http://www.timesonline.co.uk/tol/news/world/europe/article4438630.ecc</a>	Amplification } Amplification }

Fig. 15.1 Example of text mark-up

## 15.3 Results

The findings of the analysis of the web genre are presented in two stages: first a framework for the news report section of the web genre is presented in terms of the elements of the social genre/cognitive genre model. Following this, the analysis of the reader comments is presented.

### 15.3.1 *The News Article*

Table 15.4 provides a summary of the social and cognitive genre elements of the news report section of the web genre.

#### 15.3.1.1 Context

The content of each news article relates to recent events or issues, the nature of which influences the selection of textual resources (cognitive genres) that each article employs and other issues of writer stance. In most cases, the news article was primarily concerned with a recent event involving human interactants. An exception was the article about the banning of trans fats in California (Text 4). Because this involved a chemical substance, the article included a brief explanation of the process of manufacturing trans fats.

#### 15.3.1.2 Epistemology

In relation to epistemology, views about the nature of knowledge and its validation, the news articles take a “realist” view the world and of the events that they report and aim to validate their reports through ostensible displays of objectivity. The writers attempt to objectivise their reporting of real-world events or issues through a multiple perspectives approach to their content. Generally the “multiple perspectives approach” to validation of the content knowledge of the articles is attempted by:

- the reporting of multiple views of witnesses or participants (in the case of events) and multiple viewpoints (in the case of controversies); and,
- the use of multiple rhetorical purposes in the news-reporting texts – realised by the staging of textual resources (see the three Move structure in the content schema section following). Typically, the different rhetorical “angles” of the news story involve a recount of an event, followed (and enriched) by background information about the event which is then followed by a discussion or evaluation of the event including different perspectives.

#### 15.3.1.3 Writer Stance

Writer stance is connected closely to the objectivist aims of the articles and the most frequently used linguistic devices the direct or indirect reporting of participants accounts and some use of hedging devices (examples of cautious language).

Table 15.4 News report sections: social and cognitive genre elements

Article title	Context	Epistemology	Writer stance	Content schema	Cognitive genre
1. Teaching creationism in schools	Controversies about competing theories of origin	Objectivity through multiple perspectives approach	Reporting forms: e.g. “critics say/stress”, “(subject) said”	Event, evaluation	Brief recount followed by discussion
2. Israel/Hezbollah – Prisoner Exchange (July 16, 2008)	Wider historical Israel / Arab conflict	Objectivity through multiple perspectives approach	Hedges: “appeared to be”; reporting forms	Event, background, evaluation	Larger Recount containing a smaller embedded recount
3. Obama’s speech in Berlin	US election & current US/European relations	Objectivity through multiple perspectives approach	Reporting forms: “he said/called”; hedges: “seems to”	Event, evaluation	Recount followed by discussion
4. California bans trans fats	Obesity crisis	Objectivity through multiple perspectives approach	Reporting forms: “said, argued, asserted, according to”	Event, background, evaluation	Recount, explanation, discussion (with a small embedded recount)
5. Goodbye, Starbucks. Hello, coffee.	Starbucks closes of 61 of its 84 outlets in Australia	Mock objectivity through contrastive evaluation	Self mention “I”; attitude markers – “fair enough”; engagement markers “you”	Event, evaluation	Discussion
6. Radovan Karadzic stands defiant before Hague war crimes tribunal	The Bosnian Serb leader’s first indictment before The Hague tribunal	Objectivity through multiple perspectives approach	Reporting forms: “saying, refused, declared, accused, denied, said”	Event	Recount
7. Aleksandr Solzhenitsyn dies at 89	The death of the Russian writer Aleksandr Solzhenitsyn	Objectivity through detailed historical recounts	Attitude markers “stubborn, combative”; reporting forms; “said, described, recalled”	Event, background, evaluation, further Background	Recount, discussion followed by two long historical recounts

Table 15.4 (continued)

Article title	Context	Epistemology	Writer stance	Content schema	Cognitive genre
8. Russia: Georgia can “forget regaining provinces”	The war between Russia and Georgia over South Osettia	Objectivity through multiple perspectives approach	Reporting forms “said, signalled, accused, urged”; hedges “it is not clear whether . . .”	Event	Recount
9. US, Iraqi Negotiators agree on 2011 withdrawal	Negotiations between US and Iraqi governments concerning the eventual withdrawal of US troops from Iraq	Objectivity through multiple perspectives approach	Reporting forms “have said, have agreed, pledged,”	Event, background	Recount followed by recount of earlier events
10. Food riots as Zimbabwe aid ban continues	The continued ban by the Zimbabwe government on the distribution of food by aid organisations	Objectivity through multiple perspectives	Reporting forms “was announced, say, declining, said”	Event, evaluation, background	Recount, discussion, explanation

### 15.3.1.4 Content Schema

The staging of content within news article section of the sample of texts appears to follow a content schema that involves one obligatory and two optional moves. They are:

Move 1 Event (obligatory)

Move 2 Background information about the event (optional)

Move 3 Evaluation of the event (optional but common)

Move 1, termed an “Event”, is the only obligatory move that occurs in all ten texts. This is the report of something that has happened; this could be an actual series of events with human interactants or an issue that has arisen in a certain context at a certain time. Move 2, termed “Background information about the event” is an optional move that occurred in five of the ten of texts. The purpose of this move is to provide further, detailed information, ostensibly so that the reader can better understand the event in terms of its history, the reasons for its occurrence or its significance or implications. Move 3, termed “Evaluation of the event”, introduces opinions about the event. This may be the opinions of the interactants, such as in Text 2 in relation to the Palestinian terrorist released by the Israelis, or a balanced presentation of contrasting views of the event by the actual writer, such as occurs in the evaluation of Obama’s speech in Berlin (Text 3). Move 3 occurs in seven of the ten texts.

### 15.3.1.5 Cognitive Genres

The cognitive genres that are employed in each article relate to the general rhetorical aims that the texts need to draw upon. The cognitive genres, segments of text relating to a single rhetorical purpose, usually coincide with the articles content moves. Move 1, the “Event” is usually realised by Recount cognitive genre as this Move is usually concerned with presenting information that is sequential or chronological. However, the realisation of Move 2, “Background” depends on the content knowledge communicated by the article. For example, in the case of Text 2, the Background (Move 2) was realised by a brief historical Recount describing the killings carried out by a terrorist, Kantar, which led to his imprisonment by the Israelis. However, in the case of Text 4, “Background” was realised by an Explanation cognitive genre that outlined the process of the manufacture of trans fats, their use in processed foods and the role that they play in contributing to coronary heart disease. Where Move 3 (Evaluation of the event) occurs, it is realised by Discussion cognitive genre with an attempt offer an evaluation of the event or issue (reported in Move 1) by juxtaposing contrasting views.

## 15.3.2 Reader Comments

The comment section of the web genre involves short pieces of text (with an average length of 95 words over the whole sample) usually aiming to make a single point

**Table 15.5** Quantitative data about the comments posted after news articles

Text	Total number of comments of comments	Comments that positively evaluate the article content	Comments that negatively evaluate the article content	Comments about views of other participants (or other writer purpose)	Average number of words per comment
1	64	23	36	5	161
2	21	8	12	1	42
3	120	53	48	19	47
4	48	20	22	6	59
5	9	3	2	4	66
6	44	31	12	1	69
7	89	50	18	21	97
8	322	55	45	223	106
9	271	17	47	207	100
10	8	4	0	4	138

or comment. The comments express a viewpoint about some aspect of the news article, or about the views of another participant in the blog, or they seek to make some other point. The comments are usually in the form of informal written communication, like a brief email message. Table 15.5 provides a quantitative summary of the numbers of comments, the general purposes of the comments (in relation to the news article) and average number of words in the comments attached to each article.

The length of comments is sometimes constrained by limits on the numbers of characters or words that can be used in each message for some of the articles; these constraints are usually established in the online form by means of which the comment must be posted into the electronic forum. Therefore, because of the requirement for brevity, the writer can provide little background information and can indulge in only minimal rhetorical organisation of their argument because of the word limit.

The context of the comments is loosely established by the topic of the article; however, in some of the longer series of comments, the topic of the article merely provides a background for the discussion of a different issue. For example, in Text 9, the news article topic of the negotiation of the withdrawal of American troops from Iraq mostly provides a forum for the discussion of the merits of candidates in the (then) forthcoming American presidential election, usually in relation to their respective policies relating to Iraq. While news articles aimed at appearing to be objective, the reader comments are, for the most part, subjective. Self-mention and verbs signalling opinion reflect these personal and subjective elements of writer stance, e.g. “I agree ...; I think ...; I am sad to hear ...” or “If you ...”. In relation to their internal organisation, the reader comments often follow the simple pattern of:

- a statement of opinion or information; followed by,
- justification of the statement.

Sometimes the statement is made in the form of a rhetorical question. An example of this pattern is a comment made in relation to Text 2, a text that reports Israel's return of live prisoners to Hezbollah in exchange for the bodies of two of its own servicemen; the reader says:

Are the bodies of dead guerrillas preserved – for want of a better word – for potential prisoner exchanges? So they are pawns in life as well as in death?

Because of their (often extreme) brevity, the reader comments tend not draw upon cognitive genres to structure a segment of text that relates to one type of rhetorical purpose followed by a rhetorical shift into another cognitive genre. Rather they tend to be brief, single point texts, sometimes as part of an interaction where the writer challenges or responds to the comment of another reader. Thus, the reader comments, and particularly those of the latter category, may have more of a dialogic or interactive rather than a monologic quality.

## 15.4 Discussion

The target genre of the participatory news article and its constituent elements in many ways reflect the stages of development of web genres in that it includes the synthesis of an “extant genre” from another print medium (the news article) and a developing, “novel” genre that is an artefact of the online environment in which it occurs (the reader comments). Thus, the news article section tends to be an imported, unaltered text from a physical newspaper, while the reader comment section is evolving because of the capabilities that the online environment affords, including the ability to post comments publicly about an article in the same edition of the publication, comments that can then be read by the same readership as for the article. The interactive, asynchronous discussion that can arise among the those who comment too is an artefact of the online environment.

In the present stage of the development of the genre, the sample generally indicates no interaction between the writer of the news article and those writing the comments that follow, and the online news articles appear to be the same as their physical newspaper editions. An exception was Text 5 where the writer of the article participates in and responds to comments in the blog. In this article, the writer reports the closure of much of the Starbucks operation in Australia, but at the same time expresses a view in favour of the downsizing, a view which leads to the reader discussion in which he subsequently participates (defending his position). However, it must be said that this article deviated somewhat from the other texts of the sample in that the writer of the article dispensed with the news article convention of ostensible objectivity. (While discussing the advantages and disadvantages of Starbucks stores in the article, the writer's overall view was negative toward the stores and he generally welcomed their closure.)

In applying the social genre/cognitive genre model to the analysis of this particular genre, both the news article and reader comments sections are found to have

a range of conventionalised, genre-identifying features relating both to the social genre and cognitive genre models which provided the basis for the analysis.

In relation to the news article, the social genre elements identified in the study are: a realist view of the world achieved through the reporting of multiple eyewitness views of an event or multiple perspectives on an issue involving the use of a regularised schema for staging the content of the article. In relation to the content schema, the study proposes a provisional move structure in terms of: recounting an Event (compulsory), Background information about the event (optional) and Evaluation of the event (optional but common). The multiple perspectives approach of the moves of the schema tended to be mirrored in their use of particular textual (cognitive genre) resources, with the Event section commonly realised by a Recount cognitive genre, the Background using Recount or Explanation cognitive genre (depending on the subject matter of the news article) and the Evaluation using the Discussion cognitive genre. Features of metadiscourse language are mainly that of reporting language (when quoting eyewitnesses or commentators) and hedging or cautious language when presenting information or claims that could be deemed to be contestable.

On the other hand, the readers' comments section of the genre contains a series of immediate, seemingly spontaneous (and probably hastily written) personal responses to the article. Comments are often brief, limited in size by the character limit in the online form by means of which they are sometimes posted and, therefore, they are unable to include much rhetorical organisation. Unlike the news article section of the web genre, which uses devices to validate the knowledge that they report through displays of objectivity (multiple participant or commentator views and multiple rhetorical angles), the comments tend to be single-view, subjective statements broadly falling into the categories of being in favour of or against the content of the article. A third broad category of comments are those that are responses to other participants in the comment section, and it is these comments that introduce a further interactive dimension to the comment section. Rather than addressing a view to the general reading audience, the third category of comment directly addresses another comment poster by agreeing with, opposing, correcting or commenting in some other way on that person's posting. In some of the larger chains of postings, this was the largest category of comment. The general organisational pattern that is often observable is a statement of a view followed by some justification or argument that supports the comment.

The findings in relation to this genre analysis of the participatory news article can only be considered as provisional and indicative rather than representative or generalisable since the study was a small-scale, exploratory investigation. The sample was small (10 texts, 108,693 words), and multiple rater analysis of the texts and moderation of findings were not logistically possible. However, this study does raise a number of important issues relating to web genre identification and classification, issues that can broadly be grouped within two areas: first, frameworks for the analysis and classification of web genres and, secondly, the means by which their analysis and classification may be achieved.

## 15.5 Conclusion

In relation to frameworks for the analysis of web genres, this study illustrates the fact that web genres, like genres in other print media, clearly involve a range of knowledge types relating to both text and discourse. Therefore, whatever analytical frameworks are employed need to have the capacity to provide a nuanced analysis of the characteristics of the manifestations of a web genre as “the linguistic trace of a discourse process” [85, p. 163]. Thus, if “text” is to be considered in the analysis, the analytical framework would need to include systematic ways of considering the linguistic features and rhetorical organisation of text, drawing upon appropriate theories of text. If discourse (the social and cognitive operations that surround and lend meaning to the text) is also to be part of the framework for analysis, then salient knowledge related to areas of knowledge, such as context, epistemology, writer stance and the nature of human cognition (as it relates to knowledge categorisation), would also need to be included among the elements accounted for by such a framework. Thus, any attempt at a systematic, nuanced description of web genres requires a theory (of genre) that is sufficiently powerful to account for the full range of knowledge elements that the genre draws upon. Furthermore, it is important that such a theoretical framework has the capacity to accommodate the multiple views and multiple interpretations of its reader audience community since it is the readers who ultimately ratify an example text as a member of a particular web genre category. In this way, a sufficiently powerful theory of genre should have the capacity to overcome the problem previously identified in relation to web genres that different communities of users may interpret and understand the same genre in different ways [87].

The second implication of this study in relation to web genres is methodological in terms of how genre are analysed and the relative roles of human, inferential analysis and computer-mediated analysis, such as by the use of corpus software. If genres are to be operationalized in terms of multiple types of integrated knowledge that relate to both text and discourse (as was the case in this study), such an approach to genre has implications for any notion of the use of computer-mediated analysis as the primary or first-instance method identification and classification of texts in terms of genre categories. For example, if genre analysis is to encompass both text and discourse, the findings of this study suggest that this will be achieved, in the first instance, by qualitative, (human) rater analysis rather than through quantitative, computer-mediated methods, such as through the use corpus software in the creation of wordlists and concordancing:

the computer can only cope with the material products of what people do when they use language. It can only trace the textual processes whereby meaning is achieved: it cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted [84, pp. 6–7].

In delimiting the role of corpus methods in such analyses, Widdowson says that corpus linguistics provides us with the description of the text, not discourse. Although textual findings may well alert us to possible discourse significance and send us back to their contextual source, such significance cannot be read off from the data [84, p. 7].

Thus, if genres are, as Biber suggests “defined and distinguished on the basis of systematic non-linguistic criteria, and they are valid in those terms” [9, p. 39], it would seem to be incontrovertible they are discursive entities and discourse is at their core. Consequently, it would seem that a valid theory of genre requires a systematic approach to uncovering the discursive elements of the genre, which will inevitably involve qualitative rater analysis. However, it is quite possible that wordlists and concordance searches may also be employed in a supportive way to provide empirical evidence for the role of theorised textual or discursive features (for examples, see [16, 18]).

Genres as categories of texts have long been examined in other print media; however, the enterprise of examining web genres is more recent and may require consideration of both monologic text and more overtly dialogic types of interactions that, as Young [88] says, fall somewhere, between spoken and written discourse (such as the reader comments examined in this study). Thus, through the medium of the web, new genre combinations and hybrid genres are emerging, such as the target genre examined in this study. However, if ways are to be found to achieve a comprehensive and inclusive “faceted-analysis approach” [25, p. 4] to both their identification and classification, any proposal will need to involve a comprehensive, theoretical framework that can account for the different types of knowledge, knowledge that relates to both text and discourse, which intermeshes in the creation of so complex an entity.

## References

1. Adam, J.-M. 1985. Quels types de textes? *Le Français dans le Monde* 192:39–43.
2. Adam, J.-M. 1992. *Les textes – types et prototypes*. Paris: Nathan.
3. Barsalou, L.W. 1983. Ad hoc categories. *Memory and Cognition* 11:211–227.
4. Bhatia, V.K. 1993. *Analysing genre – language use in professional settings*. London: Longman.
5. Bhatia, V.K. 1998. Generic conflicts in academic discourse. In *Genre studies in English for academic purposes*, eds. I. Fortanet, S. Posteguillo, J.C. Palmer, and J.F. Coll, 15–28. Castello de la Plana: Publicacions de al Universitat Jaume.
6. Bhatia, V.K. 2002. Applied genre analysis – analytical advances and pedagogic procedures. In *Genre in the classroom – multiple perspectives*, ed. A. Johns, 279–283. Mahwah, NJ: Erlbaum.
7. Bhatia, V.K. 2004. *Worlds of written discourse – a genre based view*. London: Continuum.
8. Biber, D. 1988. *Variation across speech and writing*. Cambridge, MA: Cambridge University Press.
9. Biber, D. 1989. A Typology of English texts. *Linguistics* 27:3–43.
10. Biber, D., E. Csomay, K. Jones, and C. Keck. 2007. Introduction to the identification and analysis of vocabulary-based discourse units. In *Discourse on the move – using corpus analysis to describe discourse structure studies in corpus linguistics*, eds. D. Biber, U. Connor, T.A. Upton, vol. 28, 173. Amsterdam: John Benjamins.
11. Bloor, M. 1998. Variations in the methods sections of research articles across disciplines – the case of fast and slow text. In *Issues in EAP writing research and instruction*, ed. P. Thompson, 84–106. Reading UK: CALS, The University of Reading.
12. Borg, E. 2003. Key concepts in ELT – discourse community. *ELT Journal* 57:398–400.

13. Bruce, I. 2003. Cognitive genre prototype modelling and its implications for the teaching of academic writing to learners of English as a second language. Unpublished doctoral dissertation, University of Waikato, Hamilton, New Zealand.
14. Bruce, I. 2005. Syllabus design for general EAP courses – a cognitive approach. *Journal of English for Academic Purposes* 4:239–256.
15. Bruce, I. 2007. Defining academic genres – an approach for writing course design. In *Proceedings of the 2005 Joint BALEAP/SATEFL Conference: New Approaches to Materials Development for Language Learning*, ed. O. Alexander, 103–116. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang.
16. Bruce, I. 2008a. Cognitive genre structures in methods sections of research articles – a corpus study. *Journal of English for Academic Purposes* 7:39–55.
17. Bruce, I. 2008b. *Academic writing and genre*. London: Continuum.
18. Bruce, I. 2009. Results sections in sociology and organic chemistry articles – a genre analysis. *English for Specific Purposes* 28:105–124.
19. Carrell, P.L. 1981. Culture-specific schemata in L2 comprehension. In *Selected Papers from the Ninth Illinois TESOL/BE Annual Convention, First Midwest TESOL Conference*, eds. R. Orem and J. Haskell, 123–132. Chicago, IL: TESOL/BE.
20. Carrell, P.L. 1988. Content and formal schemata in ESL reading. *TESOL Quarterly* 21:461–481.
21. Celce-Murcia, M., and Z. Dornyei. 1995. Communicative competence – a pedagogically motivated model with content specifications. *Issues in Applied Linguistics* 6:5–35.
22. Connor, U., and A. Mauranen. 1999. Linguistic analysis of grant proposals – European Union Research Grants. *English for Specific Purposes* 18:47–62.
23. Council of Europe. 2001. *Common European framework of reference for languages – learning teaching assessment*. Cambridge, MA: Cambridge University Press.
24. Crombie, W.H. 1985. *Process and relation in discourse and language learning*. Oxford: Oxford University Press.
25. Crowston, K., and B.H. Kwasnik. 2004. A framework for creating a faceted classification for genres – addressing issues of multidimensionality. In *Proceedings of the 37th Hawaii International Conference on System Sciences*. Hawaii.
26. Dudley-Evans, A. 1986. Genre analysis – an investigation of the introductions and discourse sections of MSc dissertations. In *Talking about text – discourse analysis monographs No. 13*, ed. M. Coulthard, 128–145. Birmingham: English Language Research, University of Birmingham.
27. Dudley-Evans, T. 1989. An outline of the value of genre analysis in LSP work. In *Special language – From humans thinking to thinking machines*, eds. C. Laurén and M. Nordman, 72–79. Clevedon: Multilingual Matters.
28. Dudley-Evans, T. 1993. Variation in communication patterns between discourse communities – the case of highway engineering and plant biology. In *Language learning and success – studying through English*, ed. G.M. Blue, 141–147. London: Macmillan.
29. Dudley-Evans, T. 1994. Genre analysis – an approach to text analysis for ESP. In *Advances in written text analysis*, ed. M. Coulthard, 219–228. London: Routledge.
30. Eggins, S. 1994. *An introduction to systemic functional linguistics*. London: Pinter.
31. Feez, S. 2002. Heritage and innovation in second language education. In ed. A. Johns, *Genre in the classroom – multiple perspective*, 43–69. Mahwah, NJ: Lawrence Erlbaum.
32. Fowler, A. 1982. *Kinds of literature – an introduction to the theory of genres and modes*. Cambridge, MA: Harvard University Press.
33. Giddens, A. 1984. *The constitution of society – outline of the theory of structuration*. Berkeley, CA: University of California Press.
34. Grabe, W. 2002. Narrative and expository macro-genres. In *Genre in the classroom – multiple perspectives*, ed. A. Johns, 49–267. Mahwah, NJ: Lawrence Erlbaum.
35. Halliday, M.A.K. 1978. *Language as a social semiotic*. London: Edward Arnold.

36. Halliday, M.A.K., and R. Hasan. 1989. *Language, context and text – aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.
37. Hasan, R. 1989. The identify of a text. In *Language, text and context*, M.A.K. Halliday and R. Hasan, 97–118. Mahwah, NJ: Oxford University Press (Original work published in 1985).
38. Hearst, M.A. 1997. TextTiling – segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23:33–64.
39. Herring, S., L. Scheidt, S. Bonus, and E. Wright. 2004. Bridging the gap – a genre analysis of weblogs (DDGDD04). In *Proceedings of the Annual Hawaii International Conference on System Sciences, (Conf 37)*, 101. Hawaii.
40. Hoey, M. 1979. *Signalling in discourse. Discourse Analysis Monograph No. 6*. Birmingham: English Language Research, University of Birmingham.
41. Hoey, M. 1983. *On the surface of discourse*. London: George Allen and Unwin.
42. Hoey, M. 1994. Signalling in discourse – a functional analysis of a common discourse pattern in written and spoken English. In *Advances in written text analysis*, ed. M. Coulthard, 26–45. London: Routledge.
43. Hoey, M. 2001. *Textual interaction – an introduction to written discourse analysis*. London: Routledge.
44. Hopkins, A., and Dudley-Evans, T. 1988. A genre-based investigation of the discussion section in articles and dissertations. *English for Specific Purposes* 7:113–122.
45. Hyland, K. 2005. *Metadiscourse – exploring interaction in writing*. London: Continuum.
46. Johns, A.M. 1997. *Text role and context – developing academic literacies*. Cambridge, MA: Cambridge University Press.
47. Johns, A.M. 2001. The future is with us – Preparing diverse students for the challenges of university texts and cultures. In *Academic writing in context – implications and applications*, ed. M.A. Hewings, 30–42. Birmingham: University of Birmingham Press.
48. Johnson, M. 1987. *The body in the mind – the bodily basis of meaning imagination and reason*. Chicago, IL: University of Chicago Press.
49. Lackstrom, J., L. Selinker, and L. Trimble, L. 1973. Technical rhetorical principles and grammatical choice. *TESOL Quarterly* 15:51–57.
50. Lakoff, G. 1987. *Women fire and dangerous things – what categories reveal about the mind*. Chicago, IL: Chicago University Press.
51. Lave, J., and E. Wenger. 1991. *Situated learning – legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.
52. Lea, M.R., and B. Street. 1998. Student writing in higher education – an academic literacies approach. *Studies in Higher Education* 23:157–172.
53. Martin, J.R. 1984. Language register and genre. In *Children's writing – reader*, ed. F. Christie, 21–30. Geelong, Australia: Deakin University Press.
54. Martin, J.R. 1986. Intervening in the process of writing development. In *Writing to mean – teaching genres across the curriculum*, C. Painter, J.R. Martin, Applied Linguistics Association of Australia (Occasional Paper 9, 11–43). Bundoora: Applied Linguistics Association.
55. Martin, J.R. 1992. *English text – system and structure*. Amsterdam: John Benjamins.
56. Martin, J.R. 1994. Macro-genres – the ecology of the page. *Network* 21:29–52.
57. Martin, J.R. 1995. Text and clause – fractal resonance. *Text* 15:5–42.
58. Martin, J.R. 1997. Analysing genre – functional parameters. In *Genre and institutions – social processes in the workplace and school*, eds. F. Christie and J. Martin, 3–39. London: Cassell.
59. Martin, J.R. 2000. Design and practice – enacting functional linguistics. *Annual Review of Applied Linguistics* 20:116–126.
60. Miller, C.R. 1984. Genre as social action. *Quarterly Journal of Speech* 70:151–167.
61. Murphy, G.L., and D.L. Medin. 1985. The role of theories of conceptual coherence. *Psychological Review* 92:289–316.
62. Nwogu, K.N. 1991. Structure of science popularisations – a genreanalysis approach to the schema of popularised medical texts. *English for Specific Purposes* 10:111–123.

63. Oller, J.W. 1995. Adding abstract to formal and content schemata – results of recent work in Peircean semiotics. *Applied Linguistics* 16:273–306.
64. Paltridge, B.R. 1993. A challenge to the current concept of genre: Writing up research. Unpublished doctoral thesis, University of Waikato, Hamilton.
65. Paltridge, B. 1997. Genre frames and writing in research settings. Amsterdam: John Benjamins.
66. Paltridge, B. 2002. Genre text type and the English for Academic Purposes (EAP) Classroom. In *Genre in the classroom – multiple perspectives*, ed. A. Johns, 73–90. Mahwah, NJ: Lawrence Erlbaum.
67. Pilegaard, M., and F. Frandsen. 1996. Text type. In *Handbook of pragmatics*, eds. J. Verschueren, J.-O. Ostaman, J. Blommaert, and C.C. Bulcaen, 1–13. Amsterdam: John Benjamins.
68. Quinn, J. 1993. A taxonomy of text types for use in curriculum design. *EA Journal* 11(2):33–46.
69. Rosch, E. 1978. Principles of categorisation. In *Cognition and categorization*, eds. E. Rosch, and B.B. Lloyd, 27–47. Hillsdale, NJ: Erlbaum.
70. Sanford, A.J., and S.C. Garrod. 1981. *Understanding written language*. Chichester: Wiley.
71. Santini, M. 2005. Automatic text analysis – Gradations of text types in web pages. In *Proceedings of the 10th European Summer School in Logic, Language and Information Student Session*, ed. J. Gervain, 276–285. Edinburgh, UK.
72. Shepherd, M., and C. Watters. 1998. The evolution of cybergenres. *Proceedings of the Hawaii International Conference on System Sciences* 31(2):87–109.
73. Silva, T. 1990. Second language composition instruction – developments, issues and directions in ESL. In *Second language writing – research insights for the classroom*, ed. B. Kroll, 11–23. Cambridge, UK: Cambridge University Press.
74. Swales, J.M. 1981. Aspects of article introductions (Aston ESP Research Rep. No. 1). The University of Aston, Language Studies Unit, Birmingham.
75. Swales, J.M. 1988. Discourse communities genres and English as an international language. *World Englishes*, 7:211–220.
76. Swales, J.M. 1990. *Genre analysis – English in academic and research settings*. Cambridge, MA: Cambridge University Press.
77. Swales, J.M. 1998. *Other floors other voices – a textography of a small university building*. Mahwah, NJ: Lawrence Erlbaum.
78. Swales, J.M. 2004. *Research genres – exploration and applications*. Cambridge, MA: Cambridge University Press.
79. Van Dijk, T.A. 1980. *Macrostructures – an interdisciplinary study of global structures in discourse, interaction and cognition*. Hillsdale, NJ: Lawrence Erlbaum.
80. Ventola, E. 1985. Orientation to social semiotics in foreign language teaching. *Applied Linguistics* 5:275–286.
81. Virtanen, J. 1992. Issues of text typology – Narrative – a ‘basic’ type of text? *Text* 12:292–310.
82. Wenger, E. 1998. *Communities of practice*. Cambridge, MA: Cambridge University Press.
83. Werlich, E. 1976. *A text grammar of English*. Heidelberg: Quelle and Meyer.
84. Widdowson, H.G. 2000. On the limitations of linguistics applied. *Applied Linguistics* 21:3–25.
85. Widdowson, H.G. 2004. *Text, context and pretext*. Oxford: Blackwell.
86. Yates, J., and W. Orlikowski. 1992. Genres of organizational communication – a structural approach to studying communication and media. *Academic of Management Review* 17:299–326.
87. Yoshioka, T., J. Yates, and W. Orlikowski. 2002. Community-based interpretative schemes – exploring the use of cyter meetings within a global organization. In *Proceedings of the 35th Hawaii International Conference on System Sciences*. Hawaii.
88. Young, A. 2006. *Teaching writing across the curriculum (Prentice Hall resources for writing)*. Upper Saddle River, NJ: Pearson Prentice Hall.



**Part VI**  
**Prospect**

# Chapter 16

## Any Land in Sight?

Marina Santini, Serge Sharoff, and Alexander Mehler

What conclusions can we draw from the 15 studies included in this book? Is there hope of sorting out the complex issues of genre on the web? Is there “any land in sight”? We think so.

As emphasised in the introduction of this book, genre is a multifarious concept that lends itself to many interpretations and uses. For this reason, we included as many approaches and different views as possible.<sup>1</sup> We believe that the plurality and diversity of visions foster cross-fertilisation of ideas and that inter- and transdisciplinarity are the most productive approaches to increasing our understanding of this important concept.

### 16.1 Web Genre Benchmarks

Plurality, diversity, cross-fertilization, inter- and transdisciplinarity are key points for our future projects, as well. The book contains the gist of 15 years of empirical experience with genre and shows the way to the next generation of web genre research. In our view, the necessary next step is the construction of large and shared web genre benchmarks, i.e. web genre reference corpora that enable the objective assessment of effectiveness of various empirical and computational approaches. As empiricists, we need to test our methods and ideas. In order to test them, we need some kind of “reference” against which our different methods or ideas can be measured. For this reason, we propose building a web genre benchmark spawned by a wide and comprehensive discussion of genres on the web. Without such a benchmark, it is hard to evaluate progress.

---

M. Santini (✉)  
KYH, Stockholm, Sweden  
e-mail: marinasantini.ms@gmail.com

<sup>1</sup> Additional experiments are presented in the Special Issue on Genre of the *Journal for Language Technology and Computational Linguistics*, JLCL 24(1).

For instance, how does the list of 522 genre labels collected by Crowston et al. (Chapter 4) compare against the set of eight genres used in the KI-04 corpus used by Stein et al. (Chapter 8)? Is the 96% accuracy reported by Kim and Ross (Chapter 6) better than the 86% accuracy obtained by Sharoff (Chapter 7)? These are the questions for which we need to find answers in the upcoming phase.

For several reasons, the construction of genre reference corpora is a challenging endeavour. The three main problems that need to be discussed concern (1) the set of genre labels, (2) the process of annotation of source texts and (3) representativeness.

### ***16.1.1 Genre Labels***

One main challenge in the construction of web genre benchmarks is to convey the variety of genre classes that have been used so far, without cutting out genre labels that can be potentially useful for other information needs or research fields. Given that there is no lasting solution to the problem of diversity of genre labels, our plan is to produce corpora with stand-off annotation according to a fairly fine-grained genre palette and a set of mappings to other classification schemes. The exact composition of the source palette will have to be determined as a result of future discussion and research, but the starting point for it will be the set of labels listed in the *WebGenreWiki*.<sup>2</sup> The palette in the wiki results from an agreement between several groups of genre researchers, and, by design, it is a flat list of genre classes with reasonably fine granularity. Most of the labels used in other genre palettes can be converted to this scheme without considerable ambiguity. Naturally, this genre palette will be enhanced and refined along the way.

### ***16.1.2 Annotation***

Previous experiments have shown how assigning one single genre per document (whatever the unit of analysis) is quite artificial. The chapters in this book have well illustrated this difficulty and reported on how existing genre collections have been annotated with a variety of approaches, following differing taxonomies and nomenclatures. As genre is a multifaceted concept, influenced by elements such as perception, terminological *prestige*, membership in certain communities, and the fluidity of the language itself, certainly the next step in genre annotation is to find a way to accommodate several genre labels per document, by working out techniques to establish sensible labelling thresholds. Reliable manual annotation paired with the availability of an unlimited amount of unannotated documents on the web can be leveraged by semi-supervised classification methods that will alleviate the burden of any future annotation work.

---

<sup>2</sup> <http://purl.org/net/webgenres>

### ***16.1.3 Representativeness***

Generally speaking, corpora are designed as samples for studying a much larger whole. With respect to genres two questions naturally arise:

- Is a given corpus representative for a large number of genres?
- Is a given genre adequately represented in a given corpus?

The first question is important, as attempts to create a very big corpus from a small number of sources normally restrict the diversity of genres. Our reference corpora will be produced from a diverse collection of webpages, as already experienced for the I-EN.<sup>3</sup> For a cross-cultural concept like genre, it also makes sense to create reference web genre corpora for multiple languages.

The second question is much more challenging, as a subcorpus defined for a given genre is normally much smaller and has less variation. The BNC, for instance, is representative for a variety of genres including research articles. However, as for the genre of research articles itself, its texts were mostly taken from the *Journal of Gastroenterology and Hepatology*, so they cannot reflect the variety within this genre. Building on this experience, one of our goals is to create genre reference corpora that aim at a better representation of each genre.

## **16.2 Work Plan**

The major research efforts will be to:

1. Propose a characterisation of genre suitable for digital environments and empirical approaches shared by a number of genre experts working in different disciplines and following different schools of thought.
2. Define the criteria for the construction of genre benchmarks and draw up annotation guidelines.
3. Create several genre benchmarks in several languages, that are differing in size, corpus composition, and annotation methods, and that can be updated over time with emerging genres.

We conjecture that the construction of a shared web genre reference corpus would be the most solid legacy to future genre research.

### ***16.2.1 Benefits***

The creation of multilingual web genre benchmarks will:

1. Help researchers avoid investing large amounts of time and money coming up with proprietary and incompatible solutions instead of working with shared resources and common standards.

---

<sup>3</sup> See <http://corpus.leeds.ac.uk/internet.html>

2. Provide a common ground for genre-related research, spanning from information retrieval to discourse analysis.
3. Provide material to be used as training data for machine learning approaches for tasks such as automatic web genre identification, focused crawling, spam detection and web mining.
4. Allow more sophisticated computational genre modelling that builds upon genre relations at different units of analysis.

# Index

## A

- Academic web space, 24, 213, 255, 258, 260–263, 266, 269, 272–273
- Accuracy, 49, 90–92, 103–113, 116–118, 142–145, 155–158, 161, 163–164, 179–183, 192, 195, 213–214, 226–228, 232–233, 306, 352
- Adaptive learning, 88
- AGI, *see* Automatic genre identification
- Amateur Flash exchange, 278
- Animation, 16, 168, 279–281, 284–286, 288, 290–292, 294, 296–300
- Annotation methods, 353
- Automatic genre identification (AGI), 18, 24, 87–91, 93–95, 98–100, 104, 108, 118–119, 167

## B

- Bag-of-features model, 20, 22, 238
- Bag of words, 18, 21, 131, 138, 170, 178
- Baseline, 111, 117–119, 142, 196
- Bayes' theorem, 100–101
- Benchmarks, 7, 18, 73, 87, 227, 351–353
- Bias of a learning algorithm, 175
- Blind Accessibility Tool, 172–173
- Blog, 277, 300
  - expert, 319–320
  - group, 64, 305, 315
  - personal (diary), 24, 39, 111–114, 117, 120, 303–304, 317, 320
  - political, 304
  - thematic, 24, 303–305, 320
  - type, 304–305, 317, 319–320
- Bottom-up approach, 7, 23, 74

## C

- Canonical discriminant analysis, 205
- Categorization, 20–22, 24, 34–36, 38, 49, 58, 73, 153, 171–172, 177, 192–195, 198, 214, 259–262, 272, 323, 327–328, 330–332, 344
  - genre, 259–260
- Classification scheme, 13, 50, 91, 119, 150–152, 155, 352
- Classification by structure and content, 228, 230–233
- Classifier, 8, 13, 19, 21–22, 49–51, 57, 61, 88, 91–94, 99–100, 106, 109–110, 130–131, 137–139, 142–145, 156–157, 159–164, 167, 171, 173–175, 177, 179–183, 186, 191, 195–196, 215, 227, 229, 231, 304
- Cleaning, 17–18, 153, 158, 218, 221
- Cluster analysis, 282, 288, 290, 305, 315–316, 320
- Cognitive genres, *see* Genre
- Comment, 37, 63–65, 70, 76, 80, 98, 152, 226, 283, 303, 305, 310, 323, 333–334, 336–337, 340–343, 345
- Community
  - spoken, 327
  - written, 327
- Complex network theory, 237
- Computational models, 6, 9, 18, 22
- Conventions, *see* Genre
- Co-occurrence pattern, 307
- Copyright, 16, 52, 87, 160, 178, 197
- Corpus composition, 159, 353
- Crawl, crawling, 15–16, 93, 149, 151, 159–160, 171, 213, 217–219, 221, 256, 270–271, 273, 282–283, 354

- Creative Commons, 16, 160  
 Cross-testing, cross-test, 18, 87, 175
- D**
- Data, 4, 14, 16, 21, 24, 37–39, 41–42, 48, 50, 58, 60, 62, 66, 69, 72, 75–77, 80, 92, 130, 136–140, 142–145, 150, 159–160, 163, 170, 175–177, 180–181, 184–186, 193, 195–198, 202, 206, 212–218, 220–221, 225, 227, 229, 231, 237–238, 244, 248–250, 256–258, 260, 266, 269, 272, 279–280, 282–283, 290, 300, 305–307, 331–332, 341, 344, 354
- Deep Web, 16
- Dice coefficient, 116–117
- Digital media, 3, 23, 278–280
- Discourse, 6–7, 13, 19, 33, 35, 45, 76, 78, 82, 278, 308–312, 323–325, 327–334, 336, 343–345  
 community, 6, 13, 76, 278, 324–325, 327, 330  
 type, 324
- Document structure, 11–12, 20–22, 130–131, 133, 140, 187, 237–239, 244, 248–249
- Domain  
 sub, 216, 218–220  
 top level, 213, 216
- DOM-tree, 21, 173, 215, 239, 244, 249
- E**
- Emerging genre, 41, 74, 87, 150, 154, 297, 300, 353
- Empirical studies, 8, 22
- Evaluation, 5, 24, 49, 65, 90, 104, 111, 119, 121, 177–179, 183, 193–197, 202, 205–206, 214–215, 227, 248, 250, 295, 336–340, 343
- F**
- Factor analysis, 100, 279, 305–306, 315–316, 320
- Features  
 bag-of-words (BOW), 18, 21, 131, 138, 170  
 byte n-grams, 90, 98–99  
 character n-grams, 18, 109  
 content-based, 213  
 content-related, 211  
 co-occurrence, 306–307, 314  
 cultural reference, 282, 285–286, 288, 294–299  
 facets, 98, 100–102, 106, 109–110, 117–118, 125, 193  
 formal, 285, 300  
 function words, 15, 18, 109, 170, 194, 304  
 genre, 282, 284–286, 290–293, 297  
 genre-specific words, 109  
 harmonic descriptors, 99  
 HTML, 95, 109, 125, 140, 153, 156, 163, 170, 176–177, 181–182, 192–194, 197, 219, 225, 238, 280  
 lexico-grammatical, 326  
 linguistic, 304–305, 307–308, 325, 328–329, 334  
 n-characters n-grams, 99  
 non-linguistic, 328  
 POS trigrams, tri-grams, 18, 99, 109, 156, 158, 164  
 punctuation, 109, 156, 164, 170, 194, 199–200  
 shallow, 98–100  
 structural, 211, 213–214  
 syntactic chunks, 18, 170
- Flash, 16, 24, 37, 168, 217, 277–300
- Functional style, 14, 152, 196, 198–199, 206
- G**
- Generalization capability, 168, 174–176, 178–180, 183, 187
- Genre  
 asynchronous multi-party correspondence, 37  
 bilingual, 123  
 classificatory principle, 3  
 characterization, 3, 6–9, 11, 13, 19, 90, 100, 118–119  
 class, 259–260, 263–268  
 concept, 5–8, 23, 90, 93, 98–99  
 connectivity, 255–256, 262, 269  
 content, 34, 71, 77, 81  
 conventions, 4, 8, 33–35, 37, 40, 45  
 cross-genre link, 261  
 cross-topic link, 255, 257–258, 261, 271  
 culture, 37  
 data bases, 37  
 definition, 3, 8–9, 35, 48, 55, 57–59, 71–72, 79, 88, 90  
 digital, 278  
 dimension, 4–5  
 drift, 270–272  
 ecology, 7  
 emergence, 277–282, 288, 297–300  
 evaluation, 49, 65  
 evolution, 7–8, 13, 277  
 folksonomy, 50  
 function, 34, 52–53

- granularity, 93, 97
- identification, 49, 64, 87
- lens, 5, 82
- mapped/mapping, 111–112
- media, 37
- meso level, 11–12
- micro level, 11–12, 22
- model, modelling, 90, 94, 96, 98–104, 106, 354
- multimedia, 277–280
- municipality, 81
- mutual expectations, 33
- names, naming, 4–6, 9, 13
- nomenclature, 5, 97, 327, 352
- palette, 48, 55–56, 58–64, 88, 90–92, 96–97, 352
- PDF, 79
- perception, 55, 88
- predictability, 9
- predictivity, 4
- purpose, 3–4, 70–71
- recognition, 88
- recognizability, 47, 49–50
- regularities, 4
- regulations, 161
- robustness, 87, 90, 104
- social, 6–7, 25, 88, 97, 100, 323, 325, 329–330, 333–334, 336–337
- social tagging, 272
- software program, 260–262
- sub, 19–20, 22, 59, 61, 88, 91, 212–213
- substance, 53–54
- super, 19, 88, 211–216, 218, 226, 230, 232–233
- rhetorical, 97–98, 100, 102
- taxonomy, 5, 23, 50, 70–71, 73–75, 77–78, 174
- topic drift, 270–272
- usefulness, 49–50, 56, 63–64
- validation, 53, 56, 61
- web, 211–217, 220–233, 323, 330, 332–334, 336–337, 340, 342–345
- webconnectivity, 213–214
- web directory, 214, 216
- weblog network, 300
- web of, 269
- web, retrieval, 212
- Genre classes (genres)
  - about us page, 79
  - abstract, 79, 131, 138, 143
  - academic, 198, 212–215, 217, 221–229, 232–233, 323, 328
  - academic monograph, 138, 143
  - academic texts, 73
  - adult pages, 36–37
  - advertisement, 69, 138
  - advertising page, 79
  - archive of abstracts/archives, 79
  - argumentative, 7
  - argumentative\_persuasive, 97
  - articles, 60, 79, 122, 170
  - BBC DIYs (Do-It-Yourself), 97
  - BBC editorials, 97, 121
  - BBC feature articles, 97, 112–113
  - BBC short biographies, 97, 122
  - blogs/blogging, 13, 79, 120, 123, 138, 143, 212–215, 217, 221–226, 228–233, 333, 341–342
  - book, 79
  - business report, 138
  - calendar page, 80
  - catalogue, 123
  - “check out what a flashy page I can code”, 37
  - children’s, 124
  - code, 123
  - cognitive, 4, 7, 25, 97, 323, 325, 329–332, 334, 336–338, 340
  - comics, 37
  - commentary, 123
  - commentary page, 303
  - comments, 98, 340–342
  - commercial, 79
  - commercial homepage, 170
  - commercial page, 76
  - commercial/promotional, 124
  - communication, 96–98, 123
  - community, 124, 212–213, 215, 217, 221–233
  - company home page, 79
  - conference website, 150, 248
  - content delivery, 124
  - contributions to discussions, requests, comments; Usenet news materials, 37
  - corporate, 212–215, 217, 221–230, 232–233
  - corporate info, 37
  - corporate page, 79
  - course description, 60
  - course list, 60
  - CV, 138
  - definition page, 79
  - department site, 256, 260
  - descriptive, 7
  - descriptive\_narrative, 97
  - diary, 60

Genre classes (genres) (*cont.*)

dictionary, 123  
 directory, 79–80  
 discussion boards, 278  
 discussion page, 170  
 discussions, 73, 122, 159, 331–332, 338, 340  
 documentation, 123  
 download, 73  
 download pages, 122  
 download site, 170, 177  
 drama, 123  
 economic info, 37  
 education page, 79  
 email, 129, 138, 278, 333, 341  
 email discussion list, 300  
 entertainment, 37, 41, 124  
 entry page, 79  
 error messages, 37, 89, 94, 96, 124  
 eshops, 93, 97, 103–104, 106, 108, 120, 138, 143  
 everyday communication, 198  
 exam, 138, 143  
 executive overview/overview, 79  
 expectations, 4  
 explanation, 123, 331, 340, 343  
 explicatory\_informational, 97  
 expository, 7  
 FAQ(s), 37, 60, 79, 88–89, 97, 103, 106, 108, 112–113, 120, 123–124, 137–138, 143  
 feature, 109–110, 123, 283–286, 290–293  
 fiction, 138, 143  
 form, 34, 54, 60, 81, 123, 138  
 forum, 60, 123  
 front page, 79, 138, 143  
 full story list/list of page/stories, 79  
 games, 37  
 gateway, 79, 89, 124  
 government page, 79  
 guest books, 37, 123  
 handbook, 138  
 helps, 60, 73, 122  
 help site, 170  
 here I am, 36–37  
 highlights, 80  
 hobby page, 259, 261, 271  
 homepage, 37, 79, 82  
 home pages for the general public, 37  
 how-to page, 79  
 IDK (I Don't Know, unknown), 52, 92, 94, 106, 116  
 “I guess we have to be on the net too”, 37

index, 60, 79, 124  
 index page, 79  
 informal, private, 37  
 information, 37, 79, 123, 159, 212–213, 215–226, 228–233  
 informative, 124  
 informative advertisements, 37  
 institutional link list, 258, 261, 263, 266, 268–271  
 institutional (web) page, 258–262  
 instruction, 160  
 instructional, 7, 97  
 interactive discussion archive, 60  
 internal documents, 37  
 internet relay chat, 278  
 interview, 80, 88, 123  
 inventory, 87  
 job listing, 60  
 journal article, 79  
 journalism, 123  
 journalistic, 124, 198  
 journalistic materials, 37  
 language, 37  
 law, 123  
 lesson plan, 76, 80  
 letter, 79, 138, 140–141  
 lexicon, 123  
 limerick, 72  
 link collections, 37, 73, 122  
 link page, 79  
 links, 60  
 list, 79, 138, 143  
 listings, 121  
 list of links, 79  
 listservs, 278  
 literary, 198  
 literature, 123  
 live feeds, 13  
 macro-, 324  
 macro level, 11–12, 19  
 magazine, 79  
 magazine article, 138  
 mail, 123  
 main page, 79  
 manual page, 76  
 marginal note, 123  
 meeting notes/minutes, 79  
 memo, 80, 138, 144  
 minutes, 131, 137–138, 143  
 narrative, 7  
 navigation page, 79  
 news, 123  
 newspaper, 79

- news release, 80
- news story, 76, 80–81
- none of the above, 60
- non-government organization info, 37
- non-informative advertisements, 37
- nonprofit, 212–213, 215, 217, 221–226, 228–230, 232–233
- notes, 13
- nothing, 123
- office memo, 72
- official, 123–124, 198
- online diary, 303, 317
- online news article, 323, 333, 342
- online newspaper front pages, 97
- online posting, 303
- online shop, 177
- opinions, 98
- organization home page, 79
- organization page, 79
- other instructional materials, 60
- other listings and tables, 37
- other running text, 37
- pages with feed-back: customer dialogue; searchable indexes, 37
- participatory news article, 333–336, 342
- periodicals, 138, 143
- person, 123
- personal, 124, 212–215, 217, 221–233
- personal blog, 303–304, 317, 320
- personal documents, 37
- personal homepage, 37, 121, 138, 143, 261, 263–264, 266, 272
- personal link list, 261, 263, 266, 268–272
- personal profile, 13
- personal publication, 260–261, 268–269
- personal teaching page, 261, 271
- personal (web) page, 259–261
- personal website, 60
- picture/photo, 60
- poem, 123, 131, 138, 140–141, 143
- poetry, 60, 124
- pornographic/adult, 124
- pornography, 37
- portrait, 123
- portrayal (non-priv), 73, 122
- portrayal (priv.), 73, 122
- poster, 138
- presentation, 123
- press: news, reportage, editorials, reviews, popular reporting, e-zines, 37
- press release, 79
- private homepage, 170
- product page, 76
- product review, 70
- product for sale/shopping, 60
- propaganda, 160
- prose, 124
- protocol, 123
- public, commercial, 37
- public documents, 37
- public info, 37
- question and answer, 79
- receipt, 123
- recount, 325, 330–332
- recreation, 160–161
- reference materials, 37
- reportage, 123
- reporting, 161
- reports, 37, 331–332, 338, 340
- research project page, 260, 264
- resource page, 76, 79
- resources, 123
- review, 69–70, 123
- sales pitches, 37
- schematic drawing, 69
- scholarly article, 70
- science, 37, 123
- scientific, 124
- scientific article, 138
- scientific, legal, and public materials; formal text, 37
- searchable indices, 37
- search directory, 79
- search engine, 79
- search info, 37
- search pages, 37, 79, 93, 97, 106, 108, 121, 138, 143
- search results, 79
- search start, 60
- serious material, 37
- service page, 261
- sheet music, 138
- shopping, 124
- shops, 73, 122, 212–215, 217, 221–233
- site map, 79
- slides, 138, 143
- social media site, 300
- speech, 60
- speech transcript, 138
- sports, 37
- staff list, 260–261
- statistics, 123
- story page, 76, 81
- summary, 79
- table of contents, 60, 79
- talk, 123

- Genre classes (genres) (*cont.*)
- technical manual, 138, 143
  - technical report, 138
  - terms and conditions, 79
  - text, 324, 328
  - textbook, 73
  - thematic blog, 303–305, 320
  - thesis, 138, 140–141, 143
  - timeline, 123
  - tourism, 37
  - university subsite, 256–257
  - usenet newsgroups, 278, 300
  - user input, 124
  - weblog or blog, 60
  - welcome/homepage, 60
  - wiki, 272, 300
- Genre collections
- American blog variety of the English language, 305
  - ANC (American National Corpus), 14
  - Bank of English, 14
  - BBC web genre corpus, 93
  - BNC (British National Corpus), 14, 61, 73, 89, 158, 260
  - Brown corpus, 5, 14, 88, 150, 163
  - HARD TREC, 195
  - HGC (Hierarchical Genre Collection), 95–96, 106–107, 111–113
  - I-EN, 159
  - I-RU, 159
  - KI-04, 93, 95, 106–107, 180
  - KRYS-01, 88, 91, 137
  - LOB (Lancaster-Oslo/Bergen Corpus), 89
  - MGC (Multi-Labelled Genre Collection), 91, 96, 106–107, 111–114
  - ROMIP, 193
  - RNC (Russian National Corpus), 150–151, 155, 158–160, 163
  - SANTINIS, 92–94, 105–108, 113–114, 116
  - SPIRIT sample, 92–94, 105–106, 113–114, 116
  - Super-Genre Dataset, 215–217
  - 7-webgenre collection, 91, 93, 95, 104–105, 108–111, 117–118, 180
- Genre conventions, *see* Genre
- Genre corpora, *see* Genre collections
- Genre-enabled applications, 96, 119
- Genre-enabled prototypes, 5
- Genre evolution, *see* Genre
- Genre expectations, *see* Genre
- Genre inventory, 87, 150–151
- Genre lens, *see* Genre
- Genre model, genre modelling, *see* Genre
- Genre palette, *see* Genre
- Genre retrieval model, 24, 167–168, 172–181, 183, 186–187, 212
- Genre taxonomy, *see* Genre
- Graphics, 277, 280, 285, 334
- Graph matching, 242–243
- Graph similarity measurement, 238–239, 242, 244–245, 248–250
- H**
- Harmonic descriptor representation, 132–133, 137
- Hierarchical clustering, *see* Cluster analysis
- HTML, 16–17, 20–22, 95, 109, 125, 140, 153, 156, 163, 170, 176–177, 181–182, 192–194, 197, 219, 225, 238, 280
- Hypertext, 7, 12, 15, 17, 20–21, 24, 221, 237–239, 249
- I**
- If-then rules, 100, 104
- Inferential model, 100, 102, 105–107, 110–111, 113–118
- Inferential rater analysis, 334
- Information extraction, 5, 168–169, 171, 186, 237
- Information retrieval, 4, 36, 48, 50, 53, 130, 142, 168, 193, 197, 202, 238, 354
- Inlink, 258, 263, 266, 268–270, 272
- Internet corpora, 151, 155, 158–160
- Intertextuality, 324
- J**
- Jaccard coefficient, 116
- L**
- Link structure, 21, 211, 213–214, 218–221, 224–225, 237–238, 244, 256–258, 261, 271–273
- Logical document structure, 12, 21–22
- Logical Necessity (LN), 101–102
- Logical Sufficiency (LS), 101–102
- M**
- Machine learning (ML), 8, 18, 92, 98–100, 118–119, 137, 155–156, 164, 178, 180–181, 192–193, 354
- Manual annotation, 15, 94, 99, 114–116, 118, 352
- Mapped web genres, *see* Genre
- Metadata, 14, 48, 69, 130, 195–196, 279
- ML, *see* Machine learning (ML)
- Multi-dimensional analysis, 24, 89, 303

- Multifaceted, 71, 352
- Multi-label, multi-labelling, 90, 92, 96, 100, 104–105, 113, 118–119, 174
- Multilabel, multilabelling, *see* Multi-label, multi-labelling
- Multimedia, 277–281
- Mutual expectaions, *see* Genre
- N**
- Naive Bayes, 106, 138, 194–195, 227, 231
- Naive Bayes classifier, 106, 195
- Naïve Bayes classifier, *see* Naive Bayes classifier
- Naive Bayesian classifier, *see* Naive Bayes classifier
- Naming conventions, 13  
*See also* Genre
- Noise  
  experiments with noise, 91–92  
  structured, 91–92  
  unstructured, 92
- Nomenclature, 5, 97, 327, 352
- Normalization, 45, 102, 132
- O**
- Odds-likelihood, 100–101
- Online posting, 303
- Open Directory Project, 216
- Outlink, 214, 258, 262–263, 266, 269–272
- Overlabelling, 94, 116
- P**
- Palette, *see* Genre
- Pattern, 13, 24, 40, 54, 97–98, 100, 118, 120, 130, 132, 137, 140, 145, 157, 170, 205, 214, 221, 237, 279, 283, 292, 295, 297, 306–307, 318–320, 324–332, 334, 336, 341–343
- POS trigrams, 18, 99, 109, 156, 158, 164
- Predictivity, predictability, *see* Genre
- Preprocessing, 186, 218, 221, 226–227, 231
- Principal components analysis, 199, 282, 288, 290, 94, , 299
- Probabilities, 101–103, 176
- PROSPECTOR, 101
- Prototype theory, 7, 88, 93, 328–329
- R**
- Rainbow classifier, 131, 142–144
- Rank displacement of relevant documents, 202
- Reduction, 218, 227, 307
- Reference corpora, 18, 119, 151, 154, 351–353
- Register, 89, 279, 283, 306–307, 309, 315–316, 320, 325–326  
  spoken, 307  
  written, 307
- Rhetoric, rhetorical, 7, 97–98, 100, 102–104, 118, 120, 154, 278, 304, 324–334, 337, 340–344
- Rhetorical patterns, 97–98, 100, 118
- S**
- Scalability, 24, 105–106, 119
- Scores, 63, 102–103, 193–196, 200–201, 203, 205, 226, 229, 244, 249, 290–297, 307–317
- Search engine, 4, 15–16, 36, 39, 42–44, 47–50, 53, 56, 58, 60, 63, 66, 69, 79, 149, 151, 160, 167, 169, 171, 182, 184, 191–192, 195, 212, 216–217, 304–305
- Selection, 14, 16, 23, 39, 48, 55, 78, 93–94, 99, 124, 129, 146, 156, 164, 178, 183, 197, 199, 220, 225, 227, 261, 290, 332, 337
- Shallow features, *see* Features
- Site, *see* Website, web site
- Small world, 255–258, 270–273
- Social genres, *see* Genre
- Social network, 13, 24, 39, 88, 94, 99, 119, 255, 272, 279–283, 288, 297, 300, 330  
  analysis, 255, 279–280, 282, 288, 330
- Social process, 278–279, 300
- Spam-detection, 213, 354
- Structure  
  document, 238  
  gestalt, 331–332, 336  
  hypertext, 24, 221, 237, 239, 249  
  layout, 20, 238  
  schematic, 326–330  
  text, 238
- Stylistic differences, 33
- Support vector machine (SVM), 18, 106–107, 109, 137, 139–140, 142–145, 156, 161, 175, 179, 181, 194, 196, 206
- SVG, 280
- T**
- Task-driven search, 69
- Term frequency, 130, 133, 138, 146
- Test collections, *see* Benchmarks
- Text classification, 111, 132, 137–138, 146, 151–152, 179, 211, 324
- Text types, 7, 12, 15, 24, 73, 97, 100, 103, 149, 154–156, 160, 163, 169–170, 238, 279, 305, 315–317, 324, 328, 331–332, 334

- Tf\*idf, 132, 138
- Thesaurus, specific, 225–226, 229, 231
- Top-down approach, 7, 73, 75
- Type, 3–5, 7–8, 10, 12, 15–17, 19–20, 24, 33, 35–36, 38–40, 42–45, 47–54, 57–58, 63–66, 69, 71, 73–74, 76–78, 80, 82, 89–91, 93, 96–97, 99–100, 103, 129, 131–133, 137, 140–141, 149, 151, 154–156, 159–161, 163, 169–171, 173, 175–176, 180, 182, 184, 201, 215, 218–221, 226, 237–238, 240, 255, 257, 259, 262, 277–283, 285, 287–288, 291, 293, 296–297, 300, 303–305, 309, 315–321, 323–326, 328–334, 336, 342, 344–345
- Typification, 9
- U**
- Unit of analysis, 10, 13, 100, 211–212, 216, 305, 352
- URL, 4, 15, 20, 54, 77, 93, 156, 164, 170, 176, 193–194, 213, 215–221, 223, 238, 311
- User group, 35, 49–52, 55–56, 58–59, 61–62, 64, 66
- User profile, 279, 282–283, 288
- User validation, *see* Genre
- User warrant, 48, 55
- V**
- Variation, linguistic, 24, 279, 305–307, 315, 320
- Video, 38–39, 259, 277, 280–281, 284–285, 287–289, 291, 293, 303
- W**
- Web content mining, 11, 237
- Webcorpora, web corpora, 15, 17–18, 24, 116–117, 150–151
- Web-as-Corpus, web as corpus, 15
- Web documents, 4, 9–13, 20–23, 53, 55, 70, 73–75, 87, 92, 111, 167–168, 171, 191, 194, 201, 219, 307
- Web genre, webgenre, 6–8, 10–13, 18–22, 33, 35–36, 38, 41, 47–55, 62, 65–66, 69–75, 82, 87, 90–91, 93, 97, 100, 103–104, 106–108, 111, 113, 150, 167–170, 174–175, 177–181, 185–186, 194, 197, 211–217, 220–233, 238, 248–250, 259–261, 277, 323, 330, 332–334, 336–337, 340, 342–345, 351–354
- WebGenreWiki, 87, 352
- Web graph, 237, 256–257
- Weblog, *see* Blog
- Web mining, 216, 227, 237, 242, 249, 354
- Webometrics, 255
- Webpage, web page, 6, 10, 15–18, 20, 22, 24, 43, 47–59, 61–66, 69, 73, 76–77, 79–81, 87, 90–96, 99–100, 102, 104–109, 111–114, 116–120, 122, 124–125, 137–138, 143, 145, 149–156, 158, 161–163, 167–168, 172–174, 177, 194–195, 213–218, 220–221, 226, 238, 249, 255–256, 258–263, 266, 267, 269–272, 333–334, 353
- Website, web site, 11–12, 15, 17–22, 24, 49, 52–54, 60, 62, 64, 66, 119, 143–144, 150–151, 154, 159, 161, 237–238, 241, 248–249, 280, 286, 303–304, 306, 314 corporate, 213
- Web structure mining, 237–239
- Web usage mining, 213, 237, 239, 249
- WEGA, 5, 99, 116, 118, 168–169, 183–186, 191
- Weights, 54, 101–102, 132–133, 172, 176, 185, 193, 200, 202–203, 214, 229, 249
- World Wide Web, 50, 169, 171–172, 176, 211, 216, 255
- World-Wide Web, *see* World Wide Web
- WWW, *see* World Wide Web
- X**
- X-Site, 5, 99
- Y**
- YouTube, 277, 280, 300
- Z**
- Zerolabelling, 94, 116