Retrieval Models for Genre Classification

Benno Stein and Sven Meyer zu Eissen

Faculty of Media, Bauhaus University Weimar, Germany
{benno.stein, sven.meyer-zu-eissen}@medien.uni-weimar.de

Abstract. Genre provides a characterization of a document with respect to its form or functional trait. Genre is orthogonal to topic, rendering genre information a powerful filter technology for information seekers in digital libraries. However, an efficient means for genre classification is an open and controversially discussed issue. This paper gives an overview and presents new results related to automatic genre classification of text documents. We present a comprehensive survey which contrasts the genre retrieval models that have been developed for Web and non-Web corpora. With the concept of genre-specific core vocabularies the paper provides an original contribution related to computational aspects and classification performance of genre retrieval models: we show how such vocabularies are acquired automatically and introduce new concentration measures that quantify the vocabulary distribution in a sensible way. Based on these findings we construct lightweight genre retrieval models and evaluate their discriminative power and computational efficiency. The presented concepts go beyond the existing utilization of vocabulary-centered, genre-revealing features and open new possibilities for the construction of genre classifiers that operate in real-time.

Key words: genre analysis, retrieval models, analysis and evaluation.

1 Introduction

The term "genre" is used manifold in our culture, be it in connection with music, with literature, with entertainment, or with philosophy. In literature, for exam-

ple, more than 20 major genres and about 100 subgenres are distinguished. But, although literature of nearly any kind is available in digital form, rather few genres are distinguished with respect to digital libraries, electronic document collections, or the World Wide Web. This observation is rooted in the use case: when working with digital libraries and the like, prevalent tasks relate to text organization, text retrieval, or the clarification of some information need.

A specialized view to genre is already found at the pioneers of automatic genre classification. Roussinov et al. (2001) argue that genre should be defined in terms of purpose or function, in terms of the physical form, or in terms of the document form. And, usually a genre combines both purpose and form (Roussinov et al. 2001); similar definitions are given and discussed by (Biber 1992; Karlgren and Cutting 1994; Kessler et al. 1997). Common to all is that document genre and document content are orthogonal, i.e., documents that address the same topic can be of different genre: "The genre describes something about what kind of document it is rather than what the document is about." (Finn and Kushmerick 2003). In this way, a genre classification scheme can be oriented at the style of writing, or at the presentation style. When analyzing newspaper articles for example, typical genres include "editorial", "letter", "reportage", and "spot news".

1.1 Genre for Information Seeking and Retrieval

The current discussion about genre is coined by the new media, by digital libraries and, in first the place by the World Wide Web. In this connection the term "cyber-genre" is sometimes used, which describes the new, Web-specific genre classes as a combination of the classical elements *content* and *form*, that are supplemented by certain Web functionality (Shepherd and Watters 1998).

A major reason for the increasing interest in genre is the information overload: the exploitation of information about genre classes shall help to develop more powerful tools for information seeking and retrieval. Expectations and use cases found in the literature relate to the following fields:

- *Indexing.* Web genre can be used to tune Web crawler re-visits and inform relevance judgments for search engines (Boese and Howe 2005). Web genre identification is a key factor for reducing inadequate results of search engines, as the user would be able to specify the desired Web genre along with the keywords (Santini 2004).
- *Filtering and Ranking.* The organization of documents, bookmarks, or digital document identifiers can occur topic-centered, genre-centered, or in a combined fashion. Having identified the underlying paradigm one can pro-

vide user guidance for filing ("*This is not the correct genre!*"), give hints or special views for browsing and searching, and identify classes that are not properly organized (Stein and Meyer zu Eissen 2006). Genre information provides meta knowledge for automatic Web page abstraction, which is concerned with the preparation of Web pages in a consistent and clearly arranged form. An example is the simplification of Web pages for visually handicapped people whose access to the Internet is realized with a Braille reader.

- *User Interface.* Genre information adds a dimension to standard search interfaces. Within interactive search it can encourage the user to pursue some kind of genre-based navigation and corpus exploration (Santini 2004). Genre classes will make communications more easily recognizable and understandable by recipients (Santini 2004).

The World Wide Web is a highly dynamic information space, and, as already noted by Shepherd and Watters (2004) or by Boese and Howe (2005), the emergence of novel genres on the Web as well as the disappearance of apparently established genres happens much more rapidly than in other media (Santini 2007).

1.2 A User Study from 2004

In order to learn more about user expectations related to genre classification, we conducted a poll in 2004 at the University of Paderborn among 286 students from Information Systems, Media Science, and Computer Science (Meyer zu Eissen and Stein 2004). The students received a short introduction to genres and their use as positive and negative information filters; the questionnaire asked details about Internet search habits and usefulness assessments:

1. *Frequency of Search Engine Use.* Possible answers: "daily", "once or twice a week", "once or twice a month", "never". Rationale: Experienced search engine users have a clearer idea whether genre classification could be useful or not.
2. *Typical Query Topics.* This question, if honestly answered, suggests frequent information needs and search interests.
3. *Usefulness of Genre Classification.* Possible answers: "very useful", "sometimes useful", "not useful", "don't know". Rationale: What do prospective users expect from genre filtering as a means to satisfy an information need?
4. *Favored Genre Classes.* A selection of ten genre classes was given: publication or research article, scholar material, news, shop, link collection, help

and FAQ, personal home page, non-personal home page, discussion forum, and product presentation. Possible answers for each: "very useful", "sometimes useful", "not useful", "don't know".

5. *Additional Genre Classes.* The interviewees could specify up to three additional genre classes, labeled either as "very useful" or "sometimes useful".

6. *Comments.* A field for arbitrary comments on the idea of genre classification.

(a)

Search engine use frequency	
daily	73%
1-2x per week	23%
1-2x per month	4%
never	0%

(b)

Usefulness of genre classification	
very useful	64%
sometimes useful	29%
not useful	6%
don't know	1%

(c)

Favored genre classes	Usefulness
scholar	1.72
help/FAQ	1.53
article	1.45
discussion	1.44
shop	1.37
product information	1.36
non-personal home page	1.23
new	1.19
link collection	1.00
personal home page	0.93

Table 1. Results of a user study at the University of Paderborn in 2004: (a) shows the frequency of search engine use, (b) the expected usefulness of genre classification, (c) lists a ranking among favored genre classes; larger values indicate a higher degree in expected usefulness.

Table 1 (a) shows the search engine use frequency. From the expected usefulness assessments shown in Table 1 (b) we concluded that there is a true need to post-process query results. Table 1 (c) lists the most frequently mentioned searches under genre considerations. To make a ranked list of dedicated genre classes, scores for the usefulness of each genre class were assigned: "very useful" scored two points, "sometimes useful" scored one point, "not useful" scored no points. Additionally desired genre classes were "download site" and "Web page spam". The latter relates to paid links, sites that try to install dialers, and sites that are used to manipulate the search engine page rank. Other propositions included topics (and not genres) such as pornography. Altogether the comments were encouraging and asked for implementation.

1.3 Contributions

However, up until now its undoubted potential genre filtering was not convincing in the retrieval practice. The reason for this is threefold. First, the existing genre retrieval models are computationally too expensive to be applied in an ad-hoc manner. Second, as was also shown by Santini (2007), the proposed genre classifier technology is corpus-centered: its application to large corpora such as the Web leads to a significant degradation of the classification performance, rendering most classifiers useless for practical applications. 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/her information need.

In this paper we argue that all aspects can be addressed by the extensive use of genre-specific core vocabulary whose distribution in a document in question can be efficiently analyzed. Our contributions relate to the automatic mining of tailored core vocabularies, the use of concentration measures as a means for sensible feature quantization, and a comparative analysis of this new kind of lightweight genre retrieval model. Section 2 clarifies the notion of genre retrieval models from an information retrieval position and surveys the state of the art. Section 3 presents and analyzes the new mining technology and concentration measures to construct computationally less expensive genre retrieval models. With our findings, ad-hoc genre classification, for instance in the form of genre-enabled Web search engines, is within the realm of possibility.

2   Survey: Genre Retrieval Models in Information Systems

Genre retrieval models represent certain kinds of document models. Informally speaking, a document model captures retrieval-specific aspects of a real-world document such that an information need or a retrieval task at hand can be efficiently addressed. The terminology is not used in a consistent way: aside from the term "document model", the term "retrieval model" (Baeza-Yates and Ribeiro-Neto 1999) and sometimes the term "retrieval strategy" (Grossman and Frieder 2004) is employed. We will use the term retrieval model in the following, but at first we will give a short definition in order to establish a deeper understanding and an unmistakable usage. Figure 1 illustrates our considerations.

Definition 1 (Retrieval Model). Let D be a set of documents, and let Q be a set of information needs or queries. A retrieval model $\mathcal{R}$ for D and Q is a tuple $\langle \mathbf{D}, \mathbf{Q}, \rho_{\mathcal{R}} \rangle$, whose elements are defined as follows:

1. **D** is the set of representations of the documents D. $\mathbf{d} \in \mathbf{D}$ may capture layout aspects, the logical structure, or semantic aspects of a document $d \in D$.
2. **Q** is the set of query representations or formalized information needs.
3. $\rho_{\mathcal{R}}$ is the retrieval function and quantifies, as a real number, the relevance of a document representation $\mathbf{d} \in \mathbf{D}$ with respect to a query representation $\mathbf{q} \in \mathbf{Q}$:

$$\rho_{\mathcal{R}}: \mathbf{Q} \times \mathbf{D} \rightarrow \mathbf{R}$$

Remarks. 1. Examples for retrieval models are the vector space model, the binary independence model, or the latent semantic indexing model (Salton et al. 1975; Robertson and Sparck-Jones 1976; Deerwester et al. 1990). 2. Most retrieval models are based on the semantics and the pragmatics of a document. 3. $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$ is often specified in the form of a similarity measure. 4. Document representations and retrieval models are orthogonal concepts. However, sometimes retrieval models are associated with particular representations; e.g., the term "vector space model" is also used to denote the document representation in the form of a term vector.
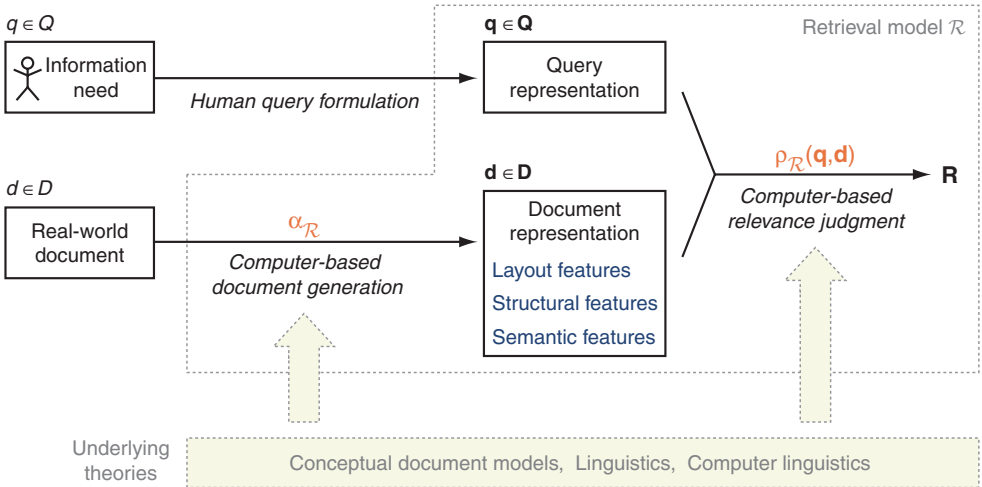

Figure 1. A conceptual view on retrieval models: in the end, an information need q is satisfied by a real-world document d. Computer-based relevance judgment requires an abstraction of q and d towards $\mathbf{q}$ and $\mathbf{d}$. The rationale for this abstraction roots in a linguistic theory and is operationalized by $\alpha_{\mathcal{R}}$ and $\rho_{\mathcal{R}}$.

The most important part of a retrieval model $\mathcal{R}$ remains hidden and cannot be made explicit in a definition, namely, the theoretical basis and the rationale behind the mapping $d \mapsto \mathbf{d}$, denoted as $\alpha_{\mathcal{R}}$ in Figure 1. Note that $\alpha_{\mathcal{R}}$ involves an inevitable simplification of d that should be:

1. quantifiable,
2. useful with respect to the information need, and
3. tailored to $\mathbf{q}$, the query representation.

Though various retrieval models have been proposed, there is little research that analyzes conceptual relations of the kind shown in Figure 1. In (Baeza-Yates and Ribeiro-Neto 1999) the authors even propose to "think first of representations of the documents", which, obviously, is a very restricted view. The narrow view on retrieval models has also been criticized in (Grossman and Frieder 2004), where the authors identify as reason for the weakness of many retrieval models their focus and simplistic use of document terms. Actually, the relation between real-world documents D and associated document representations $\mathbf{D}$ should allow any kind of transformation $\alpha_{\mathcal{R}}$. This view is also held in (Fuhr 2004), and it is illustrated here as shown in Figure 2.
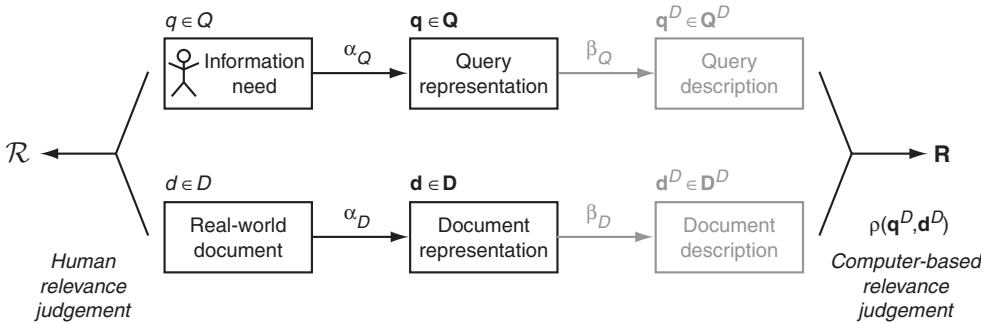

Figure 2. Conceptual model for text retrieval according to Fuhr (2004).

The letters Q, D, $\mathbf{Q}$, and $\mathbf{D}$ have the same semantics as in Definition 1. In addition we are given the sets $\mathbf{Q}^D$ and $\mathbf{D}^D$, along with the respective mappings β_Q and β_D, which realize a transformation from the representation level to a so-called description level. Essentially this third level is a data structure level and, in our definition the level is omitted since it can be derived canonically from $\mathbf{Q}$ and $\mathbf{D}$. Also note that we have omitted the mapping α_Q in Figure 1, since in existing IR applications

this transformation is left to the user. We now introduce genre retrieval models as a special form of retrieval models.

A genre retrieval model is a retrieval model that addresses queries, Q, related to a palette of genre classes in question. Hence there is a genre classifier $\gamma_{\mathcal{R}}$ instead of a retrieval function $\rho_{\mathcal{R}}$.

Definition 2 (Genre Retrieval Model). Let D be a set of documents, and let Q be a set of genre class labels, $Q = \{c_1, ..., c_k\}$. A genre retrieval model $\mathcal{R}$ for D and Q is a tuple $\langle \mathbf{D}, \gamma_{\mathcal{R}} \rangle$ whose elements are defined as follows:

1. $\mathbf{D}$ is the set of representations of the documents D.
2. $\gamma_{\mathcal{R}}$ assigns one or more genre class labels to a document representation $\mathbf{d} \in \mathbf{D}$:

$$\gamma_{\mathcal{R}} : \mathbf{D} \rightarrow \mathcal{P}(\{c_1, ..., c_k\})$$

The development of retrieval models is an active research field with various open questions. In spite of its simplicity the vector space model is quite successful; recent work focuses on probabilistic models as well as on models that rely on hidden variables (Hofmann 2001; He et al. 2001; Cai and He 2005). With respect to special purpose retrieval tasks, such as genre classification, even less is known concerning the user's information need and the adequacy of genre retrieval models. The related research, which is outlined now, combines approved and new retrieval model approaches with machine learning technology.

2.1 Research in Automatic Genre Classification

We see three main characteristics according to which research in automatic genre classification should be distinguished:

1. *Use Case.* Automatic genre classification can be useful in several respects, amongst which the search applications count the most popular ones. Other applications include the categorization of documents, bookmarks, or digital document identifiers, which can occur topic-centered, genre-centered, or in a combined fashion, or the development of special-purpose Web services that rely on genre information, such as Web page abstraction or automatic market forecast summarization (Stein et al. 2005; Stein and Meyer zu Eissen 2006).
2. *Corpus and Genre Palette.* The underlying corpus and the interesting genre palette are often determined by the use case. Nevertheless we mention them

Author	Use case	Corpus / Sample size	Genre classes Q	Document representation $\mathbf{d}$	Feature number
(Karlgren and Cutting 1994)	feasibility study	Brown corpus / 500	press, miscellaneous, non-fiction, fiction	part-of-speech, function words, text statistics	20
(Kessler et al. 1997)	feasibility study	Brown corpus / 499	reportage, editorial, scitech, legal, non-fiction, fiction	linguistic, character-level and derivative cues	55
(Yoshioka and Herman 2000)	information coordination	–	conference brochure, paper submission, hotel, travel	–	–
(Stamatatos et al. 2000)	feasibility study	Wall Street Journal / 160	editorial, letter to the editor, reportage, spot news	most frequent words of BNC, most frequent punctuation marks	38
(Antunes et al. 2001)	electronic meeting system	–	brainstorming, creating consensus, disseminating information, planning	–	–
(Rauber and Müller-Kögler 2001)	integrated search interface	own corpus / 1000	interview, emotional report, factual report, announcements, technical documents	text complexity measures, special character and punctuation, stopwords and keywords, mark-up	>300
(Dewdney et al. 2001)	feasibility study	CMU corpus / 9705	advertisement, bulletin board, FAQ, message board, radio news, Reuters newswire, television news	part-of-speech, closed-class word sets, presentation, sentence complexity, layout	300
(Finn and Kushmerick 2003)	feasibility study	own corpus / 2150	subjective versus objective, positive versus negative	bow without stop-words, part-of-speech, easily computable text statistics	45 + |bow|

Table 2. Research in the field of automatic genre classification for non-Web corpora. Note that the major part of the research has no use case in mind but reasons about genre palettes and the feasibility of an automatic classification.

as independent characteristics since corpus size and type, the number of genre classes and their granularity, or a single-label versus multi-label assumption depend on the actual application scenario. The "corpus" World Wide Web adopts an exceptional position, which is rooted in its size, its prominence, its ubiquity, but also in its cultural drift (Boese and Howe 2005). Moreover, as noted by Santini, Web documents can be considered as a sort of container of multiple texts and hence make a multi-label strategy necessary (Santini 2007).

3. *Retrieval Model.* The employed genre retrieval model defines the set of document representations $\mathbf{D}$, the complexity of its construction, $\alpha_{\mathcal{R}}$, as well as

Author	Analysis basis	Genre classes Q	Document representation d	Feature number
(Bretan et al. 1999)	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	–
(Crowston and Williams 2000)	837 documents	reproduced genre, adapted genre, home page, hotlist, Web server, interactive, unclassified	–	–
(Roussinov et al. 2001)	user study with 184 interviewees	Genre classes: home page, article, news, product information, education, FAQ	–	–
(Dimitrova et al. 2002)	–	three genre dimensions: expertise, detail, subjectivity	–	–
(Lee and Myaeng 2002)	7615 documents	FAQ, home page, reportage, editorial, research article, review, product specification	genre-specific core vocabulary	166
(Rehm 2002)	200 documents	hierarchy with three granularity levels for academic home pages	HTML meta data, presentation related tags, linguistic features	–
(Meyer zu Eissen and Stein 2004)	user study with 286 interviewees, 800 documents	article, discussion, shop, help, non-personal home page, 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	35
(Kennedy and Shepherd 2005)	321 documents	personal, corporate, organizational	HTML tags, phone, email, presentational tags, CSS, URL, link, script, genre-specific core vocabulary	25
(Boese and Howe 2005)	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	>100 + |bow|
(Lim et al. 2005)	1224 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	-
(Santini 2007)	1400 documents	blog, listing, e-shop, home page, FAQ, search page, online newspaper front page	most frequent English words, HTML tags, part-of-speech, punctuation symbols, genre-specific core vocabulary	>100
(Santini 2007)	2480 documents	[as before]	text type analysis plus a combination of layout and functionality tags	50 rules

Table 3. Research in the field of automatic genre classification for Web-based corpora and digital libraries. Typical use case is the development of richer representation forms for retrieval results in the search interface.

the achievable performance of the classifier $\gamma_{\mathcal{R}}$. Section 3 is devoted to computational and performance aspects of genre retrieval models.

Early work in automatic genre classification dates back to 1994, where Karlgren and Cutting (1994) presented a feasibility study for a genre analysis based on the Brown corpus. Several publications followed later on, investigating different corpora, using either more intricate or less complex retrieval models, stipulating other concepts of genre, or reporting on new applications; most of this work is mentioned in Table 2 (non-Web corpora) and Table 3 (Web-based corpora), following a chronological order. In Subsection 3.4, as part of our computational analysis, Table 9 compiles supplementary information: it contrasts the achieved classification performances of the mentioned approaches—as far as the authors conducted experiments—and reports on them.

3 Technology: Computational Aspects of Genre Retrieval Models

An issue of genre classification in general and Web genre classification in particular is that even humans are not able to consistently specify the genre of a given document or page. Take for instance a tutorial on machine learning, which could either be

Genre class	Description
Help	All pages that provide assistance, e.g., Q&A or FAQ pages.
Article	Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters.
Discussion	All pages that provide forums, mailing lists, or discussion boards.
Shop	All kinds of pages whose main purpose is product information or sale.
Non-personal home page	Web appearances of companies, universities, and other public institutions. I.e., home or entry or portal pages, descriptions of organization and mission, annual reports, brochures, contact information, etc.
Personal home page	Personal self-portrayals, private home pages with informal content.
Link collection	Documents which consist of link lists for the main part.
Download	Pages on which freeware, shareware, demo versions of programs etc. can be downloaded.

Table 4. The genre palette underlying our analyses is comprised of 8 classes.

classified as scholar material or as research article. Scholar material can be regarded as a super-genre that covers help, article, and discussion page. Another observation is that most product information sites are combined with a shopping interface, rendering a discrimination between shops and products impossible. However, as can be seen from Table 3, there is—to certain degree—a common understanding of genre on the Web. The genre palette underlying the analyses of Subsection 3.4 is listed in Table 4 and reflects the genre assessment of many human information miners; it has been used by other authors as well (Boese and Howe 2005; Stein et al. 2005; Santini 2007). Note, however, that the idea of genre-specific core vocabularies, the presented mining technology, as well as the concentration measures are independent of the chosen palette.

The section is organized as follows. Subsection 3.1 overviews the different kinds of features that are used to construct genre retrieval models, while Subsection 3.2 outlines two approaches to automate the mining of genre-specific core vocabulary. Subsection 3.3 presents new measures that quantify the distribution of core vocabularies in documents. In turns out that genre retrieval models that are built solely with vocabulary-centered features provide a discrimination power that can compete with the best models developed so far; Subsection 3.4 reports on the analyses.

3.1 Components of Genre Retrieval Models

With respect to the investigated features the existing literature on genre classification falls into three groups: retrieval models that rely on a subset of a document's terms (also called bag-of-words, bow) (Stamatatos et al. 2000; Lee and Myaeng 2002), retrieval models that employ linguistic features along with additional features relating to text statistics (Kessler et al. 1997), or both (Finn and Kushmerick 2003). This section gives an overview of these features. The features imply different computational efforts under the O-calculus; the presented complexities refer to n as the length (= number of words) of a document d, and to m as the number of words in a document collection D. In the following, a hash table access to the dictionary of m words is assessed as constant.

Average Word Frequency Class. This feature is of a new kind; it bases on offline acquired large-scale corpus statistics and was proposed for the first time in (Meyer zu Eissen and Stein 2004). The frequency class of a word is directly connected to Zipf's law and can be used as an indicator of a word's customariness. Let $f(w)$ denote the frequency of a word $w \in D$, and let $r(w)$ denote the rank of w in a word list of D, which is sorted by decreasing frequency. In accordance with (University of Leipzig 1995) we define the word frequency class $c(w)$ of a word $w \in D$ as

$\lfloor\log_2(f(w^*)/f(w))\rfloor$, where w^* denotes the most frequently used word in D. In the Sydney Morning Herald Corpus (Dennis 1995), w^* denotes the word "the", which corresponds to the word frequency class 0; the most uncommonly used words within this corpus have a word frequency class of 19. The analysis of the word frequency class as genre characteristic goes back to the observation that research articles use a more specialized speech than for instance advertising texts. The complexity of the language usage is reflected in the word frequency class averaged over all words in d. The computational complexity is in $O(n)$ for a document d.

Part-of-Speech Analysis. Part-of-speech (pos) analysis labels the words of a sentence according to their function or word class. Part-of-speech taggers analyze a word's morphology or its membership in a particular set. In this connection one differentiates between so-called open-class word sets and closed-class word sets, where the size of the former is not bound by a finite constant; examples are nouns, verbs, adjectives, or adverbs. Examples for closed-class word sets are prepositions and articles. For our analysis we have employed the part-of-speech tagger of the University of Stuttgart (University of Stuttgart 1996); Table 5 lists important word classes. The computational complexity for the pos-tagging of a document d is in $O(n \cdot k^2)$, with k denoting the number of tags, if the Viterbi algorithm is employed.

Nouns	verbs	relative pronouns	relative prepositions
adverbs	articles	pronouns	modals
adjectives	alphanumeric words		

Table 5. Word classes used in part-of-speech analyses

Syntactic Group Analysis. A syntactic group analysis yields linguistic features that relate to several words of a sentence. They quantify the use of tenses, relative clauses, main clauses, adverbial phrases, simplex noun phrases, etc. Since the analysis is computationally expensive, features that base on syntactic groups are commonly not part of a genre retrieval model. Dewdney et al. (2001), however, include the transition in verb tense within a sentence in their analysis. A lower bound for the complexity of syntactic group analyses is given by the part-of-speech analysis, whereas the effort for regular expression matching and the application of heuristic rules add to this lower bound.

Text Statistics. Under the label "text statistics" we comprise features that relate to the frequency of easily accessible syntactic entities: clauses, paragraphs, delimiters, question marks, exclamation marks, or numerals. Counts for these entities are put

in relation to the number of words of a document. Kessler et al. (1997) designate features of this type as "character-level cues"; Finn and Kushmerick (2003) designate such features as "hand-crafted". The computational complexity is in $O(n)$ for a document d.

Presentation-Related Features. These type of features relate to the appearance of a document. They include frequency counts as well as particular HTML-specific concepts and stylistic concepts. To the former we count the number of figures, tables, paragraphs, headlines, or captions. The latter comprises statistics related to the usage of colors, hyperlinks (anchor links, site-internal links, Internet links), URLs, mail addresses, etc. The computational complexity is in $O(n)$ for a document d but implies a considerably larger constant than simple text statistics.

Genre-Specific Core Vocabularies. Aside from word classes that relate to grammatical function, other closed-class word sets can be compiled that may be specific to a certain genre: currency symbols, help symbols ("FAQ", "Q&A", "support"), shop symbols, months, days, countries, first names, and surnames. Speaking informally, the core vocabulary of a genre class comprises terms that are significant for this class compared to the other genre classes. It is difficult to state a more precise definition: core vocabulary cannot be considered closed-class, and, its "significance" is not determined in some standard IR weighting scheme manner but requires a distribution analysis. However, the complexity for the computation of core vocabulary features is typically in $O(n)$ for a document d.

The next subsection describes the automatic mining of such core vocabularies.

3.2 Mining Genre-Specific Core Vocabularies

In order to differentiate between typically 6-10 genre classes, existing approaches squeeze various kinds of the mentioned features into a genre retrieval model. The effort to compute these features for a document d varies considerably, contributing more or less (and still questionable) discriminative power. In our opinion, features relating to genre-specific core vocabularies are underestimated. We explain this deficit as follows: First, genre-specific core vocabularies are compiled manually, following intuition. Second, the quantification of genre-specific core vocabularies is limited to simple count statistics. Third, the positive trade-off between computational effort and discriminative power of genre-specific core vocabularies is either unknown or underestimated.

For the set D of documents let $C = \{C_1, ..., C_k\}$ be an exclusive genre categorization of D. I.e.,

$$\bigcup_{i=1..k} C_i = D \text{ and } \forall C_i, C_{j,i \neq j} \in C: C_i \cap C_j = \varnothing.$$

For a genre class $C \in C$, let T_C denote the specific core vocabulary for C. Similar to Broder (2002) we argue that T_C is comprised of navigational, transactional, structural, and informative terms. 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 such as time and space. Examples include the meta information of posts in a discussion forum ("thread", "replies", "views", elements of dates) and terms that appear in addresses on home pages ("address", "street").
- *Informative Terms.* Do not appear in functional HTML elements but do, however, imply functionality. 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 C with algorithms from topic identification research (Popescul and Ungar 2000; Lawrie et al. 2001; Lawrie and Croft 2003; Stein and Meyer zu Eissen 2004). In the following, we show how these approaches are adapted for the mining of genre-specific core vocabulary.

Popescul's Method

Let C be a genre classification of D, let $P(w)$ denote the probability that term w is drawn from D, and let $P(w|C)$ denote the probability that w is drawn from the documents of some $C \in C$. Popescul and Ungar (2000) assign a score $f_C(w)$ to each term w in each class C that quantifies its class-specific frequency and predictiveness:

$$f'_C(w) = P(w|C) \cdot \frac{P(w|C)}{P(w)}$$

The first factor on the right-hand side prefers frequent terms in C, the second factor quantifies w's predictiveness with respect to C. Terms that score high are considered as candidates for T_C.

Adaptation of Popescul's Method. Although f_C prefers the frequent and predictive terms of a genre class C, it does not quantify how *representative* a term is for C. If some term w is used extensively in a single document, $P(w|C)$ is large and hence w is likely to become part of C's specific vocabulary—though w may not be representative for C. The following statistic tackles this problem by introducing information about a term's distribution:

$$f'_C(w) = P(w|C) \cdot \frac{P(w|C)}{P(w)} \cdot \frac{df_C(w)}{|C|}$$

where $df_C(w)$ denotes the number of documents from C that contain w.

Weighted Centroid Covering (WCC)

The WCC algorithm identifies characteristic terms in a set $C = \{C_1,...,C_k\}$ of document sets and was developed for the labeling of document categories (cf. Algorithm 1). Let W denote the set of terms underlying D, let $tf_C(w)$ denote the term frequency of w in C, and let $rank : W \times \mathbf{N} \rightarrow C$ be a ranking function:

$$rank(w,i) = C \quad \Leftrightarrow \quad C \text{ is at rank } i \text{ when ranked according to } tf_C(w).$$

Say, $rank(w,1)$ and $rank(w,|C|)$ return the document set in which w appears most and least frequently. In a first step, WCC selects for each term the k categories in which the term occurs most frequently. These k categories are stored along with frequency information and the associated term in a 3-tuple, which is inserted into table T. In a second step, T is sorted in descending order with respect to $tf_C(w)$, and terms from T's tuples are assigned to the genre classes in a round-robin manner. Here, the terms from tuples with the largest $tf_C(w)$ value are assigned to the associated categories while processing T top-down each round. The overall time complexity of WCC is in $O(|W| \cdot log(|W|))$.

Adaptation of WCC. Stein and Meyer zu Eissen (2004) showed that the terms assigned by WCC are both expressive and discriminating. However, as pointed out above, the genre-specific core vocabulary should be representative for the *entire* genre class. This information can be introduced into WCC's term selection proc-

ess by multiplying a tuple's score with $df_C(w) \, / \, |C|$. The modification decreases the score of terms that are used only in a small number of C's documents.

```
Input:      C     Genre categorization
            l     Number of specific terms to extract per genre,
            o     Occurrences of the same term in the vocabulary of different
                  genres.
Output:     T_C   Specific core vocabulary for each C ∈ C

  1. T = ∅;
     FOREACH C IN C DO   T_C = ∅;
  2. FOREACH w IN W DO
        FOR i=1 TO o DO
           C=rank(w, i);
           f=tf_C(w);
           insert ⟨w, f, C⟩ into T
        ENDFOR
     ENDDO
  3. Sort T with descending term frequencies;
  4. FOR round=1 TO l DO
        j=1;
        WHILE not all categories got a term DO
           let t_j=⟨w, f, C⟩ be jth tuple of T;
           IF C got no new term this round THEN
                 T_C = T_C ∪ {w};
              delete t_j from T;
           ENDIF
           j=j+1;
        ENDWHILE
     ENDFOR
  5. RETURN {T_C | C ∈ C};
```

Algorithm 1. The WCC-algorithm for topic identification provides an adequate means for core vocabulary mining.

3.3 Concentration Measures

To design a genre retrieval model $\mathcal{R} = \langle \mathbf{D}, \gamma_{\mathcal{R}} \rangle$, effective features for $\mathbf{D}$ must be found to measure the presence or absence of genre-specific core vocabulary. 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 standard measures, and hence it cannot be learned by a classifier $\gamma_{\mathcal{R}}$. Examples for Web pages on which genre-specific core vocabulary appears concentrated are personal 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). In the following, two statistics are presented that quantify different vocabulary concentration aspects.

Maximum Term Concentration. Let $d \in D$ be represented as a sequence of terms, $d = \{w_1, ..., w_m\}$, and let $W_i \subset d$ be a text window of length l in d starting with term i, say, $W_i = \{w_i, ..., w_{i+l-1}\}$. 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:

$$\mathcal{K}_{T_C}(W_i) = |W_i \cap T_C|/l, \quad \mathcal{K}_{T_C}(W_i) \in [0,1]$$

The *overall concentration* is defined as the maximum term concentration:

$$\mathcal{K}^*_{T_C}(W_i) = \max_{W_i \subset d} \mathcal{K}_{T_C}(W_i), \quad \mathcal{K}^*_{T_C}(W_i) \in [0,1]$$

Gini Coefficient. In contrast to the $\mathcal{K}_{T_C}$ 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 (see Figure 3 for an illustration).

If v_i is measured for non-overlapping windows of the same length, the Gini coefficient is computed as follows. Let $v_{\pi_1}, ..., v_{\pi_l}$ denote the sorted sequence of v_i-values, i.e. $v_{\pi_1} \leq v_{\pi_2} \leq ... \leq v_{\pi_l}$. Furthermore, let $S = \sum v_i$ for $1 \leq i \leq l$ be the sum of all values, and let $y_r = \sum v_{\pi_i}/S$ for $1 \leq i \leq r$ be the rth normalized partial sum of the sorted sequence. The Gini coefficient g and the normed Gini coefficient g^* are:

$$g = 1 - \sum (y_i + y_{i-1})/l \text{ for } 1 \leq i \leq l \text{ with } g^* = l/(l-1)g, \text{ whereas } y_0 \text{ is set to 0.}$$
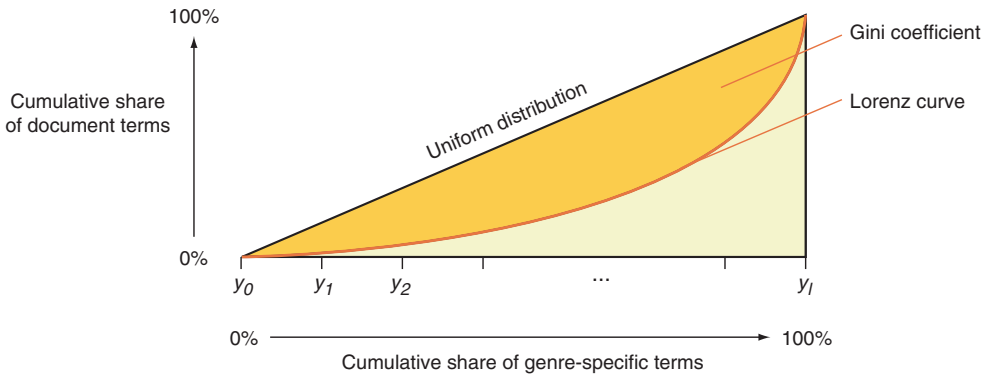
Figure 3. The Gini coefficient shows the development of the cumulative share of the genre vocabulary related to all document terms. The larger the area is between the line of uniform distribution and the Lorenz curve (= Gini coefficient g, $g \in [0,1]$) the more concentrated is the genre vocabulary.

3.4 Performance Analysis

This subsection reports on experiments and performance results that are of practical interest when constructing genre retrieval models for real-world applications. First, the core vocabulary mining methods introduced in this paper are evaluated; in a sense this is a comparison between vocabulary-based genre retrieval models. Second, vocabulary-based genre retrieval models are compared to rich feature set retrieval models as they are currently used for genre classification. The basis is a corpus consisting of about 2300 Web pages, which are uniformly distributed over 8 genre classes (Meyer zu Eissen and Stein, 2004). We also give an overview of reported performances in the literature.

Vocabulary Formation

We analyze vocabulary mining methods, statistics for vocabulary quantification, and the vocabulary distribution.

Mining Methods and Core Vocabulary Size. To get an idea of how the number of terms in T_C determines the classification performance, discriminant analyses with a genre model that consists only of the two presented concentration features were conducted for different sizes of T_C. Figure 4 shows that a vocabulary with approximately 40 terms per genre is most effective; sets with more terms can lead to over-

fitting and introduce noise. The achieved classification performances became possible through the concentration measures presented in Section 3.2. Table 6 gives examples for highly ranked terms in the genre-specific core vocabularies T_C.
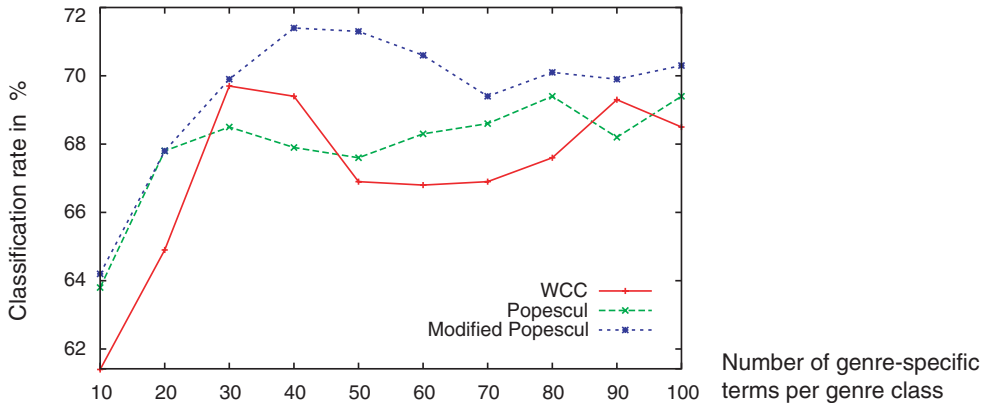

Figure 4. Classification rate depending on the size of the extracted core vocabulary. The extraction methods are WCC, Popescul, and modified Popescul. The classification rate refers to a classifier that uses concentration features only.

Rank	Discussion	Download	Help	Shop	Non-personal home page
1	forum	download	question	price	servic
2	post	version	faq	shop	busi
3	pm	instal	answer	store	develop
4	discuss	mb	find	gift	contact
5	topic	window	time	cart	new

Table 6. Highly ranked terms in the genre-specific core vocabularies TC. The terms were mined with Popescul's method and stemmed with Porter's algorithm (Porter 1980).

Core Vocabulary Quantification. Aside from an automatic core vocabulary formation the quantification of its significance is of exceptional importance. Is it the presence, the distribution, or the variance which plays the most important role? We propose the use of particular concentration measures, which capture both distributional and variance-related characteristics. Table 8 compares the concentration statistics (row 2b) to the term frequency statistic (row 2a), which is usually employed on tailored closed-class word sets.

Distribution of Core Vocabulary over Genre Classes. For this experiment, a subset of 2000 documents, uniformly distributed over the 8 genre classes, was drawn and

then split into equally sized training and test sets. The training set was used to create the genre-specific vocabularies T_C with the algorithms from Section 3.2. Due to construction, documents from a genre class C are likely to contain terms from T_C, but they will also contain terms from other genre-specific vocabularies. In this first experiment this fraction was measured for each genre class in the test set. Figure 4 shows the distribution of genre-specific core vocabularies, which were extracted with WCC. Both WCC and Popescul's method manage to identify vocabularies that constitute the major part in the respective genre classes.
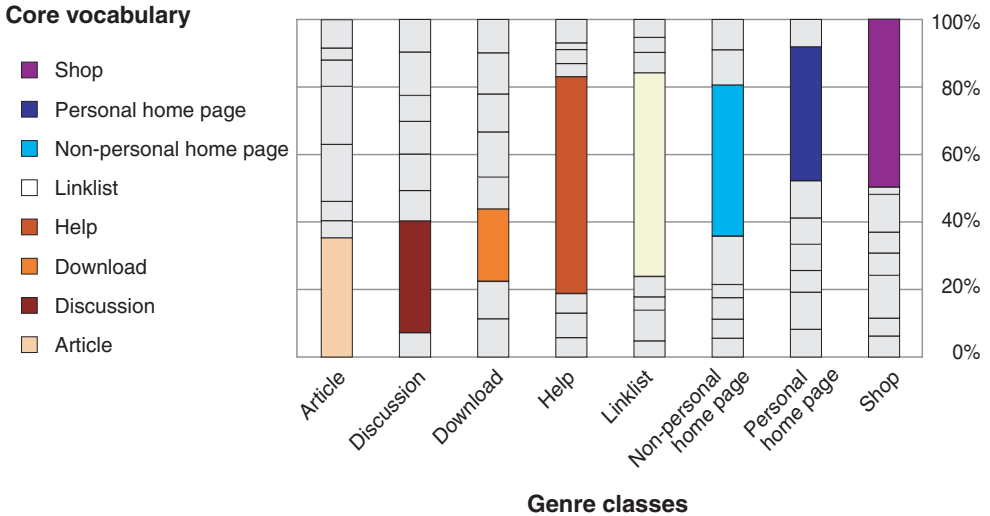

Figure 5. Distribution of the genre-specific core vocabularies over the 8 genre classes.

Retrieval Models: Vocabulary-based versus Rich Feature Set

We analyze vocabulary-based and standard genre retrieval models with respect to their generalization ability and their classification performance.

Generalization Behavior. As already mentioned, the existing generation of genre retrieval models shows poor generalization behavior. It is difficult to assess generalization behavior of a model at all; however, we can measure the amount of training data that is necessary to fit a model. Such an analysis gives us an idea of how easy the fitting process yields a corpus-specific model.

Table 7 illustrates the improvement in the fitting quality (measured by the classification performance) dependent on the size of the training corpus. Here, the test

set consists of 1000 documents for all training sets. The deviation number quantifies the increase in classification performance when more training data is available. Note that a smaller deviation indicates a better generalization behavior: the model gets fitted with smaller training sets. Also note that the vocabulary concentration measures form the most robust feature class.

	Size of training set					
	200	400	600	800	1000	Deviation
1. Part-of-speech	0.41	0.45	0.50	0.53	0.56	36.6%
2. Core vocabulary	0.62	0.66	0.66	0.68	0.70	12.9%
3. HTML	0.25	0.24	0.23	0.29	0.29	16.0%
1 + 2 + 3	0.62	0.69	0.69	0.71	0.72	16.1%
VSM (baseline)	0.35	0.52	0.59	0.64	0.68	94.3%

Table 7. Performance of classifiers that where trained with training sets of different sizes. The deviation refers to the first value and the last value in a row; a lower deviation means that the resulting genre retrieval model is less corpus-dependent and hence less susceptible for overfitting.

Classification Performance. Table 8 contrasts different genre retrieval models. On 2000 documents, uniformly distributed over the 8 genre classes, classifiers have been trained that employ specific subsets of all possible features: HTML, part-of-speech, and concentration measures. The *F*-Measure values were determined for each genre class using a 10-fold cross validated discriminant analysis, averaged over the 8 genre classes. For comparison, a Naive Bayes classifier that uses the plain text of the HTML documents has also been employed to show that our tailored core vocabulary model is more powerful than a standard text classification approach.

The discriminative power of core vocabulary features is impressive: even when they are combined with HTML and part-of-speech features, the overall classification performance is only marginally increased.

Average F-Measure for different retrieval models	
1. Part-of-speech	0.74
2a. Core Vocabulary (tf)	0.71
2b. Core vocabulary (concentration)	0.80
3. HTML	0.31
1 + 2b + 3	0.84
VSM (baseline)	0.68

Table 8. Classification performance of different genre retrieval models.

Performance Reported in Literature. Based on a literature research, Table 9 surveys classification technology and reported performances. Observe the large variance, which is rooted in the different complex genre palettes, the various sized and noisy corpora, and the different complex retrieval models. Hence, for future analyses, a commonly accepted genre corpus is necessary, and an initiative for this purpose has just been started (see http://corpus.leeds.ac.uk/serge/webgenres).

Author	Sample size / Genre classes	Classification technology	Classification Performance
(Karlgren and Cutting 1994)	500 / 4	discriminant analysis	73%
(Kessler et al. 1997)	499 / 7	logistic and multiple regression, neural networks	65%
(Stamatatos et al. 2000)	160 / 4	discriminant analysis	97%
(Dewdney et al. 2001)	9705 / 7	naive Bayes classifier, C4.5, support vector machines	92%
(Finn and Kush-merick 2003)	2150 / 2	C4.5	79% 49% (domain transfer)
(Lee and Myaeng 2002)	7615 / 7	naive Bayes classifier	90%
(Meyer zu Eissen and Stein 2004)	800 / 8	discriminant analysis, support vector machines	75%
(Kennedy and Shepherd 2005)	321 / 3	neural networks	70%
(Boese and Howe 2005)	342 / 9	additive logistic regression	90%
(Lim et al. 2005)	1224 / 16	*k*-nearest-neighbor	75%
(Santini 2007)	1400 / 7	support vector machines	90%
(Santini 2007)	2480 / 7	two-level inference based on modified Bayes rule and modus ponens	86%

Table 9. Overview over genre classification experiments, employed classification technology, and achieved classification performance found in the literature. Because of different complex genre palettes, various sized and noisy corpora, as well as different complex retrieval models the performance figures cannot be compared directly.

4 Summary and Conclusion

Efficient and reliable genre classification can play an important role in upcoming applications for information mining and retrieval. Various use cases will benefit from genre classification technology: interfaces for search and navigation in digital libraries, high-level Web services such as report and conference finders, better methods for Web page abstraction, adaptive personal information filters, or the automatic analysis of social software applications such as blogs, to mention only a few. Basis of these applications are tailored genre retrieval models, and our paper gave a comprehensive and comparative survey of the state of the art of this information retrieval branch.

Though genre retrieval models are special document retrieval models they are constructed quite differently: different kinds of simple and complex features, among others from the field of natural language processing, are combined and statistically optimized to capture the "gist" of a genre class. In this connection we observed that features relating to genre-specific core vocabularies are not used to their full potential, and our research contributes in this respect: the article presented methods to mine genre-specific core vocabularies and new statistics to measure vocabulary concentration. We showed that genre retrieval models that rely exclusively on core vocabularies reach a classification performance comparable to the best known genre retrieval models, while being an order of magnitude faster at the same time. This fact makes the ad-hoc creation of personalized genre retrieval models for various Web-based search applications possible.

Our findings, along with the experiences acquired with other genre retrieval models are compiled in Table 10. It illustrates the qualitative trade-off between the computational effort underlying different kinds of features versus the achievable discriminative power.

Computational effort	Discriminative power			
	low ⟵		⟶ high	
high	syntactic group analysis	emphasing and toning part-of-speech analysis		
	closed-class word sets	text complexity measures	URL and link analysis	
low	simple text statistics	HTML tags	averaged word frequency class	genre-specific core vocabulary

Table 10. Qualitative trade-off between the computational effort underlying different kinds of features versus the achievable discriminative power for a genre analysis.

References

Antunes, P., Costa, C., and Ferreira Dias, J., "Applying Genre Analysis to EMS Design: The example of a small accounting firm," in *Proceedings of the Seventh International Workshop on Groupware, CRIWG 2001*, IEEE CS Press, pp. 74-81.

Baeza-Yates, R., and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison Wesley, 1999.

Biber, D., "The Multidimensional Approach to Linguistic Analyses of Genre Variation: An overview of methodology and findings," in *Computers and the Humanities*, (26), 1992, pp. 331-345.

Boese E., and Howe A., "Effects of Web Document Evolution on Genre Classification," in *Proceedings of the CIKM'05*, ACM Press, 2005.

Bretan, I., Dewe, J., Hallberg, A., Wolkert, A., and Karlgren, J., *Web-Specific Genre Visualization*, 1999.

Broder, A., "A Taxonomy of Web Search," *SIGIR Forum*, 2002.

Cai, D., and He, X., "Orthogonal Locality Preserving Indexing," in *SIGIR'05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

Crowston, K., and Williams, M., "Reproduced and Emergent Genres of Communication on the World-Wide Web," *The Information Society*, (3), 2000, pp. 201-216

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R., "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, (41:6), 1990, pp 391-407.

Dennis, S., "The Sydney Morning Herald Word Database," http://www2.psy.uq.edu.au/CogPsych/Noetica/OpenForumIssue4/SMH.html, 1995.

Dewdney, N., VanEss-Dykema, C., and MacMillan, R., "The Form is The Substance: Classification of genres in text," in *Proceedings of ACL Workshop on Human Language Technology and Knowledge Management*, 2001.

Dimitrova, M., Finn, A., Kushmerick, N., and Smyth, B., "Web Genre Visualization," in *Proceedings of the Conference on Human Factors in Computing Systems*, 2002.

Finn, A., and Kushmerick, N., "Learning to Classify Documents According to Genre," in *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

Fuhr, N., "Vorlesung Information Retrieval SS'04," http://www.is.informatik.uni-duisburg.de/courses/ir_ss04, 2004.

Grossman, D., and Frieder, O., *Information Retrieval*, (Second ed.), Springer, 2004.

He, X., Cai, D., Liu, H., and Ma, W.-Y., "Locality Preserving Indexing for Document Representation," in *SIGIR'04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

Hofmann, T., "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, (42), 2001, pp.177-196.

Karlgren, J., and Cutting, D., "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis," in *Coling 94: Proceedings of the 15th. International Conference on Computational Linguistics*, (II), Kyoto, Japan, 1994, pp. 1071-1075.

Kennedy, A., and Shepherd, M., "Automatic Identification of Home Pages on the Web," in *HICSS'05: Proceedings of the 38th Annual Hawaii International Conference on System Science*, 2005.

Kessler, B., Nunberg, G., and Schütze, H., "Automatic Detection of Text Genre," in *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Somerset, New Jersey, 1997, pp. 32-38.

Lawrie, D., Croft, W., and Rosenberg, A., "Finding Topic Words for Hierarchical Summarization," in *SIGIR'01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* September 9-13, 2001, New Orleans, Louisiana, USA, pp. 349-357.

Lawrie, D.J., and Croft, W., "Generating Hierarchical Summaries for Web Searches," in *SIGIR'03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2003, Toronto, Canada, pp. 457-458.

Lee, Y.-B., and xMyaeng, Y.-B., "Text Genre Classification with Genre-Revealing and Subject-Revealing Features," in *SIGIR'02: Proceedings 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 2002, pp. 145-150.

Lim, C., Lee, K., and Kim, G., "Automatic Genre Detection of Web Documents," in *IJCNLP'04: Proceedings of Natural Language Processing*, Springer, 2005, pp. 310-319.

Meyer zu Eissen, S., and Stein, B., "Genre Classification of Web Pages: User Study and Feasibility Analysis," in *KI'04: Advances in Artificial Intelligence*, volume 3228, LNAI, Springer, 2004, pp. 256-269

Popescul, A., and Ungar, L., "Automatic Labeling of Document Clusters," http://citeseer.nj.nec.com/popescul00automatic.html, 2000.

Porter, M., "An Algorithm for Suffix Stripping," *Program*, (14:3), 1980, pp. 130-137.

Rauber, A., and Müller-Kögler, A.,"Integrating Automatic Genre Analysis into Digital Libraries," in *ACM/IEEE Joint Conference on Digital Libraries*, 2001, pp. 1-10.

Rehm, G., "Towards Automatic Web Genre Identification," in *HICSS'02: Proceedings of the 35th Hawaii International Conference on System Sciences*, IEEE Computer Society, 2002.

Robertson, and S., Sparck-Jones, K., "Relevance Weighting of Search Terms," *American Society for Information Science*, (27:3), 1976, pp. 129-146.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., and Liu, X., "Genre Based Navigation on The Web," in *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

Salton, G., Wong, A., and Yang, C., "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, (18:11), 1975, pp. 613-620.

Santini, M., "State-of-the-Art on Automatic Genre Identification," Technical report, ITRI, University of Brighton, UK, 2004.

Santini. M., *Automatic Identification of Genre in Web Pages*, PhD thesis, University of Brighton, 2007.

Shepherd, M., and Watters, C., "The Evolution of Cybergenre," in *HICSS'98: Proceedings of the 31st Hawaii International Conference on System Sciences*, 1998.

Shepherd, M., and Watters, C., "Identifying Web Genre: Hitting a Moving Target," in *Proceedings of the WWW 2004. Workshop on Measuring Web Search Effectiveness*, 2004.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G., "Text Genre Detection Using Common Word Frequencies," in *Proceedings of the 18th Int. Conference on Computational Linguistics*, Saarbrücken, Germany, 2000.

Stein, B., and Meyer zu Eissen, S., "Topic Identification: Framework and Application," in *I-KNOW'04: Proceedings of the 4th International Conference on Knowledge Management,* Journal of Universal Computer Science, Know-Center, Graz, Austria, 2004, pp. 353-360.

Stein, B., and Meyer zu Eissen, S., "Distinguishing Topic from Genre," in *I-KNOW'06: Proceedings of the 6th International Conference on Knowledge Management,* Journal of Universal Computer Science, Springer, Graz, Austria, 2006, pp. 449-456.

Stein, B., Meyer zu Eissen, S., Gräfe, G., and Wissbrock, F., "Automating Market Forecast Summarization from Internet Data," *in Fourth International Conference on WWW/Internet,* IADIS Press, Lisbon, 2005, pp. 395-402.

University of Leipzig, Wortschatz, http://wortschatz.uni-leipzig.de, 1995.

University of Stuttgart, The Decision Tree Tagger, http://www.ims.uni-stuttgart.de, 1996.

Yoshioka, T., and Herman, G., "Coordinating Information Using Genres," CCS WP 214, Massachusetts Institute of Technology (MIT), Sloan School of Management, 2000.