Genre Classification of Web Pages

—User Study and Feasibility Analysis—

Sven Meyer zu Eissen and Benno Stein
smze@upb.de, stein@upb.de

Paderborn University
Department of Computer Science
D-33095 Paderborn, Germany

Abstract Genre classification means to discriminate between documents by means of their form, their style, or their targeted audience. Put another way, genre classification is orthogonal to a classification based on the documents' contents. While most of the existing investigations of an automated genre classification are based on news articles corpora, the idea here is applied to arbitrary Web pages. We see genre classification as a powerful instrument to bring Web-based search services closer to a user's information need. This objective raises two questions:

(1) What are useful genres when searching the WWW?
(2) Can these genres be reliably identified?

The paper in hand presents results from a user study on Web genre usefulness as well as results from the construction of a genre classifier using discriminant analysis, neural network learning, and support vector machines. Particular attention is turned to a classifier's underlying feature set: Aside from the standard feature types we introduce new features that are based on word frequency classes and that can be computed with minimum computational effort. They allow us to construct compact feature sets with few elements, with which a satisfactory genre diversification is achieved. About 70% of the Web-documents are assigned to their true genre; note in this connection that no genre classification benchmark for Web pages has been published so far.

Key words: Genre Classification, Machine Learning, User Study, Information Need, Information Retrieval, WWW

1 Introduction

People who search the World Wide Web usually have a clear conception: They know what they are searching for, and they know of which form or type the search result ideally should be. The former aspect relates to the content of a found document, the latter to the presentation of its content. Basically, each delivered document constitutes an HTML file; however, in consequence of the usability and the physical nature of the World Wide Web, several favorite specializations of HTML documents emerged. A document may contain many links (e. g. a link collection), a technical text (e. g. a research article), almost no text along with several pictures (e. g. an advertisement page), or a short answer to a particular question (e. g. a mail in a help forum).

Clearly, it would be of much help if a search engine could deliver only documents of a desired—what is here called—“genre”.

The paper is organized as follows. The remaining part of this section introduces genre classification and sketches out existing work. Section 2 discusses possible genre classes and ranks them with respect to a user study. Section 3 presents standard as well as new features to make genre classification amenable to machine learning. Section 4 provides some results from different genre classification experiments. In particular we apply discriminant analysis and learning with neural networks and support vector machines to construct a genre classifier.

1.1 What Does Genre Mean?

As pointed out by Finn and Kushmerick, the term “genre” is used frequently in our culture; e. g., in connection with music, with literature, or with entertainment [7]. Roussinov et al. argue that genre can 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 [14].

Here, we are interested in the genre of HTML documents. Several definitions for document genre have been given and discussed in the past [1, 8, 9]. Common to all is that document genre and document content are orthogonal, say, documents that address the same topic can be of a different genre: “The genre describes something about what kind of document it is rather than what the document is about.” [7]. 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”, “spot news”.

1.2 What Does Genre Mean in the WWW?

In the literature on the subject there is more or less agreement on what document genre means and how different genre classes can be characterized. And, at first sight, it seems to be canonical to apply this common understanding to the World Wide Web: Certainly, “advertisement” seems to be a useful genre class, as well as “private homepage”. On second sight, however, several difficulties become apparent: Where does a presentation of a company’s mission end and where does advertisement begin? Or, does a scientific article on a private homepage belong to the same genre like a photo collection of mom’s lovely pet?

Our proposed definition of genre classes for the World Wide Web is governed by two considerations:

- Usability from the standpoint of an information miner, which can be achieved by a what we call “positive” and “negative” filtering. With the former the need for a focused search can be satisfied, while the latter simply extends the idea of spam identification to a diversified genre scheme.
- Feasibility with respect to runtime and classification performance.

The first point means that we want to support people who use the World Wide Web as a huge database to which queries are formulated.[1] The second point states that automatic genre identification shall happen on the fly, in the form of a post-processing of the results of a search engine. This aspect prevents the computation of highly sophisticated features as well as the application of a fine-grained genre scheme.[2] To get an idea which genre classes are considered useful by search engine users, we conducted a user study that is described in detail in Section 2.

1.3 Existing Work

We distinguish the existing work for computer-based genre classification with respect to the underlying corpus, say, whether it is targeted to a particular document collection—like the Brown Corpus, for example—or to the World Wide Web. In the following we outline selected papers.

Corpus-specific genre classification has been investigated among others in [9, 15, 5, 7]. The existing work can further be distinguished with respect to the interesting genre classes and the types of features that have been evaluated. Kessler et al.'s work is based on the Brown Corpus. For the characterization of genre classes they employ so-called genre facets, which are quantified by linguistic and character-level features [9]. Stamatatos et al. use discriminant analysis based on the term frequencies to identify the most discriminative terms with respect to four newspaper genre classes [15]. Dewdney et al. concentrate on different learning approaches: Naive Bayes, C4.5, and support vector machines. They employ about three hundred features including part of speech, closed-class word sets, and stemmed document terms [5]. Rehm proposes a Web genre hierarchy for academic homepages and a classifier that relies on HTML metadata, presentation related tags and unspecified linguistic features. Finn and Kushmerick distinguish between the two genres "objective" and "subjective"; they investigate three types of features sets: the document vector containing the stemmed list of a document's terms without stop-words, features from a part of speech analysis, and easily computable text statistics [7].

Genre classification and navigation related to the World Wide Web is quite new, and only very few papers have been published on this topic. Bretan et al. propose a richer representation of retrieval results in the search interface. Their approach combines content-based clustering and genre-based classification that employs simple part-of-speech information along with substantial text statistics. The features are processed with the C4.5 algorithm; however, the authors give no information about the achieved classification performance [2]. Roussinov et al. present a preliminary study to automatic genre classification: Based on an explorative user study they develop a genre scheme that is in part similar to ours and that comprises five genre groups. However, their work describes an ongoing study, and no recognition algorithm has been implemented [14]. Dimitrova et al. describe how shallow text classification techniques can be used to sort the documents according to genre dimensions. Their work describes an ongoing study, and experience with respect to the classification performance is not reported [6]. Lee

[1] There are other groups of Internet users who use the Web for amusement, for example.

[2] Crowston and Williams identified about hundred genre classes on the World Wide Web [3] .

and Myaeng define seven genre types for classifying documents from the World Wide Web. Aside from the genre “Q&A” and “Homepage” Lee and Myaeng use also the newspaper-specific genres “Reportage” and “Editorial”. The operationalized feature set is based on a list of about hundred document terms tailored to each genre class [10].

2 User Study and Genre Selection

Although we have an idea of potentially useful genres, a user study should give insights into the importance of dedicated genre classes. Moreover, it can be used as a basis to select genres for building test collections. As a matter of course, selected genres influence feature selection for automatic classification.

2.1 User Study

As discussions with colleagues on the helpfulness of different Web page genre classes were manifold, we decided to interrogate a bigger number of search engine users. We developed a questionnaire that should shed light on search engine use, usefulness of genre classification, and usefulness of genre classes. In detail, we were interested in the following points.

(1) *Frequency of Search Engine Use.* We expect that experienced search engine users have a clearer idea whether genre classification could be useful or not. We asked the interviewees how often they use search engines. Possible answers were “daily”, “once or twice a week”, “once or twice a month”, and “never”.

(2) *Typical Topics for Queries.* As already pointed out, our target audience should use the Internet not only for entertainment, but also as information source. To get an idea what the interviewees search for on the Internet, we let them specify up to 3 typical search topics.

(3) *Usefulness of Genre Classification.* With this question we wanted to figure out if genre filtering is considered as useful in general, i.e. if genre filtering helps to satisfy the user’s information need. Possible answers were “very useful”, “sometimes useful”, “not useful”, and “don’t know”.

(4) *Favored Genre Classes.* We proposed ten genre classes that we found interesting: publications/articles, scholar material, news, shops, link collections, help and FAQ, private portrayals, commercial portrayals, discussion forums, and product presentations. For each of these genres, the interviewees could specify the usefulness in terms of “very useful”, “sometimes useful”, “not useful”, and “don’t know”.

(5) *Additional Useful Genre Classes.* We also wanted to find out which additional genre classes could be interesting for the users. Therefore, a set of up to three additional genre classes could be specified and classified into “very useful” and “sometimes useful”.

(6) *Comments.* We also gave the interviewees the possibility to comment on the idea of genre classification.

To give the respondents an idea of genre classification, we gave them a 2-minute introduction to genres and their use as positive and negative information filters. As we

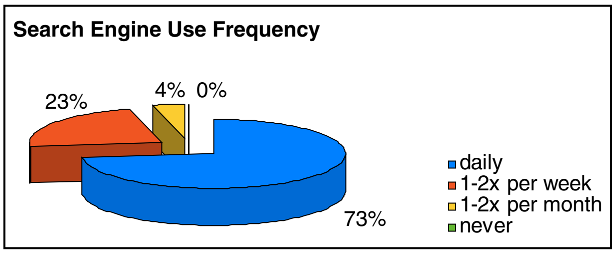

Figure 1. Frequency of search engine use. About three quarters of the interrogated students use a search engine on a daily basis.

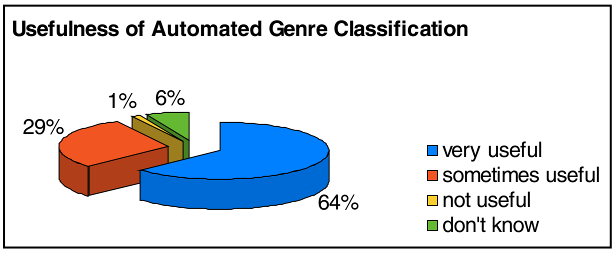

Figure 2. Usefulness of genre classification.

expect students to frequently use search engines, we asked 286 of them in our university to complete the proposed form. Figure 1 shows that we met the right audience: about three quarters of the students use search engines on a daily basis, and nearly the remaining quarter at least once a week.

The most frequently mentioned searches comprise scholar material, shopping and product information, help (discussions and troubleshooting), entertainment (music/games/films/humorous material/news), downloads, health, and programming (in this order). The fact that 64% of the students think that genre classification is very useful, and that another 29% find it sometimes useful shows that there is a strong need to post-process query results (cf. Figure 2).

To make up a ranked list of dedicated genre classes with respect to their usability, we assigned scores on the usefulness of each genre class: “very useful” scored 2 points, “sometimes useful” scored one point, “not useful” scored 0 points. We added the scores for each proposed genre and divided it by the number of interviewees that did not tick “don’t know” on that genre class. The results are depicted in Figure 3: scholar material scores best, while private portrayals were not judged as very useful by the interviewees.

Additional genres that were significantly often proposed include Web page spam and download sites. As the given comments and some given specifications of spam let conclude, spam comprises in this context (a) paid links, (b) sites that try to install dialers, and (c) sites that are only used to improve a site’s ranking in search engines. Other propositions included topics (and not genres) like pornography. The comments were encouraging and often asked for operationalization.

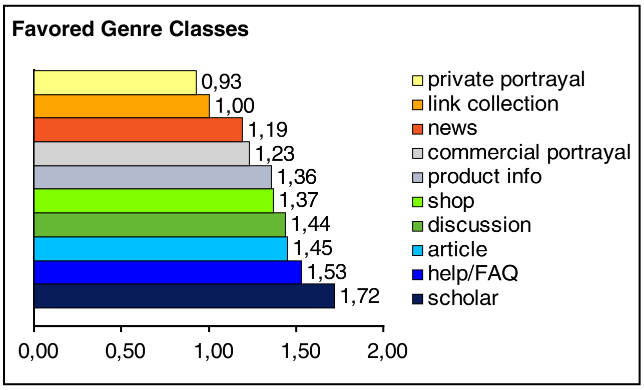

Figure 3. The favored genre classes. Higher values indicate a greater expected usefulness.

2.2 Genre Selection

An inherent problem of Web genre classification is that even humans are not able to consistently specify the genre of a given page. Take for example a tutorial on machine learning that could be either classified as scholar material or as article. In general, scholar material can be seen as a super-genre that covers help, article, and discussion pages; therefore scholar material was not chosen as a genre on its own. Another finding is that most product information sites are combined with a shopping interface, which renders a discrimination between shops and products impossible.

To cut a long story short, we finally ended up with the following eight genre classes:

(1) *Help.* All pages that provide assistance, e. g. Q&A or FAQ pages.
(2) *Article.* Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters.
(3) *Discussion.* All pages that provide forums, mailing lists or discussion boards.
(4) *Shop.* All kinds of pages whose main purpose is product information or sale.
(5) *Portrayal (non-priv).* 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.
(6) *Portrayal (priv).* Private self-portrayals, i. e., typical private homepages with informal content.
(7) *Link Collection.* Documents which consist of link lists for the main part.
(8) *Download.* Pages on which freeware, shareware, demo versions of programs etc. can be downloaded.

Although not every document can be rigorously assigned to a single class, our scheme reflects the genre assessment of many human information miners: A scientific article or a link collection, for instance, is still distinguished as such, independently of the domain holder's form of organization where the document is hosted.

Finally, it should be noted that genre classification of Web pages is at its beginning. Upcoming research may concentrate on relations between genres, or even on the development of domain-specific genre ontologies [13].

3 Features for Genre Classification

With respect to the investigated features the existing literature on genre classification falls into three groups: Classifiers that rely on a subset of a document's terms (sometimes called bag-of-words, BOW) [15, 10], classifiers that employ linguistic features along with additional features relating to text statistics [9], or both [7]. This section gives an overview of these features. In particular we introduce new features that are based on the frequency class of a word.

3.1 Word Frequency Class

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 $\mathcal{C}$ be a text corpus, and let $|\mathcal{C}|$ be the number of words in $\mathcal{C}$. Moreover, let $f(w)$ denote the frequency of a word $w \in \mathcal{C}$, and let $r(w)$ denote the rank of w in a word list of $\mathcal{C}$, which is sorted by decreasing frequency. [3]

In accordance with [12] we define the word frequency class $c(w)$ of a word $w \in \mathcal{C}$ as $\lfloor \log_2(f(w^*)/f(w)) \rfloor$, where w^* denotes the most frequently used word in $\mathcal{C}$. In the Sydney Morning Herald Corpus [4], 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 intuition to use the word frequency class as feature is the expectation that articles use a more specialized speech than e.g. shops. The complexity of speech is expected to be reflected in the average word class.

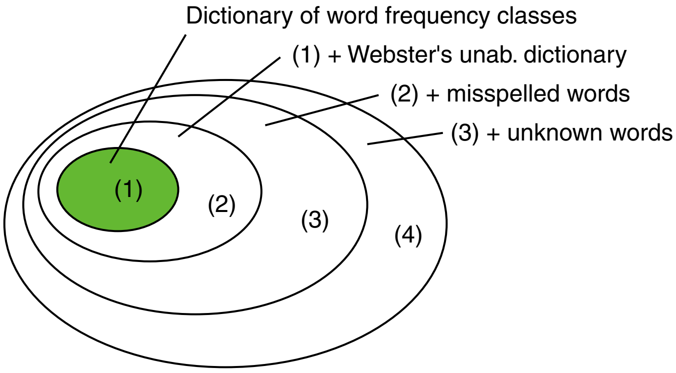
Dictionary of word frequency classes
(1) + Webster's unab. dictionary
(2) + misspelled words
(3) + unknown words
(1)
(2)
(3)
(4)

Figure 4. The figure shows the inclusion relation of the used word sets. Note that the sets (3) and (4) are only implicitly defined, by means of the Levenshtein distance and the "not-found" predicate respectively.

Based on the Sydney Morning Herald Corpus, which contains more than 38,000 articles, word frequency classes for about one hundred thousand words have been computed. This dictionary is shown as set (1) in Figure 4. The other sets in the figure evolve

[3] Zipf's law states that $r(w) \cdot f(w)$ is constant.

in a natural manner as supersets of (1): Webster's unabridged dictionary (2), the set of misspelled words (3), and the set of unknown words (4). Observe that set (3) comprises all words from the sets (1) and (2) as well as words found in the Levenshtein distance of one [11]. We use these sets to define the following features:

- average word class
- average number of misspelled words
- average number of words not found in Webster's unabridged dictionary

3.2 Syntactic Group Analysis

A syntactic group analysis yields linguistic features that relate to several words of a sentence. Such analyses quantify the use of tenses, relative clauses, main clauses, adverbial phrases, simplex noun phrases, etc. Since the identification of these features is computationally expensive, we have omitted them in our analysis. Dewdney et al., however, also include the transition in verb tense within a sentence in their analysis [5].

3.3 Part-of-Speech Analysis

Part-of-speech analysis groups 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 former do not consist of a finite number; 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 [16]. Table 1 and 2 list the actually used word classes.

3.4 Other Closed-Class Word Sets

Aside from word classes that relate to grammatical function, we have also constructed other closed-class word sets 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.

3.5 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. designate features of this type as "character-level cues" [9]; Finn and Kushmerick designate such features as "hand-crafted" [7].

Table 1. Feature set A consists of 25 features. The averages are taken with respect to the total word count within a Web document.

	Feature type	Feature set A
(1)	Presentation related	avg. # of <p> tags
(2)		avg. # of <ul> tags
(3)		avg. # of tags
(4)		avg. # of anchor links
(5)		avg. # of links same domain
(6)		avg. # of links foreign domain
(7)		avg. # of mail links
(8)		avg. # of <img> tags
(9)		avg. # of <tr> tags
(10)	Closed word sets	avg. word frequency class
(11)		avg. # of currency symbols
(12)		avg. # of help symbols
(13)		avg. # of shop symbols
(14)		avg. # of date symbols
(15)		avg. # of first names
(16)		avg. # of surnames
(17)		avg. # of words that do not appear in Webster's dictionary
(18)	Text statistics	avg. # of question marks
(19)		avg. # of letters
(20)		avg. # of digits
(21)		avg. # of dots
(22)		avg. # of semicolons
(23)		avg. # of colons
(24)		avg. # of commas
(25)		avg. # of exclamation marks

3.6 Presentation-Related Features

This 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), URL specifications, mail addresses, etc.

3.7 Constructed Feature Sets

As our concern is genre classification of search results, the classification should be done "on the fly", as a post-processing step. Since a user usually waits actively for search results, the features must be computed quickly. We propose a split of the mentioned features with respect to computational effort as follows.

(1) *Features with Low Computational Effort.* These features comprise text statistics, which can be acquired at parse time through simple counters.

Table 2. Feature set B extends feature set A by ten additional features. The averages are taken with respect to the total word count within a Web document.

	Feature type	Feature set B
(1)-(25)	identical to feature set A	
(1)	Part of speech	avg. # of nouns
(2)		avg. # of verbs
(3)		avg. # of rel. pronouns
(4)		avg. # of prepositions
(5)		avg. # of adverbs
(6)		avg. # of articles
(7)		avg. # of pronouns
(8)		avg. # of modals
(9)		avg. # of adjectives
(10)		avg. # of alphanumeric words

(2) *Features with Medium Computational Effort.* All features that are word-related and that require dictionary lookups or non-trivial parsing. These are closed-class word sets, word frequency class and presentation related features.

(3) *Features with Higher Computational Effort.* This class comprises features that rely on grammar analyses. Syntactic group analysis features and part-of-speech related features fall in this category.

It should be clear that feature category (1) is not powerful enough to discriminate between the genre classes solely. As a consequence, we built a feature set that comprises features of (1) and (2), and a feature set that makes use of all three feature classes. The Tables 1 and 2 show the details.

4 Experimental Setting and Analysis

Since no benchmark corpus is available for our concern, we compiled a new corpus with Web documents and analyzed statistical properties of the two feature sets with respect to them. We employed classifiers in the form of neural networks (MLP) and support vector machines to test the achievable classification performance. Moreover, we analyzed the classification performance for genre-specific searches and typical user groups. The following subsections outline our experiments.

4.1 Corpus Compilation

The compiled corpus of Web documents is described in Table 3. For the experiments, we used a subset of randomly drawn documents that is equally distributed over the aforementioned genres (100 documents each) and thus comprises about 800 documents.

Each element in the corpus represents a single HTML document; documents that are composed of frames and Flash elements were discarded. We then generated two distinct representations of each corpus according to the feature sets. The first feature

Table 3. Composition of the Web document corpus.

Genre	# of Documents
link collection	204
help	136
shop	169
portrayal non-priv.	171
portrayal priv.	127
articles	123
download	152
discussion	127
sum	1209

set comprises 25 attributes (see Table 1), the second feature set extends the first by additional ten features (see Table 2).

4.2 Statistical Analyses

We conducted a discriminant analysis (linear model, incremental variable selection according to Wilks Lambda, a-priori probability uniformly distributed) to get an idea of the classification performance of the selected features. Table 4 shows a confusion matrix that belongs to feature set B. The results range from acceptable to very good—articles and download pages are detected with a very high precision, and shop, discussion, link collections as well as private portrayal documents are detected with a good performance. Only non-private portrayals and help loose roughly a bigger fraction to the other genres—the diversity of non-private portrayals is immense. However, about 70% classification performance for cross-validated data (ten-fold) on a huge corpus appears still very good to us.

Table 4. Ten-fold cross-validated confusion matrix. It shows the percentage of correctly classified documents on the diagonal and summarizes the percentage of misclassified documents with respect to other genres. The average classification performance is about 70%.

	Shop	Portrayal (priv)	Portrayal (non-priv)	Article	Link Collection	Help	Discussion	Download	total
Shop	66.9%	3.0%	11.2%	5.3%	7.7%	3.0%	1.2%	1.8%	100.0%
Portrayal (priv)	0.0%	67.7%	3.1%	8.7%	15.7%	2.4%	2.4%	0.0%	100.0%
Portrayal (non-priv)	7.0%	4.1%	57.9%	5.8%	18.1%	2.9%	2.3%	1.8%	100.0%
Article	0.0%	1.6%	3.3%	81.3%	8.1%	3.3%	0.8%	1.6%	100.0%
Link Collection	0.5%	4.4%	10.8%	11.3%	67.6%	1.0%	3.4%	1.0%	100.0%
Help	2.2%	2.2%	5.1%	19.1%	10.3%	55.1%	2.2%	3.7%	100.0%
Discussion	2.4%	0.0%	3.1%	5.5%	7.1%	7.9%	68.5%	5.5%	100.0%
Download	2.0%	1.3%	5.9%	5.3%	2.5%	1.3%	2.0%	79.6%	100.0 %

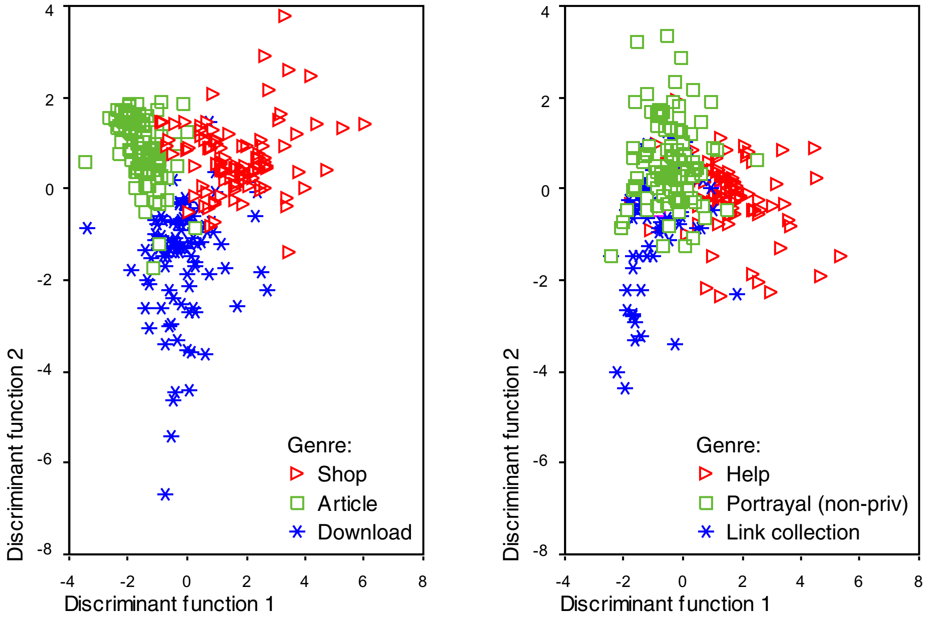

Figure 5. The left figure shows a scatter plot of the genres Shop, Article, and Download, which can be separated quite well. The right figure refers to the genres Help, Portrayal (non-priv), and Link collection, which are much harder to become identified. The underlying feature set is B.

The scatter plot in Figure 5 (left) shows that shopping sites, articles, and download pages can be separated quite well. Figure 5 (right) illustrates that help pages, non-private portrayals, and link collections overlap 5-11% each.

4.3 Classification Results

We split the corpus into test sets and training sets and, based on both feature sets, learned a classifier with both MLP neural networks and support vector machines. Table 5 comprises the classification results on the test sets. On average, for the one-against-all classification situation, which is reported in the first row of Table 5, support vector machines turned out to be better than neural networks.

It should be noted that official genre classification benchmarks for Web pages are not available. For this reason, but also to make our results reproducible for other researchers, we will make our corpus available on request.

4.4 Specialized Classifiers

Aside from general classification performance, we are interested in the question how efficient classifiers for single genre classifications and typical user profiles can be built. Assumed that a user starts several queries on the same topic, an intelligent search assistant could figure out to which predefined user profile a user probably belongs and apply

Table 5. The table shows the classification performance of specially crafted single-genre classifiers (first row) and three profile classifiers (remaining rows). The bigger boxes symbolize an aggregation of the genre classes that stand atop of them. Within a box, the upper value shows the classification performance for feature set A, while the lower value shows the performance for feature set B.

<table>
<tr><th></th><th>Shop</th><th>Portrayal (priv)</th><th>Portrayal (non-priv)</th><th>Download</th><th>Discussion</th><th>Article</th><th>Link Collection</th><th>Help</th></tr>
<tr><td>Isolated Genre Identification</td><td>79.7%
82.5%</td><td>78.0%
77.5%</td><td>72.3%
75.1%</td><td>83.4%
87.5%</td><td>76.1%
77.5%</td><td>70.7%
72.7%</td><td>63.0%
62.3%</td><td>68.2%
72.5%</td></tr>
<tr><td>Profile “Edu”</td><td colspan="5">75.5%
77.8%</td><td>72.2%
76.1%</td><td>78.7%
81.8%</td><td>60.5%
62.8%</td></tr>
<tr><td>Profile “Geek”</td><td colspan="3">54.9%
55.6%</td><td>60.6%
62.9%</td><td>62.2%
63.8%</td><td>58.4%
61.2%</td><td>59.4%
62.8%</td><td>74.3%
77.2%</td></tr>
<tr><td>Profile “Private”</td><td>73.4%
77.8%</td><td>80.1%
82.9%</td><td colspan="6">64.8%
66.7%</td></tr>
</table>

the corresponding classifier. We conducted classification experiments for the following three user profiles.

(1) *Edu.* This profile is of educational nature and comprises articles, link collections, and help sites.
(2) *Geek.* Geeks are mainly interested in downloads, discussions, articles, link collections, and help sites.
(3) *Private.* This group comprises individuals that surf the net for shopping and for reading private portrayals.

The performance of the compiled classifiers with respect to both of the feature sets is given in Table 5.

Summary and Outlook

We see genre classification as a promising concept to improve the search efficiency and to address the information need of many users that use the World Wide Web as a database. While in the past an automatic detection of genre classes has been demonstrated for newspaper corpora, there is the question whether genre classification can also be applied to the Internet.

A user study has shown the need for advanced page filtering and gave hints on the importance of dedicated Web genre classes. Taken the viewpoint of an Internet information miner we propose the following eight genres: help, article, discussion, shop, portrayals of companies and institutions, private portrayal, link collection, and download. We show that with a small set of features, which captures linguistic and presentation-related aspects, text statistics, and word frequency classes, acceptable classification results can be achieved: Our analysis reveals that about 70% of the documents are assigned correctly.

Users pointed out that there is a need for Web page spam filtering. Interesting questions here are what types of Web page spam exist and how they can be identified.

Currently we are developing a search engine that shall combine both topic search and genre search. Key questions here emerge in connection with the presentation of filtered results, say, the construction of a user friendly interface.

Bibliography

[1] D. Biber. The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. In *Computers and the Humanities*, volume 26, pages 331–345, 1992.

[2] I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. Web-specific genre visualization, 1999.

[3] K. Crowston and M. Williams. The effects of linking on genres of web documents. In *HICSS*, 1999.

[4] S. Dennis. The sydney morning herald word database. `http://www2.psy.uq.edu.au/CogPsych/Noetica/ OpenForumIssue4/SMH.html`, 1995.

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

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

[7] A. Finn and N. Kushmerick. Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[8] J. Karlgren and D. 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.

[9] B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. In P. R. Cohen and W. 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.

[10] Y.-B. Lee and S. Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proc. 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150. ACM Press, 2002. ISBN 1-58113-561-0.

[11] V. Levenshtein. Binary codes capable of correcting deletions insertions and reversals. *ISov Phys Dokl*, 6:707–710, 1966.

[12] U. of Leipzig. Wortschatz. `http://wortschatz.uni-leipzig.de`, 1995.

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

[14] D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu. Genre based navigation on the web. In *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

[15] 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.

[16] University of Stuttgart. The decision tree tagger. `http://www.ims.uni-stuttgart.de`, 1996.