

Effects of Web Document Evolution on Genre Classification

Elizabeth Sugar Boese and Adele E. Howe
Computer Science Dept., Colorado State University
Fort Collins, CO 80523 U.S.A.
<http://www.cs.colostate.edu/~boese/Research>
sugar@acm.org

ABSTRACT

The World Wide Web is a massive corpus that constantly evolves. Classification experiments usually grab a snapshot (temporally and spatially) of the Web for a corpus. In this paper, we examine the effects of page evolution on genre classification of Web pages. *Web genre* refers to the type of the page characterized by features such as style, form or presentation layout, and meta-content; Web genre can be used to tune spider crawling re-visits and inform relevance judgments for search engines. We found that pages in some genres change rarely if at all and can be used in present-day research experiments without requiring an updated version. We show that an old corpus can be used for training when testing on new Web pages, with only a marginal drop in accuracy rates on genre classification. We also show that features found to be useful in one corpus do not transfer well to other corpora with different genres.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Experimentation

Keywords

Genre, text classification, corpora

1. INTRODUCTION

Web page genre is a relatively untapped source of search information. Standard topic queries can be enhanced by specifying page genre. Instructors may look for pages with good questions or tutorials for their students. Academics may want scholarly articles on some topic. Shoppers may want reviews about specific products.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
Copyright 2005 ACM 1-59593-140-6/05/0010 ...\$5.00.

Additionally, genre may inform how often pages must be revisited by spiders. Cho and Garcia-Molina [3] estimated that 50% of the Web changes every 50 days. However, the distribution of change is domain dependent – with approximately 40% change every day in .com sites and only 1 to 2% in .edu and .gov sites [3]. This affects the efficiency of spider crawls for search engines. Systems trade off between crawling new areas on the Web and revisiting pages that have already been crawled to incorporate changes in the page. Knowing the genre of a Web page (through automated classification) may help determine whether the page is likely to change over time, and to what degree.

Studies of Web genre have focused on identifying the types of genres that are useful for the Web and the techniques for automated classification of genre (e.g., [5, 4, 7, 9]). In each case, the authors collect Web page examples of each of their genres in a corpus; the corpus is analyzed and used for evaluating classification. The rate of change in the Web calls into question whether conclusions drawn from these corpora are likely to be valid in the future or whether the corpora might be effectively used to evaluate subsequent research.

In this paper, we present an analysis of two corpora for genre classification of Web pages. Because of the advantages of having commonly accepted corpora to support evaluation and the possible utility of genre classification for tuning Web spider revisits, we focus primarily on analyzing the effect of the passage of time on classification. We seek to answer the following three questions:

1. How much do Web pages change over time within each genre?
2. Can we effectively train a genre classifier on an old corpus of Web pages?
3. How well do the features selected based on a corpus with X set of genres transfer to another corpus with different genres?

The contribution of this paper is an in-depth analysis of two corpora used previously in genre classification. We obtained versions of the pages in the corpora from specific dates (via a Web archive) and analyzed the evolution of documents within each genre, the use of old corpora for training and testing on new pages for genre classification, and the efficacy of feature transfer between corpora for genre classification.

We found that the rate of change of Web pages does indeed vary between genres, which supports genre use for optimiz-

ing spiders. We also found that, to a large extent, classification transfers from old to new corpora, thus supporting continued use of old corpora in evaluation experiments and in classification of genres, but that, not surprisingly, informative features are specific to the genre set, suggesting that new corpora need to be developed to support extended genre sets.

2. CORPORA

Studies of Web genre have used six corpora, as listed in Table 1. Four of the six corpora are considered to be old, meaning the pages were downloaded before 2000. Two corpora were identified as recent, downloaded in 2003/2004. The WebKb corpus [4] is freely available on the Internet at the CMU website. The three other old corpora are no longer available according to the authors. Crowston and Williams [6] published a list of the URLs used for their FAQ corpus, which allows it to be reconstructed, but it only contained one genre. The Meyer zu Eissen corpus was e-mailed to us by the authors. The Boese corpus was available as well, but we could not find enough old versions of its pages to support our retrospective study.

2.1 Web Knowledge Base (WebKb) Corpus

According to the headers of the WebKb corpus files, the dataset was downloaded in Nov/Dec of 1996 [4]. It contains 4,518 Web pages classified to one of six genres commonly found on computer science department websites (course, department, faculty homepage, project, staff homepage and student homepage) and an *other* category containing 3,764 Web pages. Each genre contains a set of pages from four universities (Cornell, Texas, Washington and Wisconsin) and a set from various universities.

The original WebKb corpus contains many empty Web page files and error messages, which were removed. The WebKb corpus also contained a few incorrectly formatted pages, including one page with the file:// protocol, four pages with the ftp:// protocol, and three pages without the full URL. We also removed all pages in the *other* category. After pruning errors and pages with Flash or HTML frames, the WebKb corpus sized down from 4,518 to 4,249 pages.

Classification experiments on WebKb have been done using three of the university sets in conjunction with the set from various universities for training, and testing on the held-out university set [4]. Because of the low numbers of department and staff pages, most experiments following the held-out university test set procedure had to classify based only on the four other genres.

We adopted a different methodology for classification; we use only one page from each Internet address¹ and classify on selections from five of the genres (staff still does not contain enough pages). By limiting the number of pages from the same Internet address within a genre, we prevent bias towards a particular university's style of pages or repetition of words.

2.2 Meyer zu Eissen Corpus

Meyer zu Eissen and Stein [9] developed a corpus of 1,209 pages across 8 genres. They selected genres based on a user

¹An Internet address example would be "cs.colostate.edu", as opposed to the domain name which is "colostate.edu".

study to determine the most useful genres according to students who use search engines.

We cleaned the corpus following the procedures listed previously. We also ensured unique Internet addresses within each genre, but this wasn't really an issue in this corpus. We decided to remove the "discussion" genre when we found only one discussion page available from 1999 (see next section for details of aging).

2.3 Creating Temporally Synchronized Corpora

Our experiments are designed to assess the effects of page evolution on genre classification. To control for the differences in when the corpora were downloaded, we used 1999 and 2005 as our benchmark years for analysis. We selected the most recent 1999 version available through the Internet Archive website (www.archive.org) and downloaded all new versions in April of 2005. This allowed us to see the amount of change across corpora over an equal time frame.

To obtain 1999 versions of the Web pages for both corpora, we used the Internet Archive website. Since 1996, the Internet Archive has been building an 'Internet Library' of on-line digital documents. For Web pages, the Internet Archive also maintains copies of the pages after updates through time. Not all Web pages are accessible through the archive, due to robots.txt exclusions, JavaScript issues, server-side image maps and pages not crawled by Alexa Internet. Therefore, no claims can be made concerning the exact origination date of a page, nor the accessibility of a page at a point in time before the first entry of the page in the archive.

To create a recent version of each corpus, we downloaded each URL in April 2005. We used a 30 second time-out period for downloading. We recorded the HTTP status codes and last modification dates of each page, if available. However, custom error pages designate a HTTP server response of 200 "OK", but the actual Web page returned was an error page. These are considered to be 'soft-errors' because a Web page is delivered instead of an HTTP error status code.

We cleaned each corpus by removing all empty pages, server-customized error pages (e.g., 'soft-404', permission and server errors), dynamic pages (e.g., document.write scripts), ftp pages (uses the ftp:// protocol) and re-directs (e.g., "permanently moved" and HTTP refresh meta-tags under 30 seconds). One page that used the redirect meta-tag was kept in the corpus, because it appeared to be a legitimate news refresh of the page. Pages with frames and Flash programs were also removed, since the focus this of research is text and HTML classification, not multi-media and multi-page analysis. In the process of cleaning the corpora, we discovered that most pages under 500 bytes were errors and pages under 1KB were either custom error messages or pages with very low information content. Fletcher found that Web pages under 5KB and over 200KB had a lower signal to noise ratio [8]. The only exception we found was the personal homepage, where a legitimate Web page may simply contain the person's name and contact information.

Table 2 shows the distribution of Web pages and the number of unique Internet addresses across genres for both the old and new corpora for WebKb. The list of URLs used for both the old and new WebKb corpora are available at: <http://www.cs.colostate.edu/~boese/Research/Corpora.html>. For the Meyer zu Eissen's corpus, our process resulted in downloading 327 pages from 1999 and 761 pages from 2005.

	Research	# Genres	# Pages	Download Date	Source
old	Craven et al. (1998)	6	8,282	Nov '96	Internet
	Crowston and Williams (1997)	100+	837	Feb '96	Not available
	Dewe et al. (1998)	11	1,358	early '97	Not available
	Crowston and Williams (1999)	1	70		URLs listed in publication
recent	Meyer zu Eissen and Stein (2004)	8	1,209	Jan '04	From authors
	Boese (2005)	10	343	'03-'04	From authors

Table 1: Web page corpora for genre

Genre	# Docs			# Docs without errors/flash/frame (N_w)			# Unique Internet Addresses in N_w		
	orig. '96	old '99	new '05	orig. '96	old '99	new '05	orig. '96	old '99	new '05
course	930	589	349	875	502	189	127	94	50
dept.	182	169	115	170	151	104	157	142	101
faculty	1,124	999	687	1,063	897	573	179	153	95
project	504	460	242	474	382	210	168	139	80
staff	137	111	43	126	95	37	39	33	21
student	1,641	1,382	472	1,541	1,151	327	101	81	33
Totals	4518	3710	1908	4249	3178	1440	771	642	380

Table 2: For WebKb, the number of documents for the original corpus was based on the corpus available on the Internet, including empty files. Number of 2005 documents is based on the HTTP status code from accessing the documents in April 2005. Number of 1999 documents is based on the availability of the URLs within the Internet Archive in 1999 or earlier.

3. FEATURES

Exploitation of genre requires automated classification of documents. For classification, documents are represented as sets of features. Web page features can be derived from three overlapping categories: style, form and content [1]. These three categories encompass the categories previously identified by Web genre researchers. We used the following feature types in our experiments:

Style

- Readability
- Part of speech statistics

Form

- Text statistics
- HTML analysis

Content

- Bag-of-words (BOW)
- Words in HTML title tag and URL
- Number types
- Closed-world sets
- Punctuation

A general description of each category and its feature sets are provided below. A full list of the features used in our experiments is available at:

<http://www.cs.colostate.edu/~boese/Research/Corpora.html>.

Style refers to the structure and readability of the page. Style features include: readability scales (e.g., Flesh Index score), sentence information (e.g., use of questions, passive sentences), word usage (e.g., use of conjunctions, pronouns) and sentence beginnings (e.g., prepositions, articles). A total of 45 style features were used in our experiments, derived from the output of the UNIX *style* program.

The *form* of a Web page consists of the presentation layout or format of the page. It includes text statistics (e.g., #

of paragraphs) and HTML analysis (e.g., emphasis tags, images, links). We analyzed 42 features related to form, some of which overlap with the style and content feature sets.

The *content* is composed predominately of term frequency counts (or binary counts or normalized percentages) of terms that appear in the page. A term is a group of characters separated by white-space or other non-alpha-numeric symbols. Binary count maintains either a zero for terms that do not appear or a one for terms that do appear in the page. Examples of non-alphabetic terms are phone numbers, dates and times, other numerical values, symbols and closed-word sets.

Closed-word sets are word groups that have a finite (usually small) number of related terms. For example, we analyzed the months of the year, days of the week, salutations and seasons.

We also analyzed punctuation, number of emphasized terms, terms found in the HTML title tag combined with terms from the URL, and numbers of particular HTML tags to determine presentation content such as the number of images and tables on the page. Emphasized terms are those specified in bold, headings, and/or italics as specified by HTML tags. We applied a Porter Stemmer to the words in the page and the title and URL words [11].

Our list of features is not exhaustive; there are other features commonly used in genre classification research such as: part-of-speech tagging [2], stop-word lists for term reduction and link text analysis [1, 4]. Although part-of-speech tagging has been found to be successful for genre classification, it is computationally more expensive than simple binary counts of term frequency. We used a subset of POS statistics for determining the style features of the page.

Use of stop-word lists have been used to reduce the feature space. Stop-word lists range from 100 to over a thousand words. However, some studies of genre classification

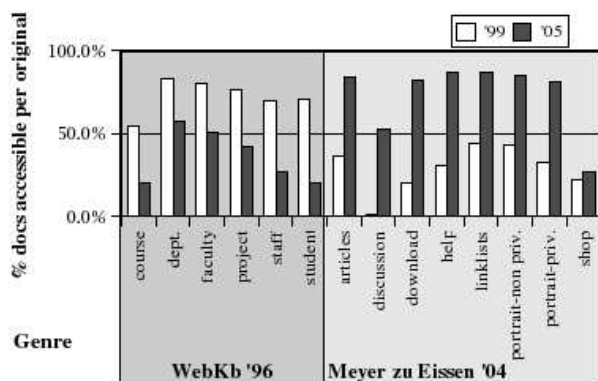


Figure 1: The accessibility of pages for the WebKb and Meyer zu Eissen corpora. Percentages are expressed with respect to the number of URLs listed in the original corpus.

have found stop words to be important for distinguishing between genres [10, 12]. For example, Nigam et al. (2000) found the word “my” to be very helpful in discriminating the personal homepage, based on Web pages available in 1999. We therefore analyzed the use of stop-words to help determine the style of the Web page text.

Link text is the text appearing in an HTML *href* tag. In our previous research [1], we achieved high classification accuracy by combining as a feature the text from a link pointing to a Web page with the text from the HTML title tag and URL components of the page. In this paper, we only considered the combination of the title text with the URL components. We could not use the link text information because it was unavailable for the Meyer zu Eissen corpus. For future work, we intend to investigate gathering link text information for all the corpora we have compiled.

4. EXPERIMENTS

We address three issues concerning the effect of time on Web pages with respect to genre: the degree of evolution of Web pages, whether we can use old corpora on new pages for genre classification, and whether selected features from one corpus could be transferred to another corpus with different genres.

4.1 Evolution

We chose to look at three measurements of change: the *accessibility* of the pages today, the *last modification dates* of pages accessed in 2005, and the degree of *page similarity* from 1999 and 2005 using the cosine metric. These measurements become increasingly complicated and informative. The first simply checks existence of the page, the second considers the amount of time between changes, and the third assesses how much the page changes with respect to genre. The following three subsections describe these in turn.

4.1.1 Accessibility

The base measure of change is the persistence of the URL, i.e., whether the page is even available at different points in time. Figure 1 shows the accessibility of URLs for the WebKb and Meyer zu Eissen corpora in 1999 and 2005.

For the WebKb corpus, while 82% of the pages originally downloaded in 1996 were still available in 1999, the percentage had dropped to 42% by 2005. Persistence varied across genre. A chi-square test comparing number of documents found in the original corpus versus number found in 2005 was highly significant ($\chi^2 = 213.59$, $p < .0001$), suggesting that accessibility is dependent on genre. Specifically, only 54% of the URLs from the course genre were still accessible in 1999 (via Internet Archive), where the other genres were all between 70% and 80%. Many of the genre differences are not surprising. For example, many course URLs contain the semester and/or year of the course - which is no longer salient after the course finishes. The student genre had the largest drop of 50%, which reflects the fact that students are expected to graduate and move on; therefore, the student homepage genre should have a high turn-over rate. The department genre had the least drop in accessibility between 1999 and 2005, which is also expected. The errors on most the department Web pages were due to a change of Internet address for the URL.

For the Meyer zu Eissen corpus, the trends are similar if somewhat more pronounced: the further from the original date, the lower is accessibility. Meyer zu Eissen was created in 2004. Thus, 1999 is a retrospective look; in 1999, only 406 (or 33%) of the URLs already existed. Given the massive expansion of the Web in the intervening years, it is hardly surprising that the sampling from 2004 would have a higher representation of new pages. In 2005, a year after its creation, most of the pages (63%) still existed.

Accessibility indicates how useful a corpus is for evaluation and training of classifiers. Additionally, crawlers can tailor their effort based on which types of pages previously indexed are more likely to still be accessible and which pages may no longer be accessible. If the crawler knows a page is from a genre with a high turn-over rate, it can visit the page more often to verify its existence or remove it from the index. This could be used to prevent search engines from reporting relevant Web pages that no longer exist.

4.1.2 Last Modification Dates

As with accessibility, last modification date provides a crude estimate of change - just when rather than how much. To assess rate of change, we queried each URL for its last modification date. Not all HTTP heading requests respond with a last modified date; we recorded as many as we could get, as indicated in the last section.

We found that some genres haven’t changed in years and others have a tendency to change more frequently. A chi-square of last modification year by genre shows that the date is statistically significantly dependent on genre ($\chi^2 = 103.2$, $p < .0001$ for WebKb and $\chi^2 = 192.8$, $p < .0001$ for Meyer zu Eissen).

Figure 2 shows a plot of last modification year for a subset of genres from the WebKb and Meyer zu Eissen corpora. The *articles* genre was selected to show that most of the articles do not change once published on the Web. *Discussion* Web pages tend to contain threads of messages that are updated as new posts are made. *Faculty* falls in between in frequency of updating. The difference in change date should be roughly predictable from a model of these observed changes dates per genre.

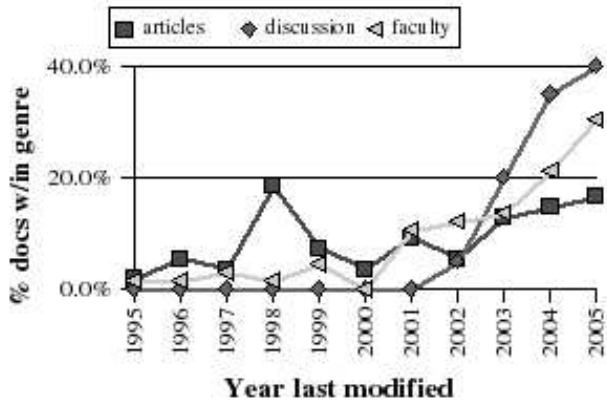


Figure 2: Last modified dates on three genres from WebKb and Meyer zu Eissen corpora.

4.1.3 Web Page Similarity

The third experiment analyzed the similarity of pages within genres over time. We used the well known cosine similarity metric:

$$\cos = \frac{o * n}{\|o\| * \|n\|}$$

where o is the feature vector for 1999 pages and n is the feature vector for 2005 pages. This can be used to determine how much change occurs in pages *within a particular genre*. The cosine measure is useful because it can scale to any number of dimensions and normalizes the feature space for more efficient processing. Table 3 displays the cosine similarity metrics for the WebKb corpus and shows that within a genre similarity is high, suggesting that most pages within a genre do not change a great deal over time.

Genre	Avg. Cosine	# docs
course	0.86	25
dept.	0.88	92
faculty	0.94	30
project	0.92	51
student	0.99	6

Table 3: Cosine values comparing 1999 and 2005 versions of WebKb pages.

4.2 Analysis of Genre Classification

Genre can enhance Web search or help tune the performance of Web spiders. Neither role can be accomplished unless genre can be automatically detected. In this section, we consider two related issues: Can an old corpus be used to train and test new pages? Are the features identified from an old corpus still informative for new data?

4.2.1 Classification Accuracy

As a baseline, we trained and tested on the 2005 corpora using stratified 10-fold cross-validation. We used WEKA’s LogitBoost algorithm for classification [13]. Classification resulted in 79.6% and 74.8% accuracy on average for WebKb and Meyer zu Eissen respectively, with a lot of variability across the genres as shown in Tables 4 and 5.

For WebKb, student homepages were most difficult to classify accurately. Most of the mis-classified student pages were classified to the faculty genre. We hypothesized that this could be partially due to the lack of pages in the student genre. To test this hypothesis, we tested classification accuracy when limiting each genre to only 31 pages, and obtained an average classification accuracy of 65%. The next step was to look at the new Web pages for students. Out of the 31 used in the classification experiments, 14 of them stated the student was currently a professor or researcher somewhere else, listed their publications and indicated when/where they finished their Ph.D. Many faculty pages contain similar information. Consequently, more careful classification or definition will be needed for student pages. An additional five contained low information content and should have been removed (no text, directory listing and server error message). The confusion shows that some pages may need to evolve (switch genres) as their owners switch their status.

We also trained classifiers on the 1999 corpora and tested them on the 2005 corpora to determine whether classifiers can be expected to continue to be accurate. We achieved an accuracy of 76.3% (as shown in Table 6), which is only slightly lower than the 79.6% using just the new corpus. Table 7 shows a similar effect on the Meyer zu Eissen corpus. Tables 8 and 9 summarize classification results for several variations on the corpora. Although the accuracy rates are similar for all versions of the Meyer zu Eissen corpus, the true positive percentages for each genre varied across training and test sets. We found that the number of documents used in training partially affected these results.

4.2.2 Persistence of Features Over Time

Automated genre classification depends on identifying salient features. In this analysis, we looked at which features were most informative at different points in time and assessed whether features continued to be informative over time.

As our first pass at the analysis, we used all the terms found for the page’s feature vector. However, we found this caused some over-fitting, where terms such as a particular university’s name would appear after feature selection methods were employed. Because of this, we performed an initial feature selection based on thresholding. Terms had to appear in at least 10% of the pages within a genre to qualify for our feature vectors. This was still problematic when the number of pages within a genre were less than 30; so we also required a minimum of 3 for the thresholding.

We analyzed the features found at the two time points under our thresholding and under feature selection as in the WEKA package. Tables 10 and 11 show the total numbers of features found at the two points in time under our thresholding and feature selection, the percentage of features that were common to both time periods, and the classification rates given all features or just the features that were common to both time periods. The relatively high rate of classification given the common features suggests that genre as defined by automated classification is quite stable over a six year period.

4.3 Feature Transfer to Different Corpora

To assess the usage of corpora for evaluation, we also examined whether there was any commonality of features across corpora. We looked at the features chosen by fea-

% classified as						accuracy: 79.6%	
course	dept	faculty	project	student	# docs		
85.7	0.0	8.2	4.1	2.0	course	49	
1.0	90.1	2.0	5.9	1.0	dept	101	
2.1	4.2	80.2	7.3	6.3	faculty	96	
3.8	12.5	8.7	73.8	1.2	project	80	
6.4	3.2	25.8	16.1	48.4	student	31	

Table 4: Stratified 10-fold cross-validation using LogitBoost classification on 2005 WebKb corpus.

% classified as							accuracy: 74.8%	
art.	down	help	links	pub	priv	shop	# docs	
71.6	1.2	4.9	3.7	7.4	11.1	0.0	articles	81
0.0	87.9	1.7	1.7	6.9	0.9	0.9	download	116
4.1	3.1	77.3	4.1	10.3	0.0	1.0	help	97
2.5	1.2	3.7	79.5	9.9	3.1	0.0	linklists	161
6.0	6.0	3.7	10.4	59.7	10.4	3.7	p-public	134
4.3	0.0	0.0	5.4	11.8	78.5	0.0	p-priv	93
0.0	2.6	2.6	7.7	28.2	0.0	59.0	shop	39

Table 5: Stratified 10-fold cross-validation using LogitBoost on 2005 Meyer zu Eissen corpus.

% classified as					accuracy: 76.3%		
course	dept	faculty	project	student	# docs		
83.7	2.0	4.1	10.2	0.0	course	49	
1.0	87.1	2.0	9.9	0.0	department	101	
1.0	6.2	75.0	11.4	6.2	faculty	96	
2.5	7.5	12.5	76.2	1.2	project	80	
6.1	6.1	24.2	27.3	36.4	student	33	

Table 6: Training on 1999 WebKb corpus with 80 pages per genre and testing on 2005 version using LogitBoost.

% classified as							accuracy: 72.0%	
art.	down	help	links	pub	priv	shop	# docs	
70.4	1.2	2.5	11.1	8.6	4.9	1.2	articles	81
1.7	83.6	1.7	0.9	9.5	0.0	2.6	download	116
3.1	4.1	55.7	11.3	15.5	2.1	8.2	help	97
0.6	0.6	0.6	87.6	8.1	0.6	1.9	linklists	161
4.5	3.7	0.0	14.2	63.4	0.7	13.4	p-public	134
3.2	0.0	2.1	8.6	20.4	61.3	4.3	p-priv	93
2.6	0.0	0.0	10.2	12.8	2.6	71.8	shop	39

Table 7: Training on 1999 Meyer zu Eissen corpus (326 instances) and testing on 2005 version (395 instances) using LogitBoost.

	all old	all new	train/test	old train	new test
course	88.2%	85.7%	83.7%	88.2%	83.7%
depart.	81.7%	90.1%	87.1%	84.5%	91.1%
faculty	78.6%	80.2%	75.0%	79.9%	79.2%
project	76.8%	73.8%	76.3%	76.1%	75.0%
student	68.8%	48.4%	36.4%	65.0%	54.5%
Method	x-valid	x-valid	train/test	x-valid	x-valid
% Accuracy	79.1%	79.6%	76.3%	79.4%	79.9%
Variance	0.059	0.060	0.064	0.059	0.061

Table 8: WebKb true positive percentage rates for each genre and the method of classification, overall accuracy based on the percent of correctly classified and variance. X-valid represents stratified cross-validation. Train/test experiment contained no duplicate URLs in training and testing corpora.

	all old	all new	train/test	old train	new test
articles	56.8%	71.6%	70.4%	56.8%	71.6%
download	72.4%	87.9%	83.6%	72.4%	87.9%
help	77.8%	77.3%	55.7%	77.8%	77.3%
linklists	72.0%	79.5%	87.6%	72.0%	79.5%
p-priv	69.4%	59.7%	63.4%	69.4%	59.7%
p-public	88.9%	78.5%	61.3%	88.9%	78.5%
shop	47.1%	59.0%	71.8%	47.1%	59.0%
Method	x-valid	x-valid	train/test	x-valid	x-valid
% Accuracy	69.6%	74.8%	72.0%	69.6%	74.8%
Variance	0.062	0.050	0.058	0.062	0.050

Table 9: Meyer zu Eissen true positive percentage rates for each genre and the method of classification, overall accuracy based on the percent of correctly classified and variance. X-valid represents stratified cross-validation. Train/test experiment contained no duplicate URLs in both training and testing corpora.

ture selection for the WebKb and Meyer zu Eissen corpora. We identified the common features for each corpora based on the features chosen during feature selection for both the 1999 and 2005 corpora. We found 33 features common to the old and new WebKb corpora, and 19 features common to the old and new Meyer zu Eissen corpora. Table 12 shows the common features in both corpora; only one feature overlapped between the two corpora with different genres. Note that most features are from BOW.

Common features in WebKb '99/'05			Common features in Meyer zu Eissen '99/'05	
assign	interest	syllabu	element	download
california	lectur	undergradu	your	window
cours	me	webmast	instal	question/q&a
depart	my	Lix	is	SMOG
faculti	our	TU:comput	link	% passive sent.
fax	ph	TU:cs	or	TU:com
grade	[number]	TU:depart	far	TU:download
group	professor	TU:scienc	scienc	HTML:# links
he	public	HTML:# images	site	HTML:# TR tags
homework	research	HTML:# links		
instructor	staff	HTML:# tables		

Table 12: Common features selected using WEKA’s Correlation-based Feature Subset Selection algorithm. Terms listed are found in both the 1999 and 2005 subsets for the corpus. TU = word appearing in either the HTML title tag or the URL (or both).

A closer look at the data shows, not surprisingly, that the BOW terms are not transferable. However, the style readability metrics are all somewhat similar, and the fact that both WebKb and Meyer zu Eissen corpora selected a readability measure and number of links could be considered transferable. Additionally, the number of table tags (selected in WebKb) and the number of tr tags designating

a table row within a table (selected in Meyer zu Eissen corpus) are fairly similar as well. Predominately, however, the features required for genre classification tend to be domain specific.

5. FUTURE WORK

In this project, we have looked at primarily temporal factors that affect the value of corpora for genre classification. In future work, we intend to expand our analysis to further identify critical factors that should be accounted for in the construction of corpora. For example, many other features may be salient as in [1], as well as additional closed-word sets such as “Ph.D.” and “M.S.”.

Feature selection and thresholding clearly exert an influence and should be further explored. Although we found 5% too small for thresholding, 10% was chosen but problematic on small corpora. A formula is needed that appropriately thresholds features selected based on the size of the corpus.

Another issue concerns which pages should be disregarded. At present, we considered discarding pages under a particular size. However, for an expanded genre set such thresholding may be inappropriate.

Finally, the genre corpora include a restricted set of genres, pages and judgments about what constitutes a genre. We intend to develop a larger corpus with a wider set of genres and test the subjectivity of the assignment of genres to pages. This might also permit assigning a single page to multiple genres.

WebKb	Number of Features			% of common features		Classification		
						w/all features		w/common features
	1999	2005	in both	1999	2005	1999	2005	2005
Thresholding	627	737	555	88.6%	75.3%	75.7%	71.4%	70.2%
Feature selection	53	63	33	75.0%	52.4%	79.1%	79.6%	74.5%

Table 10: Feature analysis over time for WebKb corpus

Meyer zu Eissen	Number of Features			% of common features		Classification		
						w/all features		w/common features
	1999	2005	in both	1999	2005	1999	2005	2005
Thresholding	5,156	4,456	3,975	77.0%	89.3%	69.7%	73.3%	71.7%
Feature selection	51	53	19	36.5%	35.2%	69.6%	74.8%	73.3%

Table 11: Feature analysis over time for Meyer zu Eissen corpus

6. CONCLUSIONS

Genre classification of Web pages is still a relatively new and unexplored domain. Genre knowledge can be used for improving the efficiency of a spider updating its knowledge base and for query relevance in search engines.

Genre classification of Web pages depends on how the corpus is organized. Using unique Internet addresses within each genre can help prevent bias towards particular website styles, and testing should be done on Internet addresses that have not yet been seen by the classifier. Proper thresholding for feature selection is required to prevent over-fitting. Classifying very small files with little information content (e.g., links and images only) is difficult and may not achieve the desired goals of relevance to users' information needs.

We showed how one genre could evolve over time to another genre (student to faculty homepage). One specific metric isn't enough for validation; for example, students who become faculty members have a high cosine similarity between old and new versions of a Web page yet they've evolved into a different genre.

Predominately, the features required for genre classification tend to be domain specific. This is problematic for attempting to use the features from one corpus in other corpora.

Most importantly, we have demonstrated that although the Web, in general, changes rapidly, the definition of genre, as manifest in classifiers, appears remarkably stable. Genre can be used for both tuning spiders and for enhancing search. Conclusions drawn from evaluation on old corpora are likely still to hold. Classifiers trained on old documents in corpora still work well on updated versions of the pages.

7. ACKNOWLEDGMENTS

We would like to thank Meyer zu Eissen and Stein for providing their corpus to us. We would also like to thank the anonymous reviewers for their comments and suggestions.

8. REFERENCES

- [1] E. S. Boese. Stereotyping the web: Genre classification of web documents. Master's thesis, Colorado State University, Fort Collins, CO, 2005.
- [2] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of ANLP '92, 3rd Conference on Applied Natural Language Processing*, Trento, Italy, 1992.
- [3] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In C. San Francisco, editor, *VLDB '00: Proceedings of the 26th International Con.*, 2000.
- [4] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the 15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, Madison, W, 1998.
- [5] K. Crowston and M. Williams. Reproduced and emergent genres of communication on the world-wide web. In *Proceedings of HICSS-97*, Kilea, HI, 1997.
- [6] K. Crowston and M. Williams. The effects of linking on genres of web documents. In *Proceedings of 32nd Annual Hawaii International Conference on System Sciences (HICSS-99)*, Kilea, HI, 1999.
- [7] J. Dewe, J. Karlgren, and I. Bretan. Assembling a balanced corpus from the internet. In *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Copenhagen, 1998.
- [8] W. Fletcher. Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton, editors, *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*. American Association for Applied Corpus Linguistics, 2004.
- [9] S. Meyer zu Eissen and B. Stein. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004)*, Ulm, Germany, 2004.
- [10] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- [11] M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.
- [12] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics*, Luxembourg, 2000.
- [13] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 2000.