Web Genre Analysis: Use Cases, Retrieval Models, and Implementation Issues

Benno Stein and Sven Meyer zu Eissen and Nedim Lipka

Bauhaus University Weimar
Faculty of Media / Media Systems
99421 Weimar
<first name>.<last name>@medien.uni-weimar.de

Abstract People who search the World Wide Web often have a multifaceted understanding of their information need: they know what they are searching for, and they know of which form or type the desired documents should be. The former aspect relates to the content of a desired document (= topic), the latter to the presentation of its content and the intended target group. Due to the different user groups and the technical means of the World Wide Web several favorite specializations of Web documents emerged: a document may contain many links (e. g. a link collection), scientific text (e. g. a research article), almost no text but pictures (e. g. an advertisement page), or a short answer to a specific question (e. g. a mail in a help forum). These examples suggest that it can be of much help if the retrieval process is capable to address a user's information need regarding to—what is called here—"genre" or "Web genre".

This chapter contributes to Web genre analysis. It presents relevant use cases, discusses existing and new technology for the construction of Web genre retrieval models, and outlines implementation aspects for a genre-enabled Web search. Special focus is put on the generalization capability of Web genre retrieval models, for which we present new evaluation measures and, for the first time, a quantitative analysis.

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.

1.1 Contributions

Section 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 3 introduces methods for mining tailored core vocabularies as well as particular statistics as a means for sensible feature quantization. Section 4 then investigates the generalization capability of our genre retrieval model and presents new kinds of experiments and analysis methods.

Section 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.[1]

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 Subsection 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. Subsection 2.1 illustrates topic-centered search technology which has been empowered by genre labeling. Subsection 2.2 shows the role of genre information in vertical search tasks. Subsection 2.3 reports on a feasibility study

[1] www.webis.de/research/projects/wega

dealing with the identification of the governing classification principle in a document collection. Longterm objective is the development of smart document classification tools. In Subsection 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.[2]

2.1 Genre-enabled Web Search

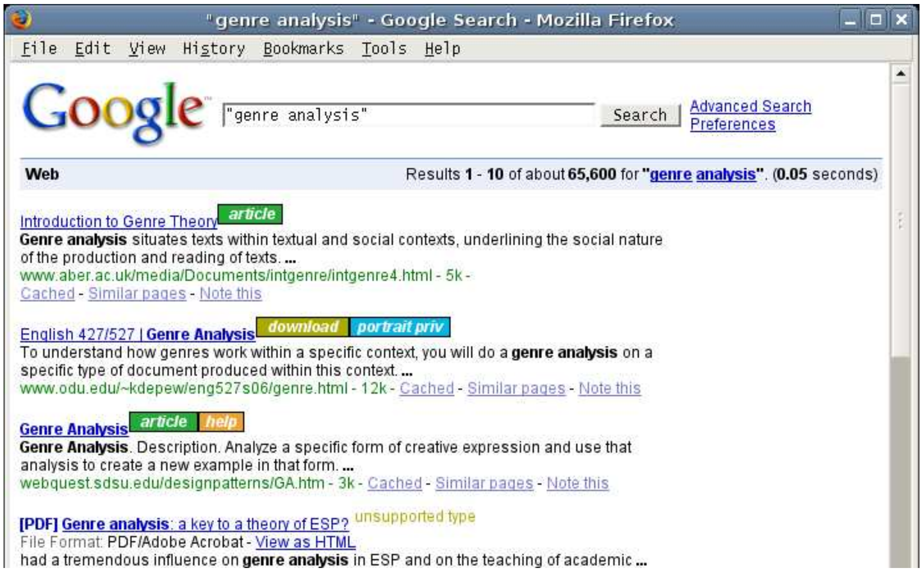
Search engines are the most influential and important applications for the World Wide Web. It stands to reason that an integration of genre-enabling technology may evolve into the most popular Web genre application. Such an integration can happen according to two different paradigms, namely filtering 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 Figure 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 1.

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:

- *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 [35]: collecting candidate documents, report filtering, time and money identification, and phrase analysis along with template filling (see Figure 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

[2] 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.



Results **1** - **10** of about **65,600** for **"genre analysis"**. (**0.05** seconds)

Introduction to Genre Theory article
Genre analysis situates texts within textual and social contexts, underlining the social nature of the production and reading of texts. **...**
www.aber.ac.uk/media/Documents/intgenre/intgenre4.html - 5k - Cached - Similar pages - Note this

English 427/527 | **Genre Analysis** download portrait priv
To understand how genres work within a specific context, you will do a **genre analysis** on a specific type of document produced within this context. **...**
www.odu.edu/~kdepew/eng527s06/genre.html - 12k - Cached - Similar pages - Note this

Genre Analysis article help
Genre Analysis. Description. Analyze a specific form of creative expression and use that analysis to create a new example in that form. **...**
webquest.sdsu.edu/designpatterns/GA.htm - 3k - Cached - Similar pages - Note this

[PDF] Genre analysis: a key to a theory of ESP? unsupported type
File Format: PDF/Adobe Acrobat - View as HTML
had a tremendous influence on **genre analysis** in ESP and on the teaching of academic **...**

Figure 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”.

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.

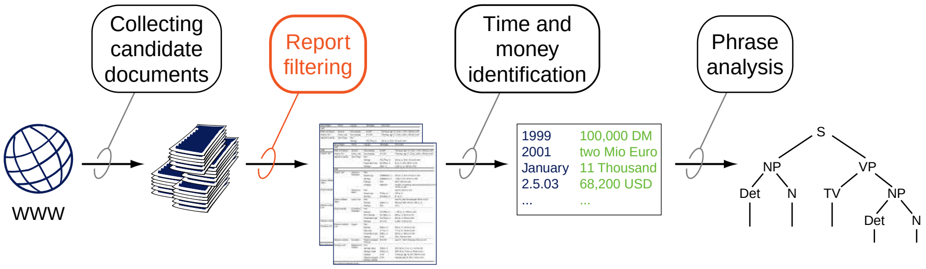


Figure 2. A four stage approach to market forecast summarization. The second step, “Report filtering”, is achieved with a genre analysis.

– *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.

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?"*

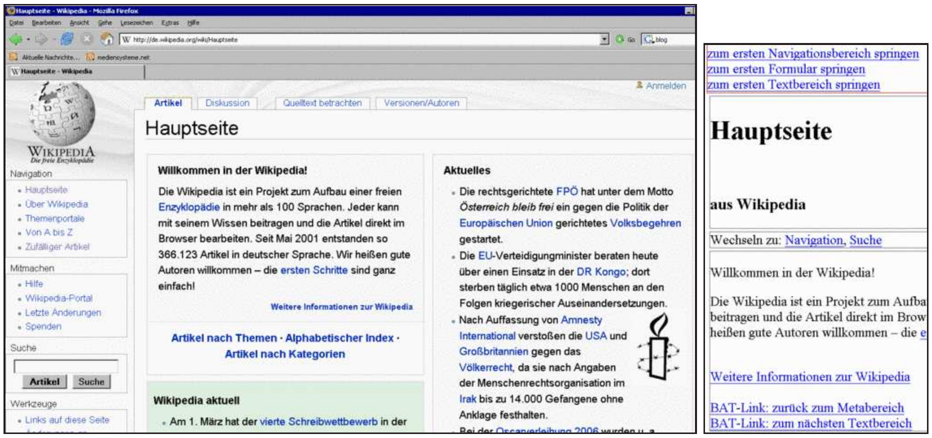
The question can be answered in the following five steps, where essentially a model fitting problem is solved. Let D be a set of documents, and let $\mathcal{C}$ be a categorization of D that is either governed by topic or by genre considerations.

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.

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



Hauptseite

Willkommen in der Wikipedia!

Die Wikipedia ist ein Projekt zum Aufbau einer freien Enzyklopädie in mehr als 100 Sprachen. Jeder kann mit seinem Wissen beitragen und die Artikel direkt im Browser bearbeiten. Seit Mai 2001 entstanden so 366.123 Artikel in deutscher Sprache. Wir heißen gute Autoren willkommen – die ersten Schritte sind ganz einfach!

Weitere Informationen zur Wikipedia

Artikel nach Themen · Alphabetischer Index · Artikel nach Kategorien

Wikipedia aktuell

- Am 1. März hat der vierte Schreibwettbewerb in der

Aktuelles

- Die rechtsgerichtete FPÖ hat unter dem Motto *Österreich bleib frei* ein gegen die Politik der Europäischen Union gerichtetes Volksbegehren gestartet.
- Die EU-Verteidigungminister beraten heute über einen Einsatz in der DR Kongo; dort sterben täglich etwa 1000 Menschen an den Folgen kriegerischer Auseinandersetzungen.
- Nach Auffassung von Amnesty International verstoßen die USA und Großbritannien gegen das Völkerrecht, da sie nach Angaben der Menschenrechtsorganisation im Irak bis zu 14.000 Gefangene ohne Anklage festhalten.
- Bei der Oscarverleihung 2006 wurden u. a.

zum ersten Navigationsbereich springen
zum ersten Formular springen
zum ersten Textbereich springen

Hauptseite

aus Wikipedia

Wechseln zu: Navigation, Suche

Willkommen in der Wikipedia!

Die Wikipedia ist ein Projekt zum Aufba
beitragen und die Artikel direkt im Brow
heißen gute Autoren willkommen – die e

Weitere Informationen zur Wikipedia

BAT-Link: zurück zum Metabereich
BAT-Link: zum nächsten Textbereich

Figure 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.

pages for visually handicapped people whose access to the World Wide Web is managed with 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 3 illustrates the use of the Blind Accessibility Tool add-on BAT that has been developed in our working group.[3]

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.

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: $\mathbf{d}$ de-

[3] www.webis.de/research/projects/bat

Author **Analysis basis**	**Web genre palette** Q	**Document representation d**
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
Lee and Myaeng (2002) 7615 documents	FAQ, home page, reportage, editorial, research article, review, product specification	genre-specific core vocabulary
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, 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 (2005) 321 documents	personal, corporate, organizational	HTML tags, phone, email, presentational tags, CSS, URL, link, script, genre-specific core vocabulary
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
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
Freund et al. (2006) 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 (2007) 1400 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 (2007) 2480 documents	[as before]	text type analysis plus a combination of layout and functionality tags

Table 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.

fines 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, $Q = \{c_1, \ldots, c_k\}$, 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 $\mathbf{d} \in \mathbf{D}$:*

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

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 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 [16, 41, 34, 1, 25, 8, 12].

Genres *on the Web* have been investigated since 1999 [4]. Table 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. were not included since the authors provided suggestions rather than a technical specification about their genre retrieval models [6, 29, 9].

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.

Bias	**Type I** **Search Space Size**		**Type II** **Search Space Exploration**	
Synonyms	exclusive bias	Rendell (1986)	preferential bias	Rendell (1986)
	representational bias	Quinlan (1993)	procedural bias	Quinlan (1993)
	restriction bias	Mitchell (1997)	search bias	Mitchell (1997)
	Strength weak <————————> strong		*Strength* weak <————————> strong	
	Generalization capability low <————————> high		*Training data sensibility* low <————————> high	

Table 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).

Insufficient Generalization Capability The authors of the approaches listed in Table 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.[4] It turned out that the retrieval precision 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

[4] 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].

authors use different names for both types of biases, Table 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 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 measured by the "stability" and "efficiency" of its construction process with respect to training samples; Subsection 4.2 introduces the theoretical means and reports on respective experiments.

High Computational Effort Table 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 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.

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 $\mathcal{C} = \{C_1, \ldots, C_k\}$ 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 [23, 17, 18, 36]. 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, $d = \{w_1, \ldots, w_m\}$, and let $W_i \subset d$ be a text window of length l in d starting with term i, say, $W_i = \{w_i, \ldots, w_{i+l-1}\}$. 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:

   $$\kappa_{T_C}(W_i) = \frac{|W_i \cap T_C|}{l}, \qquad \kappa_{T_C}(W_i) \in [0,1]$$

   The overall concentration is defined as the maximum term concentration:

   $$\kappa^*_{T_C} = \max_{W_i \subset d} \kappa_{T_C}(W_i), \qquad \kappa^*_{T_C} \in [0,1]$$

2. *Gini Coefficient.* In contrast to the κ_{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 $\nu_i = |T_C \cap W_i|$. Let A denote the area between the uniform distribution line and the Lorenz curve of the distribution of ν_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 ν_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 Kim and Ross (in this book) also propose features that consider the word distribution in a document.

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.

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.

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 $\alpha_1 : D \to \mathbf{D}_1$ and $\alpha_2 : D \to \mathbf{D}_2$ 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:*

$$agree(\gamma_{\mathcal{R}_1}, \gamma_{\mathcal{R}_2}) \;:=\; P\left(\gamma_{\mathcal{R}_1}(\mathbf{d_1}) = \gamma_{\mathcal{R}_2}(\mathbf{d_2})\right),$$

where $\mathbf{d}_1 \in \mathbf{D}_1$ and $\mathbf{d}_2 \in \mathbf{D}_2$ are representations of the same document $d \in D$.

I.e, 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 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 $\alpha : D \to \mathbf{D}$ 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:*

$$e_{\gamma_{\mathcal{R}_1}, D_2} \;:=\; P\left(\gamma_{\mathcal{R}_1}(\mathbf{d}_2) = c^*\right),$$

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 $c^* \in (Q_1 \cap Q_2)$.

I.e., 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, α, 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 α 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.

4.3 Experiments

We now discuss the generalization capability of genre retrieval models regarding the measures introduced in the Definitions 2 - 4. Our empirical analysis illustrates

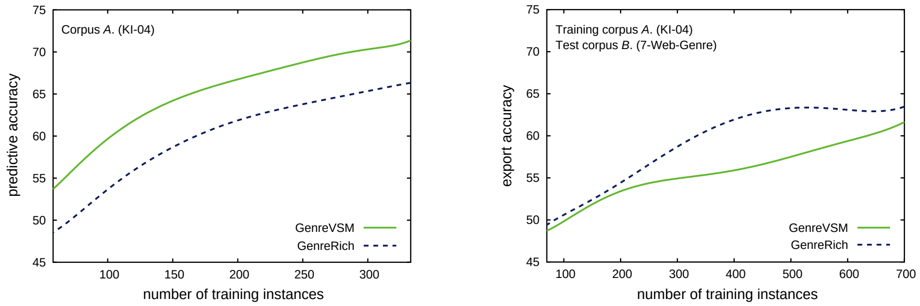


Figure 4. Predictive accuracy (left) and export accuracy (right) of the retrieval models GenreVSM and GenreRich, depending on the size of the training set, which is always drawn from corpus A (KI-04). The predictive accuracy is estimated on a test set of corpus A, while the export accuracy is estimated on a test set of corpus B (7-Webgenre collection).

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.[5] These corpora are sketched in Table 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:

1. GenreVSM. The vector space model using $tf \cdot idf$ term weighting scheme, comprising about 3500 features.
2. GenreVoc. A genre retrieval model based on the core vocabulary analysis as introduced in Section 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 Figure 4: the left plot shows the predictive accuracy of the retrieval models GenreVSM and GenreRich—trained

[5] KI-04 can be downloaded from www.webis.de/research/corpora. In the experiments the extended version of this corpus (1200 Web pages) was used.

on and applied to documents of corpus A containing 1200 documents. The right plot shows the export accuracy of these classifiers with respect to corpus B containing 600 documents, with $Q_A \cap Q_B = \{$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 Figure 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.

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 Figure 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 Figure 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:

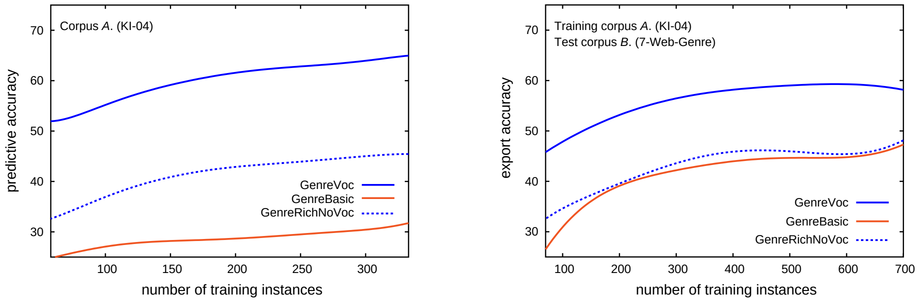


Figure 5. Predictive accuracy (left) and export accuracy (right) of the retrieval models GenreVoc, GenreRichNoVoc, and GenreBasic, using the same settings as in the experiments shown in Figure 4.

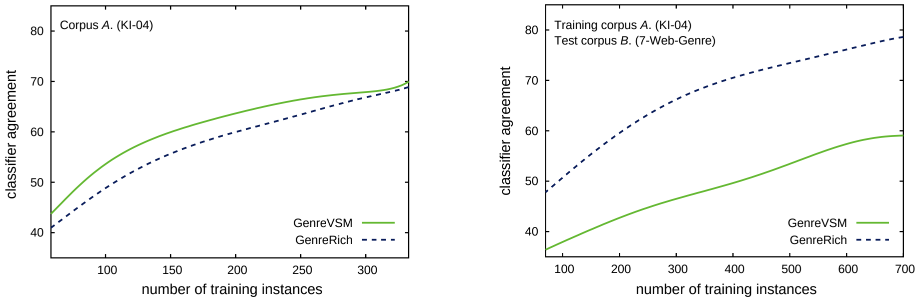


Figure 6. Classifier agreement of the retrieval models GenreVSM and GenreRich, depending on the size of the training set, which is always drawn from corpus A. In the left plot the agreement is analyzed on corpus A, while in the right plot the agreement is analyzed on corpus B.

- 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 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.

Observe in the left plot in Figure 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 Figure 6): the agreement of classifiers under the GenreRich model is much better than the agreement of classifiers under the GenreVSM model. I.e., 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.

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 1, and the classification results are integrated as genre labels into a standard result list (see Figure 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 AIsearch we employ a filter-based interface paradigm where documents are visualized as nodes of a hyperbolic graph, which can be faded in and out.[6] 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 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.

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.[7] 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 Figure 8

[6] www.webis.de/research/projects/aisearch

[7] www.webis.de/research/projects/wega

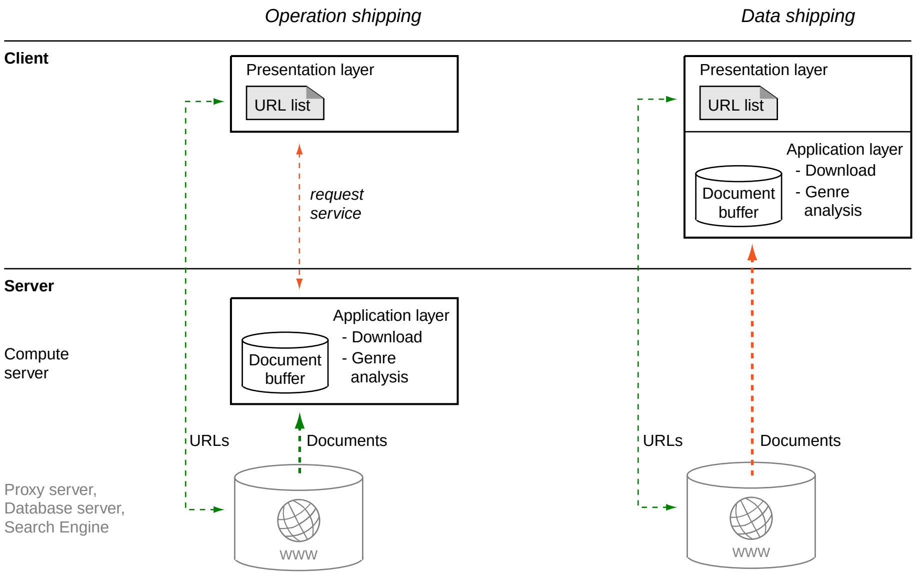


Figure 7. The genre-enabled Web search WEGA was implemented according to two different software engineering paradigms: operation shipping (left) and data shipping (right).

and Figure 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.

Characteristic	**Operation shipping**	**Data shipping**
Language	Java	JavaScript
Code size	medium	small
Runtime		
Feature computation	342 [kB/s]	134 [kB/s]
Classification	<1 [**d**/ms]	<1 [**d**/ms]

Table 3. Operation shipping versus data shipping: comparison of computational characteristics of the associated implementations.

The data shipping paradigm came within the realms of possibility with the new elements for light-weight genre retrieval models, outlined in Subsection 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 3 contrasts selected characteristics of the implementations.

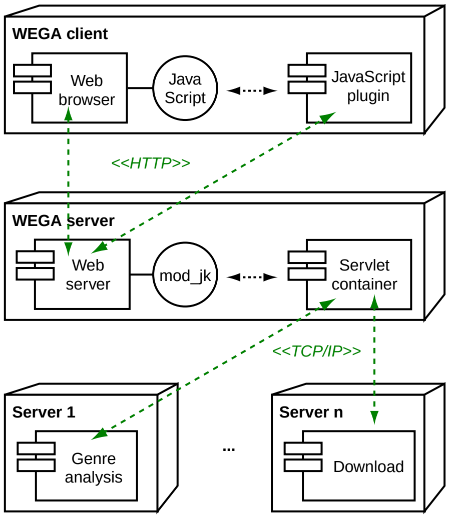


Figure 8. The deployment diagram of WEGA under the operation shipping paradigm. A servlet container provides the necessary components for the analysis service.

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 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.

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 γ 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

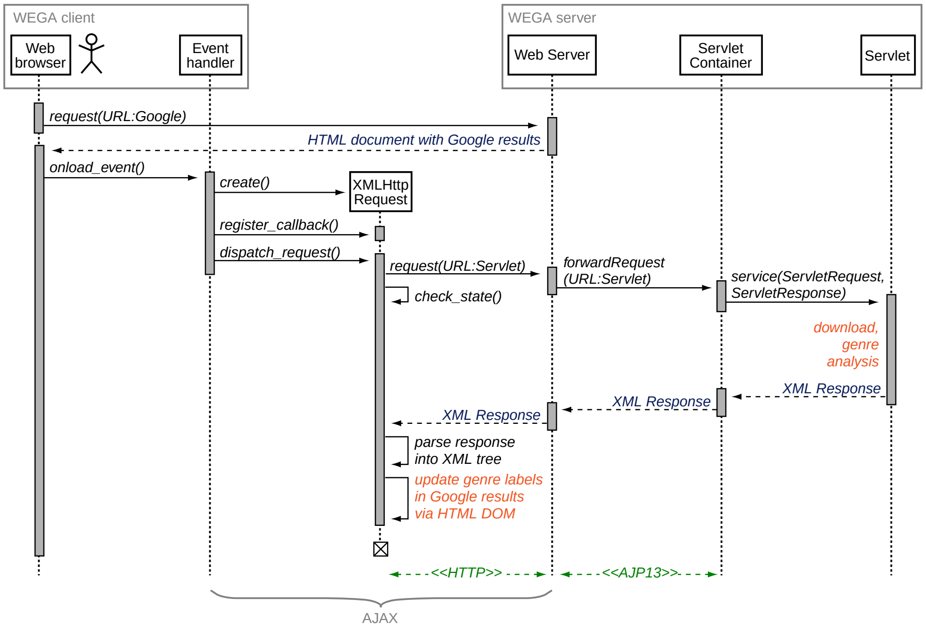


Figure 9. The sequence diagram of WEGA under the operation shipping paradigm. The middle part of the diagram show the an synchronous interaction realized with AJAX.

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.

Bibliography

[1] Pedro Antunes, Carlos J. Costa, and Joao Ferreira Dias. Applying genre analysis to ems design: The example of a small accounting firm. In *Proceedings of the Seventh International Workshop on Groupware, CRIWG 2001*, pages 74–81, Darmstadt, Germany, 2001. IEEE CS Press.

[2] Stephen D. Bay. Multivariate discretization of continuous variables for set mining. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 315–319, New York, NY, USA, 2000. ACM. ISBN 1-58113-233-6. doi: http://doi.acm.org/10.1145/347090.347159.

[3] Elisabeth Sugar Boese and Adele E. Howe. Effects of web document evolution on genre classification. In *Proceedings of the CIKM'05*. ACM Press, November 2005.

[4] Ivan Bretan, Johan Dewe, Anders Hallberg, Niklas Wolkert, and Jussi Karlgren. Web-specific genre visualization, 1999.

[5] Andrei Z. Broder. A Taxonomy of Web Search. *SIGIR Forum*, 2002.

[6] Kevin Crowston and Marie Williams. Reproduced and Emergent Genres of Communication on the World-Wide Web. *The Information Society*, 16 (3):201–216, 2000.

[7] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[8] Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. The form is the substance: Classification of genres in text. In *Proceedings of ACL Workshop on HumanLanguage Technology and Knowledge Management*, 2001.

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

[10] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and Unsupervised Discretization of Continuous Features. In A. Prieditis and S. Russell, editors, *Proceedings of the 12th International Conference on Machine Learning*, pages 194–202, Menlo Park, CA, July 1995. Morgan Kaufmann.

[11] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the International Joint Conference on Uncertainty in AI (IJCAI)*, pages 1022–1027, 1993.

[12] Aidan Finn and Nicholas Kushmerick. Learning to Classify Documents According to Genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[13] Luanne Freund, Charles L. A. Clarke, and Elaine G. Toms. Towards genre classification for IR in the workplace. In *Proceedings of the 1st international conference on Information interaction in context*, pages 30–36, New York, NY, USA, 2006. ACM. ISBN 1-59593-482-0.

[14] Jussi Karlgren and Douglass Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th. International Conference on Computational Linguistics, Coling 94*, volume II, pages 1071–1075, Kyoto, Japan, 1994.

[15] Alistair Kennedy and Michael Shepherd. Automatic Identification of Home Pages on the Web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, HICSS-38*, 2005.

[16] Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic detection of text genre. In Philip R. Cohen and Wolfgang Wahlster, editors, *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*, pages 32–38, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[17] Dawn Lawrie, W. Bruce Croft, and Arnold L. Rosenberg. 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, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 349–357, 2001.

[18] Dawn J. Lawrie and W. Bruce Croft. 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, July 28 - August 1, 2003, Toronto, Canada*, pages 457–458, 2003.

[19] Yong-Bae Lee and Sung Hyon Myaeng. 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*, pages 145–150. ACM Press, 2002. ISBN 1-58113-561-0. doi: http://doi.acm.org/10.1145/564376.564403.

[20] C. S. Lim, K. J. Lee, and G. C. Kim. Automatic Genre Detection of Web Documents. In Su, Tsujii, Lee, and Kwong, editors, *Proceedings of Natural Language Processing, IJCNLP 2004*, pages 310–319. Springer, 2005.

[21] Sven Meyer zu Eißen and Benno Stein. Genre classification of web pages: User study and feasibility analysis. In Susanne Biundo, Thom Frühwirth, and Günther Palm, editors, *KI 2004: Advances in Artificial Intelligence*, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269, Berlin Heidelberg New York, September 2004. Springer. ISBN 0302-9743.

[22] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997. ISBN 0070428077.

[23] Alexandrin Popescul and Lyle H. Ungar. Automatic Labeling of Document Clusters. `http://citeseer.nj.nec.com/popescul00automatic.html`, 2000.

[24] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[25] Andreas Rauber and Alexander Müller-Kögler. Integrating automatic genre analysis into digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 1–10, 2001.

[26] Georg Rehm. Towards Automatic Web Genre Identification. In *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS'02)*. IEEE Computer Society, January 2002.

[27] Larry A. Rendell. A General Framework for Induction and a Study of Selective Induction. *Machine Learning*, 1:177–226, 1986.

[28] S. E. Robertson and Karen Sparck-Jones. Relevance Weighting of Search Terms. *American Society for Information Science*, 27(3):129–146, 1976.

[29] Dmitri Roussinov, Kevin Crowston, Mike Nilan, Barbara Kwasnik, Jin Cai, and Xiaoyong Liu. Genre based navigation on the web. In *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

[30] Gerard Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, 1975.

[31] Marina Santini. Common criteria for genre classification: Annotation and granularity. In *Proceedings of the ECAI-Workshop TIR-06*, Riva del Garda, Italy, 2006.

[32] Marina Santini. *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton, 2007.

[33] Jürgen Sellentin. *Konzepte und Techniken der Datenversorgung für komponentenbasierte Informationssysteme*. PhD thesis, University of Stuttgart, Germany, 1999.

[34] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th Int. Conference on Computational Linguistics*, Saarbrücken, Germany, 2000.

[35] Benno Stein and Michael Busch. Density-based Cluster Algorithms in Low-dimensional and High-dimensional Applications. In Benno Stein and Sven Meyer zu Eißen, editors, *Second International Workshop on Text-Based Information Retrieval (TIR 05)*, Fachberichte Informatik, pages 45–56. Universität Koblenz-Landau, September 2005.

[36] Benno Stein and Sven Meyer zu Eißen. Topic Identification: Framework and Application. In Klaus Tochtermann and Hermann Maurer, editors, *Proceedings of the 4th International Conference on Knowledge Management (I-KNOW 04), Graz, Austria*, Journal of Universal Computer Science, pages 353–360, Graz, Austria, July 2004. Know-Center.

[37] Benno Stein and Sven Meyer zu Eissen. Distinguishing Topic from Genre. In Klaus Tochtermann and Hermann Maurer, editors, *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06), Graz*, Journal of Universal Computer Science, pages 449–456. Springer, September 2006.

[38] Benno Stein and Sven Meyer zu Eißen. Retrieval models for genre classification. *Scandinavian Journal of Information Systems (SJIS)*, 20(1): 91–117, 2008. ISSN 0905-0167.

[39] Peter D. Turney. Technical note: Bias and the quantification of stability. *Machine Learning*, 20(1-2):23–33, 1995.

[40] P. E. Utgoff. Shift of Bias for Inductive Concept Learning. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach: Volume II*, pages 107–148. Kaufmann, Los Altos, CA, 1986.

[41] Takeshi Yoshioka and George Herman. Coordinating Information Using Genres. CCS WP 214, Massachusetts Institute of Technology (MIT), Sloan School of Management, Cambridge, MA 021392, August 2000.