CLASSIFYING WEB PAGES BY GENRE
A Distance Function Approach

Jane E. Mason, Michael Shepherd and Jack Duffy
Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

Keywords: Information retrieval, Web genre classification, Web page genres, Web page representation, *n*-gram analysis.

Abstract: The research reported in this paper is part of a larger project on the automatic classification of Web pages by their genres, using a distance function classification model. In this paper, we investigate the effect of several commonly used data preprocessing steps, explore the use of byte and word *n*-grams, and test our classification model on three Web page data sets. Our approach is to represent each Web page by a profile that is composed of fixed-length *n*-grams and their normalized frequencies within the document. Similarly, each of the genres in a data set is represented by a profile that is constructed by combining the *n*-gram profiles for each exemplar Web page of that genre, forming a centroid profile for each Web page genre. We use a distance function approach to determine the similarity between two profiles, assigning each Web page the label of the genre profile to which its profile is most similar. Our results compare very favorably to those of other researchers.

1 INTRODUCTION

The extraordinary growth in both the size and popularity of the World Wide Web has generated a growing interest in the identification of Web page genres, and in the use of these genres to classify Web pages. Web page genre classification is a potentially powerful tool for filtering the results of online searches.

The research reported in this paper is part of a larger project on the automatic classification of Web pages by their genres, using a distance function classification model. The goal of this phase of the research is threefold. First, we want to investigate the effects of typical data preprocessing steps when using an *n*-gram approach to Web page representation: are preprocessing steps beneficial or detrimental to the classification accuracy achieved by our model? Second, we want to do an initial exploration as to whether the strength of our classification model lies in the byte *n*-gram representation of the Web pages and Web genres, or in the architecture of the model itself; experiments with both byte *n*-grams and word 1-grams provide valuable insight. Finally, we want to evaluate our Web page genre classifier by comparing our results to those of other researchers on the same or similar data sets, and determine whether our approach warrants further investigation.

The remainder of the paper proceeds as follows. Section 2 gives an overview of related work, and Section 3 reviews our methodology, including a description of the classification model, distance function, and data sets. Section 4 describes the experiments and discusses the results, while Section 5 presents our conclusions and the direction of our future work. Tables and figures are given in the Appendix.

2 RELATED WORK

In order to classify Web pages by genre, it is necessary to identify features that effectively characterize each Web page and genre. Analyzing genre in academic and research settings, (Swales, 1990) noted that members of a particular genre tend to exhibit similar patterns of content, style, structure, and intended audience; analyzing genre in an online setting, (Shepherd and Watters, 1998) proposed a similar characterization of cybergenres, using the properties of content, form, and functionality. (Crowston and Kwasnik, 2004) also explore genre as a multidimensional phenomenon. They suggest the use of a *facetted* approach to Web page representation in which models are built with facets representing the Web page from many conceptual perspectives, such as content, structure, language, and source. Many different combinations and representations of features

E. Mason J., Shepherd M. and Duffy J.
CLASSIFYING WEB PAGES BY GENRE - A Distance Function Approach.
DOI: 10.5220/0001837706460653
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

or facets have been tested by researchers. For example, (Asirvatham and Ravi, 2001) obtained encouraging results using feature sets with components based on content, form, and functionality to classify Web pages. (Lim et al., 2005) used five distinct feature sets; they found that the best combination of feature sets included the URL, HTML tags, token information, most frequently used words and punctuation, and chunks (multi-word expressions). (Dong et al., 2008) also examined the use of feature sets with combinations of content, form, and functionality; their results indicated that using combinations of the attributes content, form, and functionality always gave better results than using only one of the attributes.

The representations of Web pages used in the genre classification task are typically based on those used in text classification, although they may be augmented with information unique to Web pages, such as HTML tags. For example, (Rehm, 2002) uses linguistic features combined with HTML meta data and presentation related tags. (Santini, 2007) developed feature sets that include different combinations and numbers of features, such as HTML tags, part-of-speech tags, common word frequencies, genre-specific facets, and other attributes. (Meyer zu Eissen and Stein, 2004) combine genre-specific vocabulary and closed-class word sets with text statistics, part-of-speech information, and HTML tags, while (Boese and Howe, 2005) use a bag of words representation augmented with other information that includes HTML tags, and URL information. See (Stein and Meyer zu Eissen, 2008) for a detailed chronological overview of the document representations used for genre classification on Web-based corpora.

In our research, we represent Web pages using *n*-gram profiles. Each *n*-gram can be thought of as the contents of a fixed-size sliding window as it is moved through the text. Since the research of (Shannon, 1948), *n*-grams have been widely used in natural language processing and statistical analysis. Closely related to our *n*-gram techniques is the work on *n*-gram based text classification by (Cavnar and Trenkle, 1994), and the work on authorship attribution by (Kešelj et al., 2003). In each of those cases, the document is represented by a profile of character *n*-grams, and the distance between two documents is determined using a distance measure. We note, however, that our method of creating genre profiles is very different, as explained in Section 3.1. (Houvardas and Stamatatos, 2006), working on the problem of authorship attribution, proposed a selection technique for variable-length character *n*-grams in which each *n*-gram is compared with similar *n*-grams in the feature set and the most important of them is kept. After feature selection, a support vector machine (SVM) is then trained on this reduced feature set, and then applied to the test set. (Kanaris and Stamatatos, 2007) use this new feature selection technique for variable-length character *n*-grams and apply it to the problem of Web page genre identification. They test two models: the first uses feature sets of variable-length character *n*-grams from the textual content of each Web page, whereas the second model augments the first model with structural information about the most frequent HTML tags.

3 METHODOLOGY

3.1 Classification Model

We partition each data set into a training and test set. Our approach is to represent each Web page in the test set using a profile consisting of the L most frequent fixed-length *n*-grams and their normalized frequencies within the document. These *n*-gram profiles are produced using the Perl package `Text:Ngrams`[1]. Although our main focus is on using byte *n*-gram representations of the Web pages, we also experiment with the use of word *n*-grams of length 1. The byte *n*-grams are raw character *n*-grams in which no bytes are ignored, including the whitespace characters, thus some of the structure of a document is captured using byte *n*-grams. Character *n*-grams, on the other hand, use letters only and typically ignore digits, punctuation, and whitespace. Word *n*-grams are word sequences of length *n*.

Each genre in a data set is also represented using a profile of *n*-grams and their normalized frequencies. A genre profile is constructed by combining the *n*-gram profiles for each Web page of that genre (from the training set) to form a centroid genre profile. Each genre profile initially consists of the L most frequent *n*-grams from each of the Web pages (of that particular genre) in the training set. Because of the large number of unique *n*-grams, combining the Web page profiles to create a genre profile typically results in genre profiles much larger than the Web page profiles. Thus, once all of the initial genre profiles have been created, the *n*-grams in each genre profile are sorted by frequency, and the genre profiles are then truncated to the size of the smallest of the genre profiles. Our use of centroid genre profiles differs from the work of (Kešelj et al., 2003) on authorship attribution, in which *n*-gram author profiles are constructed in the same manner as *n*-gram document profiles.

[1] `http://users.cs.dal.ca/~vlado/srcperl/Ngrams/`

In order to assign a genre label to a Web page from the test set, we construct an n-gram profile for the Web page and compare this profile to the n-gram profile for each genre in the data set. The Web page is assigned the label of the genre profile to which its profile is closest (most similar), according to a distance measure. For our experiments, the distance between two n-gram profiles is computed using the formula suggested by (Kešelj et al., 2003). The distance (dissimilarity) between two profiles is defined as

$$d(P_1,P_2) = \sum_{m\in(P_1\cup P_2)} \left(\frac{2\cdot(f_1(m)-f_2(m))}{f_1(m)+f_2(m)} \right)^2, \quad (1)$$

where $f_1(m)$ and $f_2(m)$ are the frequencies of n-gram m in the profiles P_1 and P_2 respectively.

3.2 Data Sets

In order to evaluate our results by comparing them with those of other researchers, we use established data sets for which published results are available. Note that the unit of analysis for each of the data sets is the individual Web page, and each Web page is labeled with one, and only one, genre label. The 7-Genre and KI-04 data sets are available online[2].

The 7-Genre data set, constructed by and described in (Santini, 2007), contains 1400 English Web pages, and is evenly balanced with 200 Web pages in each of seven genres. These genres are BLOG, ESHOP, FAQ, ONLINE NEWSPAPER FRONT PAGE, LISTING, PERSONAL HOME PAGE, and SEARCH PAGE. The granularity of the collection is consistent, with the exception of the LISTING genre, which can be decomposed into the subgenres CHECKLIST, HOTLIST, SITEMAP, and TABLE.

In order to compare our results with those of (Dong et al., 2008), we also use a subset of the 7-Genre data set, which we call the 4-Genre data set. This data set consists of the genres used by Dong et al. (ESHOP, FAQ, ONLINE NEWSPAPER FRONT PAGE, and PERSONAL HOME PAGE), however we note that Dong et al. used only 170 Web pages in each genre, and that for the PERSONAL HOME PAGE genre, they did not use the pages from the 7-Genre collection. We do not have access to the exact data set they used, however by using the 4-Genre data set, we believe we can make a fair comparison of our results.

The KI-04 corpus, constructed by (Meyer zu Eissen and Stein, 2004) has eight genres suggested by participants in a user study on the usefulness of Web page genres. The original corpus includes some empty Web pages, and so we follow the lead of (Santini, 2007) in using the 1205 non-empty pages. The number of Web pages per genre ranges from 126 to 205; see Table 1 in the Appendix for details. The Web pages in this data set include supplementary tagged information, such as the title, genre, and a plain text summary, which must be removed prior to processing.

[2] http://www.itri.brighton.ac.uk/~Marina.Santini/#Download

4 RESULTS AND DISCUSSION

4.1 Experiments

The experiments reported here are performed on the 4-Genre, 7-Genre, and KI-04 data sets. The experiments on each data set are run using 10-fold cross validation. This provides robustness against overfitting and gives additional strength to the statistical analysis. The results for all of the iterations are averaged to give the final results. We use classification accuracy as the evaluation metric.

The parameters varied in these experiments are the number of n-grams used in the Web page profiles, the type of the n-gram, the length of the n-gram, and the preprocessing performed on the data set. The number of n-grams used in the Web page profiles ranges from 5 to 5000, using increments of 5 from 5 to 50, increments of 50 from 50 to 1000, and increments of 1000 from 1000 to 5000. The experiments are run with word n-grams of length 1, and with byte n-grams of lengths 5 to 7; the length for the byte n-grams was chosen based on the results of our earlier work (Mason et al., 2009). In each case, the n-grams in each Web page profile are the L most frequent n-grams in the Web page.

In order to investigate the effect of some common preprocessing steps, we experiment with different levels of preprocessing on the data sets. These levels are no preprocessing, removing only JavaScript code, removing both HTML tags and JavaScript code, and removing stopwords as well as HTML tags and JavaScript code. The first three levels of preprocessing are tested on byte n-grams; the third level is also tested on word 1-grams. The final preprocessing level, which includes the removal of stopwords, is only tested using word 1-grams. In each case, the supplementary tagged information has been removed from the KI-04 data set as a pre-preprocessing step.

4.2 Classification Results

Tables 3-5 in the Appendix give summaries of the best results for each data set for word 1-grams and byte 5-

grams, 6-grams, and 7-grams, at each level of preprocessing. Tables 6-8 give confusion matrices for the best results for each data set. The tables show the percentage of correctly classified Web pages on the diagonal and summarize the percentage of misclassified Web pages with respect to other genres. Figures 1-3 show the accuracy for Web page profiles from size 5 to 5000 for word 1-grams and byte 6-grams at each level of preprocessing. Results for byte 5-grams and 7-grams follow a similar pattern to those of byte 6-grams.

4.2.1 Effect of Preprocessing Steps

As shown in Tables 3-5 in the Appendix, we did not find a particular level of preprocessing to be superior on every data set, however when using Web page profiles composed of byte *n*-grams, removing the JavaScript code gave slightly better results on the 4-Genre and 7-Genre data sets than not performing any preprocessing, but the difference is not statistically significant on the 7-Genre data set ($p = 0.68$). Aside from this case, there is a statistically significant difference between doing no processing, removing Javascript code, and removing both HTML and JavaScript code on each of the data sets. Figures 1-3 show that when using Web page profiles composed of fewer than 1000 byte *n*-grams, removing both the HTML tags and JavaScript code from the data sets clearly gives better classification accuracy than doing no processing or removing only the JavaScript code. Not surprisingly, there is a tradeoff between data preprocessing and Web page profile sizes. Note that the percentage of Web pages in the 4-Genre, 7-Genre, and KI-04 data sets that contain JavaScript code before preprocessing is 82%, 78%, and 53% respectively.

When using Web page profiles composed of word 1-grams, removing stopwords has a beneficial effect on the 7-Genre and KI-04 data sets, but not on the smaller 4-Genre data set. Statistically, there is a significant difference between these two levels of preprocessing on each data set ($p < 0.02$), but Figures 1-3 show that the effect on the classification accuracy is appreciable only on the 4-Genre data set.

4.2.2 Word vs. Byte *n*-grams

Byte *n*-grams are raw character *n*-grams in which no bytes are ignored, thus some of the structure of a document is captured when byte *n*-grams are used to construct Web page profiles. For this reason, we expected that using byte *n*-grams would achieve better classification results than using word *n*-grams, which use only the textual content of the Web page. This is the case for the 7-Genre data set, however for the 4-Genre and KI-04 data sets, the best classification accuracy was achieved using Web page profiles built from the 150 most frequent word 1-grams in each Web page. Whereas the best results using byte *n*-grams are found using between 700 and 5000 *n*-grams in each Web page profile, word 1-grams give the best results at between 100 and 350 1-grams; using larger word 1-gram profiles clearly degrades performance. Because using word 1-grams allows for substantially smaller Web page profiles than does using byte *n*-grams, the use of word 1-grams warrants further investigation.

4.2.3 Comparison of Results

Table 2 in the Appendix gives a comparison of our best results with those of other researchers, as reported in the literature, using the same or similar data sets. The other researchers have used a variety of different features and machine learning methods. On the 4-Genre data set, we compare our results with those of (Dong et al., 2008), who use a Naive Bayes classifier. Their best accuracy is 97.0%, using a feature set size of 100 that combines the attributes of content and form. Our best accuracy on the 4-Genre data set is 99.5%, using Web page profiles constructed of the most frequent 150 word 1-grams. In this case the data set had the HTML tags and JavaScript code removed as preprocessing steps. We note that Dong et al. used a slightly different version of the data set that included noise pages.

On the 7-Genre data set, (Santini, 2007), who uses an SVM classifier with feature sets that include HTML tags, part-of-speech tags, and genre-specific facets, achieves a best accuracy of 90.6%. (Kanaris and Stamatatos, 2007) also use an SVM classifier, but their feature set includes a combination of variable-length character *n*-grams and structural information from HTML tags; their best accuracy on this data set is 96.5%. Our best accuracy on the 7-Genre data set is 94.5%, using Web page profiles constructed of the most frequent 4000 byte 5-grams, after removing the JavaScript code as a preprocessing step.

Using the KI-04 data set, (Santini, 2007) achieves an accuracy of 68.9% while (Meyer zu Eissen and Stein, 2004) get an accuracy of 70.0%. They use an SVM classifier on a balanced subset of 800 Web pages from the data set. (Boese and Howe, 2005) use WEKA's LogitBoost algorithm for classification on a subset of the KI-04 data set, and achieve an accuracy of 74.8%. (Kanaris and Stamatatos, 2007) report a best accuracy of 84.1%. Our best accuracy on the KI-04 data set is 97.6%. This was achieved using Web page profiles constructed of the most frequent 150 word 1-grams, after removing stopwords, HTML tags, and JavaScript code from the data set. This accu-

racy was also achieved using Web page profiles constructed of the most frequent 700 byte 5-grams; in this case the data set had both the HTML tags and the JavaScript code removed during preprocessing.

5 CONCLUSIONS

The research reported in this paper is on the automatic classification of Web pages by their genres, using a distance function classification model. The goal of this phase of the research was threefold. First, we wanted to investigate the effects of typical data preprocessing steps when using an *n*-gram approach to Web page representation. As discussed in Section 4.2.1, we did not find a particular level of preprocessing to be superior on every data set. Our model is able to achieve very high classification accuracy using byte *n*-gram Web page profiles even when no preprocessing of the data set is performed.

Second, we wanted to do an initial exploration of whether the strength of our classification model lies in the byte *n*-gram representation of the Web pages and genres, or in the architecture of the model itself. We achieved high classification accuracy on each data set using Web page profiles constructed of the most frequent *L* word 1-grams in each Web page, indicating that the strength of the model comes less from the byte *n*-gram approach, and more from the model's architecture. Further investigation is needed.

Finally, we wanted to evaluate our Web page genre classifier by comparing the results of our experiments to the results of other researchers, as reported in the literature, on the same or similar data sets. As discussed in Section 4.2.3, our classification accuracy is either in the same range as, or higher than results reported by other researchers who are using different features and different machine learning methods. These results are very encouraging, and indicate that our model warrants further investigation.

The major contribution of this research is to show that our distance function classification model is a viable approach to the classification of Web pages by genre, and that achieving high classification accuracy with this model is not dependent on the use of byte *n*-grams, on the level of preprocessing of the data, or on the use of a particular data set. Future work will include refining the model and expanding the scope of the work by using more challenging data sets, including highly unbalanced and multi-labeled data sets.

REFERENCES

Asirvatham, A. and Ravi, K. (2001). Web page classification based on document structure. *IEEE Nat. Conv.*

Boese, E. and Howe, A. (2005). Effects of web document evolution on genre classification. In *Proc. 14th ACM International Conf. on Information and Knowledge Management (CIKM '05)* , pages 632–639.

Cavnar, W. and Trenkle, J. (1994). N-gram-based text categorization. *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR-94*.

Crowston, K. and Kwasnik, B. (2004). A framework for creating a facetted classification for genres: addressing issues of multidimensionality. *Proc. 37th Annual Hawaii International Conf. on System Sciences*.

Dong, L., Watters, C., Duffy, J., and Shepherd, M. (2008). An examination of genre attributes for web page classification. *Proc. 41st Annual Hawaii International Conf. on System Sciences (HICSS-41)*.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. *Proc. 12th International Conf. on Artificial Intelligence: Methodology, Systems, Applications*, pages 77–86.

Kanaris, I. and Stamatatos, E. (2007). Webpage genre identification using variable-length character n-grams. *19th IEEE International Conf. on Tools with Artificial Intelligence (ICTAI 2007)*, 2:3–10.

Kešelj, V., Peng, F., Cercone, N., and Thomas, T. (2003). N-gram-based author profiles for authorship attribution. In *Proc. Conf. Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264.

Lim, C., Lee, K., and Kim, G. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management*, 41(5):1263–1276.

Mason, J., Shepherd, M., and Duffy, J. (2009). An n-gram based approach to automatically identifying web page genre. *Proc. 41st Annual Hawaii International Conf. on System Sciences (HICSS-42)*.

Meyer zu Eissen, S. and Stein, B. (2004). Genre classification of web pages. *Proc. 27th German Conf. on Artificial Intelligence (KI-2004), Ulm, Germany*.

Rehm, G. (2002). Towards automatic web genre identification. *Proc. 35th Annual Hawaii International Conf. on System Sciences (HICSS-35)*, 04:101.

Santini, M. (2007). *Automatic identification of genre in web pages*. PhD thesis, University of Brighton, U.K.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27:379 – 423, 623 – 656.

Shepherd, M. and Watters, C. (1998). The evolution of cybergenres. *Proc. 31st Annual Hawaii International Conf. on System Sciences (HICSS-31)*, 02:97.

Stein, B. and Meyer zu Eissen, S. (2008). Retrieval models for genre classification. *Scandinavian Journal of Information Systems*, 20(1):93–119.

Swales, J. (1990). *Genre analysis*. Cambridge University Press New York.

APPENDIX

Table 1: Genre densities for the KI-04 corpus.

Genre	Number of Web pages	Genre	Number of Web pages
PORTRAIT (PRIVATE)	126	DOWNLOAD	151
ARTICLE	127	PORTRAIT (NON-PRIVATE)	163
DISCUSSION	127	SHOP	167
HELP	139	LINK COLLECTION	205

Table 2: A comparison of the best classification accuracy results for several researchers.

Researchers	4-Genre	7-Genre	KI-04
Dong et al. (2008)	97.0%		
Santini (2007)		90.6%	68.9%
Meyer zu Eissen and Stein (2004)			70.0%
Boese and Howe (2005)			74.8%
Kanaris and Stamatatos (2007)		96.5%	84.1%
Mason et al.	99.5%	94.5%	97.6%

Table 3: Summary of best results for the 4-Genre data set.

Preprocessing	Best Accuracy {Number of *n*-grams in each Web page profile}			
	word 1-gram	byte 5-gram	byte 6-gram	byte 7-gram
no preprocessing		98.6% {5000}	98.7% {5000}	98.2% {5000}
removed JavaScript		98.7% {5000}	98.7% {5000}	98.2% {5000}
removed HTML, JavaScript	99.5% {150}	99.4% {1000}	99.4% {1000}	99.4% {4000}
removed stopwords, HTML, JavaScript	98.7% {150}			

Table 4: Summary of best results for the 7-Genre data set.

Preprocessing	Best Accuracy {Number of *n*-grams in each Web page profile}			
	word 1-gram	byte 5-gram	byte 6-gram	byte 7-gram
no preprocessing		93.5% {4000}	93.5% {5000}	92.9% {5000}
removed JavaScript		94.5% {4000}	94.1% {5000}	93.1% {4000}
removed HTML, JavaScript	90.9% {100}	93.3% { 500}	93.3% { 750}	93.4% { 850}
removed stopwords, HTML, JavaScript	91.1% {100}			

Table 5: Summary of best results for the KI-04 data set.

Preprocessing	Best Accuracy {Number of *n*-grams in each Web page profile}			
	word 1-gram	byte 5-gram	byte 6-gram	byte 7-gram
no preprocessing		97.1% {5000}	97.2% {2000}	97.2% {3000}
removed JavaScript		97.0% { 850}	97.0% {2000}	97.0% {5000}
removed HTML, JavaScript	96.3% {350}	97.6% { 500}	97.2% { 350}	97.2% { 350}
removed stopwords, HTML, JavaScript	97.6% {150}			

Table 6: 10-fold cross-validated confusion matrix for the best results for the 4-Genre data set,using Web page profiles of 150 word *n*-grams of length 1; HTML tags and JavaScript code were removed from the data set as preprocessing steps. The table shows the percentage of correctly classified Web pages on the diagonal and summarizes the percentage of misclassified Web pages with respect to other genres.

Actual Genre	Assigned Genre			
	ESHOP	FAQ	FRONT PAGE	HOME PAGE
ESHOP	**99.0%**	1.0%	0.0%	0.0%
FAQ	0.0%	**100.0%**	0.0%	0.0%
FRONT PAGE	0.0%	0.0%	**100.0%**	0.0%
HOME PAGE	0.5%	0.5%	0.0%	**99.0%**

Table 7: 10-fold cross-validated confusion matrix for the best results for the 7-Genre data set, using Web page profiles of 4000 byte *n*-grams of length 5; JavaScript code was removed from the data set as a preprocessing step. The table shows the percentage of correctly classified Web pages on the diagonal and summarizes the percentage of misclassified Web pages with respect to other genres.

Actual Genre	Assigned Genre						
	BLOG	ESHOP	FAQ	FRONT PAGE	LISTING	HOME PAGE	SEARCH PAGE
BLOG	**99.0%**	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%
ESHOP	0.5%	**88.5%**	0.0%	0.0%	6.0%	2.0%	3.0%
FAQ	0.0%	0.0%	**99.0%**	0.0%	1.0%	0.0%	0.0%
FRONT PAGE	0.0%	0.0%	0.0%	**100%**	0.0%	0.0%	0.0%
LISTING	1.5%	2.0%	0.0%	1.5%	**87.0%**	3.0%	5.0%
HOME PAGE	0.0%	0.5%	0.0%	0.0%	3.5%	**94.5%**	1.5%
SEARCH PAGE	0.0%	0.5%	0.0%	0.0%	3.0%	1.5%	**95.0%**

Table 8: 10-fold cross-validated confusion matrix for the best results for the KI-04 data set, using Web page profiles of 700 byte *n*-grams of length 5; HTML tags and JavaScript code were removed from the data set as preprocessing steps. The table shows the percentage of correctly classified Web pages on the diagonal and summarizes the percentage of misclassified Web pages with respect to other genres.

Actual Genre	Assigned Genre							
	PORTRAIT (PRIV.)	ARTICLE	DISCUSSION	HELP	DOWNLOAD	PORTRAIT (NON-PRIV.)	SHOP	LINK COLL.
PORTRAIT (PRIV.)	**97.6%**	0.8%	0.0%	0.8%	0.0%	0.0%	0.0%	0.8%
ARTICLE	0.0%	**99.2%**	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%
DISCUSSION	0.0%	0.0%	**97.0%**	1.5%	0.0%	1.5%	0.0%	0.0%
HELP	0.7%	1.4%	0.0%	**97.2%**	0.0%	0.7%	0.0%	0.0%
DOWNLOAD	0.0%	0.0%	0.0%	0.0%	**100%**	0.0%	0.0%	0.0%
PORTRAIT (NON-PRIV.)	0.0%	0.6%	0.0%	0.6%	0.0%	**98.2%**	0.0%	0.6%
SHOP	0.0%	0.0%	0.0%	0.6%	1.3%	1.8%	**95.7%**	0.6%
LINK COLLECTION	0.0%	0.5%	0.0%	0.5%	0.5%	1.0%	1.0%	**96.5%**

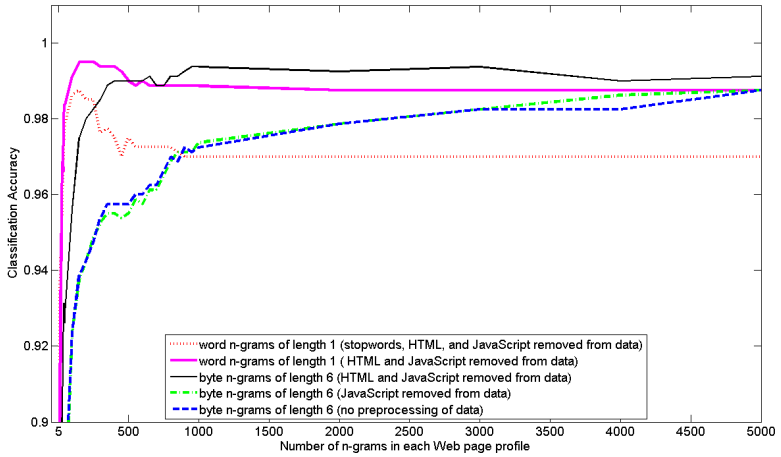

Figure 1: Classification accuracy for the 4-Genre data set; size of Web page *n*-gram profiles ranges from 5 to 5000.

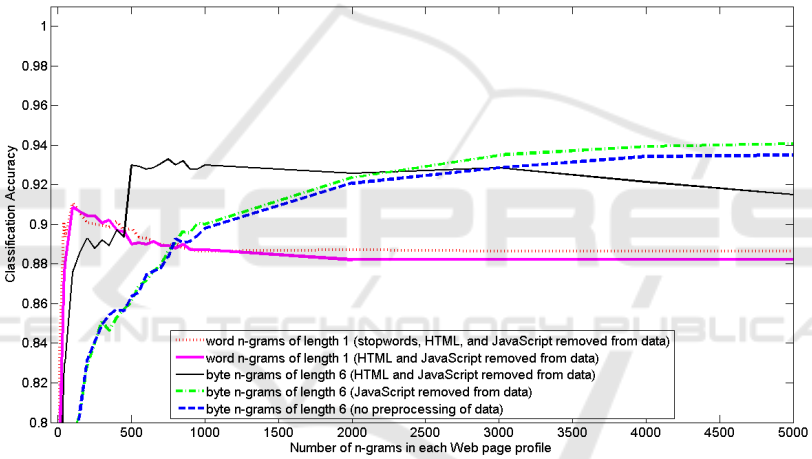

Figure 2: Classification accuracy for the 7-Genre data set; size of Web page *n*-gram profiles ranges from 5 to 5000.

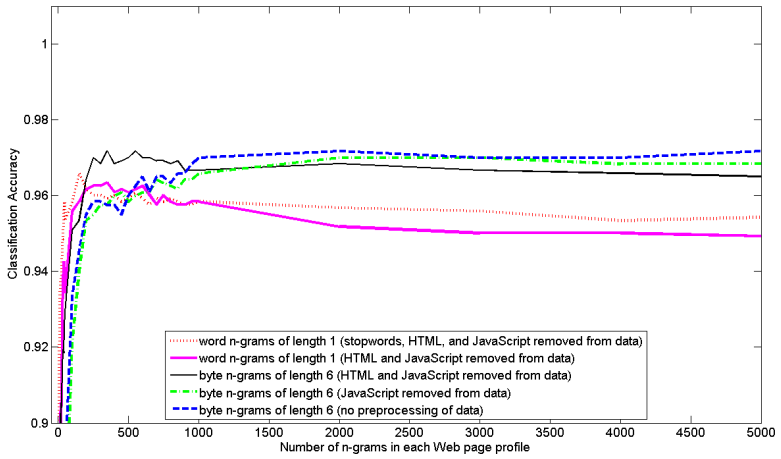

Figure 3: Classification accuracy for the KI-04 data set; size of Web page *n*-gram profiles ranges from 5 to 5000.