Webpage Genre Identification Using Variable-length Character *n*-grams

Ioannis Kanaris and Efstathios Stamatatos
University of the Aegean
{kanaris.i; stamatatos}@aegean.gr

Abstract

An important factor for discriminating between webpages is their genre (e.g., blogs, personal homepages, e-shops, online newspapers, etc). Webpage genre identification has a great potential in information retrieval since users of search engines can combine genre-based and traditional topic-based queries to improve the quality of the results. So far, various features have been proposed to quantify the style of webpages including word and html-tag frequencies. In this paper, we propose a low-level representation for this problem based on character n-grams. Using an existing approach, we produce feature sets of variable-length character n-grams and combine this representation with information about the most frequent html-tags. Based on two benchmark corpora, we present webpage genre identification experiments and improve the best reported results in both cases.

1. Introduction

One basic property of a document is its genre. Although there is no agreed definition of genre, it relates closely to a distinctive type of discourse. Documents of the same genre share a set of communicative purposes (purpose, form of transmission, etc) [17]. As a result, documents of the same genre have stylistic (rather than thematic) similarities. Unfortunately, there is no consensus about the complete set of different genres (and sub-genres). Nowadays, webpages form a new kind of genre taking advantage of a new communication medium and different type of interaction with the receiver (hypertext). Most of the webpage genres can hardly find an antecedent in the traditional paper documents (e.g., personal home pages, blogs).

So far, search engines support topic-based queries. There is only limited control over the genre of the results (e.g., Goggle allows different searches for the web, discussion forums, and news). Automatic webpage genre identification could improve significantly the performance of information retrieval systems. Despite topic-related terms, users would be able to define the type of webpage they look for. Alternatively, results from a topic-related set of keywords could be grouped according to their genre so that the user can easily navigate through them and find the type of information that matches their interests.

Studies on webpage genre identification focus on the definition of appropriate quantification of style so that to reveal genre properties. Following the practice of similar work on text-genre identification [7, 15], topic-neutral features are usually proposed (e.g., 'function' words, closed-class words etc). Apart from textual information, structural (html-based) information is usually taken into account. To this end, different sets of manually defined html tags have been proposed [1, 10, 12].

A crucial factor for estimating the applicability of the proposed approaches is the objective evaluation. Unfortunately, the majority of the previous studies do not provide a reliable comparison with other approaches. The main reason for this is that, until recently, there were no publicly-available benchmark corpora for this task. Another reason is that there is not agreed sense of webpage genres and each study focuses on a different set of genres [2, 9, 10] or a specific genre and its sub-genres (e.g., home pages) [5].

The aim of this paper is to explore the use of novel features, i.e., character *n*-grams, for representing webpage genres. Character *n*-grams (contiguous n characters) are able to capture nuances of style including lexical and syntactical information. In addition, they are robust to noisy texts. Note that webpages are very noisy including irregular use of punctuation and spelling errors. The character *n*-gram representation has also been applied to authorship identification [6, 16], another type of style-based classification task, with promising results. Based on an existing approach for *n*-gram feature selection [3], we produce a variable-length character *n*-gram representation suitable for the task of webpage genre identification. Moreover, in contrast to previous work, we propose a fully-automated approach for extracting structural information. That is, the procedure of finding the most useful html tags for distinguishing among genres requires no manual effort.

The proposed models are objectively evaluated in two benchmark corpora already used for webpage genre identification. Hence, we are able to compare our models with previous work based on two different genre palettes. A series of experiments show that our approach is quite effective and significantly improves the best reported results.

The rest of this paper is organized as follows. Section 2 gives an overview of the previous work on webpage genre identification. Section 3 describes the extraction of textual information, based on character *n*-grams of variable-length, and structural information, based on html tag frequencies. The benchmark corpora used in this study and the result of the experiments are presented in Section 4. Finally, Section 5 summarizes the main points of this paper and proposes future work directions.

2. Previous work

The previous studies of webpage genre identification differ significantly with respect to two factors: *i*. the feature set they use to represent the content and structure of webpages, and *ii*. the genre palette (the set of different genres). The following review of previous work is presented in chronological order. Moreover, this review focuses on studies related to webpage genre classification, rather than text genre classification.

Lee and Myaeng [8] focus on distinguishing genre-related features from subject-related features. To this end, they use a corpus annotated for both genre and topic and eliminate features that are closely related to the subject of webpages. Using a set of seven genres (reportage, editorial, technical paper, critical review, personal homepage, Q&A, and product specification) they report 87% precision/recall values.

Meyer zu Eissen and Stein [10] provided a corpus of eight genres (see KI-04 in Table 1) following a user study on genre usefulness. Moreover, they examined various feature sets attempting to combine different kinds of information including presentation-related features (html tag frequencies), closed word features (names, dates, unknown words, etc.), text statistics (punctuation mark frequencies) and syntactic features (part-of-speech frequencies). Using discriminant analysis, they report 70% average classification accuracy.

Lim *et al.* [9] focused on the usefulness of information found in different parts of the webpages (title, body, anchor text etc). Based on a corpus of 15 genres (personal home page, public home page, commercial home page, bulletin collection, link collection, image collection, simple table/lists, input pages, journalistic material, research report, official materials, FAQs, discussions, product specification, informal texts) they indicated that the main body and anchor text information is the most effective.

Kennedy and Shepherd [5] emphasized on a specific genre and its sub-genres. Using a neural network they attempted to discriminate between home pages from non-home pages. On a second level, they classify home pages into three categories (personal, corporate, and organization). Their feature set comprises features about the content (e.g., common words, meta tags), form (e.g., number of images), and functionality (e.g., number of links, use of Javascript). The best reported results were for personal home pages.

Boese [1] examined the effects of webpage evolution on genre classification. She found that webpages in some genres change rarely while webpages of other genres change continuously. This information could be very useful to spiders for tuning the frequency of visits. She used features combining style (e.g., part-of-speech frequencies), form (e.g., html tag frequencies), and content information (e.g., bag-of-words).

Finn and Kushmerick [2] investigate different feature sets for genre classification, including bag-of-words (BOW), part-of-speech frequencies, and text statistics (e.g., sentence length) and how they can be combined to improve performance. They focused on two genres, articles and reviews, and examined whether an article is subjective or objective and whether a review is positive or negative. However, these classes better match the task of sentiment analysis rather than the task of genre identification.

Last but not least, Santini [12] studied webpage genre classification based on three different feature sets including frequencies of common words, part-of-speeches and part-of-speech trigrams, html tags, punctuation marks etc. She built a corpus of seven genres (see 7genre in Table 1) and achieved 90.6% classification accuracy. She also evaluated her approach to another corpus (see Table 2) and used the produced models to predict the genre of unclassified webpages. However, the latter experiments provided discouraging results.

3. Representing webpages

In order to represent the content of webpages we use two sources of information: textual and structural information. The procedure of extracting this information is described in the following subsections.

3.1. Textual information

To extract information from the textual content of the webpages, we first remove any html-based information. Then, we use character *n*-grams to quantify the textual information. Houvardas and Stamatatos [12] propose a feature selection method for character *n*-grams of variable-length aiming at authorship identification. In this study, we use this approach in order to represent

webpages as a 'bag-of-character *n*-grams'. Having a big initial set of variable-length *n*-grams (e.g., composed by equal amounts of fixed-length *n*-grams), the main idea is to compare each *n*-gram with similar *n*-grams (either immediately longer or shorter) and keep the dominant *n*-grams based on their frequency of occurrence. All the other *n*-grams are discarded and they don't contribute to the classification model. To this end, we need a function able to express the significance of an *n*-gram. We view this function as 'glue' that sticks the characters together within an *n*-gram. The higher the glue of a *n*-gram, the more likely for it to be included in the set of dominant *n*-grams. For example, the 'glue' of the *n*-gram |the_| will be higher than the 'glue' of the *n*-gram |thea|[1].

3.1.1. Extraction of dominant *n*-grams. To extract the dominant character *n*-grams in a corpus we modified the algorithm *LocalMaxs,* originally introduced by Silva and Lopes [14] for extracting multiword terms (i.e., word *n*-grams of variable length) from texts. It is an algorithm that computes local maxima comparing each *n*-gram with similar (shorter or longer) *n*-grams. Given that:

- $g(C)$ is the glue of *n*-gram C, that is the power holding its characters together.
- $ant(C)$ is an antecedent of an *n*-gram C, that is, a shorter string composed by *n*-1 consecutive characters of C.
- $succ(C)$ is a successsor of C, that is, a longer string of size *n*+1, i.e., composed by C and one extra character either on the left or right side of C.

Then, the dominant *n*-grams are selected according to the following rules:

$$if\ (C.length > 3)$$
$$g(C) \geq g(ant(C)) \wedge g(C) > g(succ(C)),\ \forall ant(C), succ(C)$$
$$if\ (C.length = 3)$$
$$g(C) > g(succ(C)),\ \forall succ(C)$$

In this study, we only consider 3-grams, 4-grams, and 5-grams as candidate *n*-grams since they can capture both sub-word and inter-word information and keep the dimensionality of the problem in a reasonable level. Note that, according to the proposed algorithm, 3-grams are only compared with successor *n*-grams. Moreover, 5-grams are only compared with antecedent *n*-grams. So, it is expected that the proposed algorithm will favor 3-grams and 5-grams against 4-grams.

[1] We use '|' and '_' to denote *n*-gram boundaries and a single space character, respectively.

3.1.2. Representing the *glue*. To measure the glue holding the characters of an *n*-gram together we adopt the *Symmetrical Conditional Probability* (SCP) proposed by Silva *et al.* [13]. The SCP of a bigram |xy| is the product of the conditional probabilities of each given the other:

$$SCP(x,y) = p(x \mid y) \cdot p(y \mid x) = \frac{p(x,y)}{p(x)} \cdot \frac{p(x,y)}{p(y)} = \frac{p(x,y)^2}{p(x) \cdot p(y)}$$

Given a character *n*-gram $|c_1 \ldots c_n|$, a dispersion point defines two subparts of the *n*-gram. Hence, an *n*-gram of length *n* contains *n*-1 possible dispersion points (e.g., if * denote a dispersion point, then the 3-gram |the| has two dispersion points: |t*he| and |th*e|). Then, the SCP of the *n*-gram $|c_1 \ldots c_n|$ given the dispersion point $| c_1 \ldots c_{n-1} * c_n|$ is:

$$SCP((c_1 \ldots c_{n-1}), c_n) = \frac{p(c_1 \ldots c_n)^2}{p(c_1 \ldots c_{n-1}) \cdot p(c_n)}$$

Essentially, this is used to suggest whether the *n*-gram is more important than the two substrings defined by the dispersion point. The lower the SCP, the less important the initial *n*-gram. The SCP measure can be easily extended so that to account for any possible dispersion point (since this measure is based on fair dispersion point normalization, will be called *fairSCP*). Hence, the *fairSCP* of the *n*-gram $|c_1 \ldots c_n|$ is as follows:

$$fairSCP(c_1 \ldots c_n) = \frac{p(c_1 \ldots c_n)^2}{\frac{1}{n-1} \sum_{i=1}^{i=n-1} p(c_1 \ldots c_i) \cdot p(c_{i+1} \ldots c_n)}$$

3.2. Structural information

The structural information of a webpage is encoded into the html tags used in this page. To quantify this structural information, we use a BOW-like approach based on html-tags instead of words. That is, we first extract the list of all the html tags that appear at least three times in the entire collection of webpages and represent each webpage as a vector using the frequencies of appearance of these html tag in that page.

Then, the ReliefF algorithm for feature selection [11] is applied to the produced dataset to reduce the dimensionality. This algorithm is able to detect conditional dependencies between attributes and is noise tolerant. That way, we avoid any manual definition of useful html tags and the whole procedure is fully-automated. In addition, the extracted html tags can be easily adjusted to the particular properties of a specific corpus.

Table 1. The two corpora of webpage genres used in this study.

7Genre		KI-04	
Genres	**Pages**	**Genres**	**Pages**
BLOG	200	ARTICLE	127
E-SHOP	200	DOWNLOAD	151
FAQs	200	LINK COLLECTION	205
ONLINE NEWSPAPER FRONTPAGE	200	PORTRAYAL-PRIVATE	126
LISTING	200	DISCUSSION	127
PERSONAL HOME PAGE	200	HELP	139
SEARCH PAGE	200	PORTRAYAL-NON PRIVATE	163
		SHOP	167

4. Webpage genre identification experiments

4.1. Corpora

Recently, several webpage corpora appropriate for evaluating genre identification approaches have become available. Although, there are significant differences about the methodology of defining the genres and populating the webpages under each genre, such benchmarks provide an objective testing ground for comparing different approaches. In this paper, we use two benchmark corpora, already used by several previous studies. Details about these corpora[2] can be found in Table 1.

7genre: This corpus was built in early 2005 and consists of 1,400 English webpages categorized in 7 genres following the criteria of 'annotation by objective sources' and consistent genre granularity' (with the exception of the LISTING genre which may be decomposed into many sub-genres) [12]. This is a balanced corpus, meaning that the webpages are equally distributed among the genres.

KI-04: This corpus was built in 2004 and comprises 1,205 English webpages classified under 8 genres. These genres were suggested by a user study on genre usefulness [10]. The distribution of webpages in genres is not balanced.

Note that for some genres included in the aforementioned corpora there are great similarities (e.g., PERSONAL HOME PAGE and E-SHOP of 7genre are quite similar with PORTRAYAL-PRIVATE and SHOP of KI-04, respectively). However, the exact relationships between all the genres defined in these two corpora are not clear.

As already mentioned, several researchers used these corpora to evaluate their methods. The so far reported results are summarized in Table 2.

[2] Both corpora were downloaded from: http://www.nltg.brighton.ac.uk/home/Marina.Santini/

Table 2. Best accuracy results for the two corpora used in this study (* The results repored in [1] refer to a sub-corpus of KI-04).

Approach	7Genre	KI-04
[12]	90.6%	68.9%
[1]	-	74.8%*
[10]	-	70.0%
Proposed in this paper		
Textual only	96.2%	82.8%
Textual + structural	96.5%	84.1%

4.2. Results

Three sets of experiments were conducted. First, only textual information was used. The initial set of features comprised equal amounts of three fixed-length character *n*-gram subsets (e.g., an initial feature set of 3,000 features includes the 1,000 most frequent character 3-grams, the 1,000 most frequent character 4-grams, and the 1,000 most frequent character 5-grams). The initial feature sets included 3,000 up to 24,000 *n*-grams, with a step of 3,000 features (i.e., 1,000 from each of the three fixed-length *n*-gram subsets) were examined. Then, the feature selection approach described in Section 3.1 was applied to the initial set and the resulting feature sets of variable-length character *n*-grams were used to represent the webpages of each corpus. Finally, a Support Vector Machine (SVM), a machine learning algorithm able to deal with high-dimensional and sparse data [4], was used for the classification. The (microaverage) classification results of stratified 10-fold cross-validation evaluation for 7genre and KI-04 are presented in Figures 1 and 2, respectively. As can be seen, the character *n*-gram representation is very effective and the produced results are better than the best reported results for both corpora (see Table 2).

Experiments using structural information exclusively (following the procedure described in Section 3.2) revealed that html-based information has not enough descriptive power in order to achieve good results.

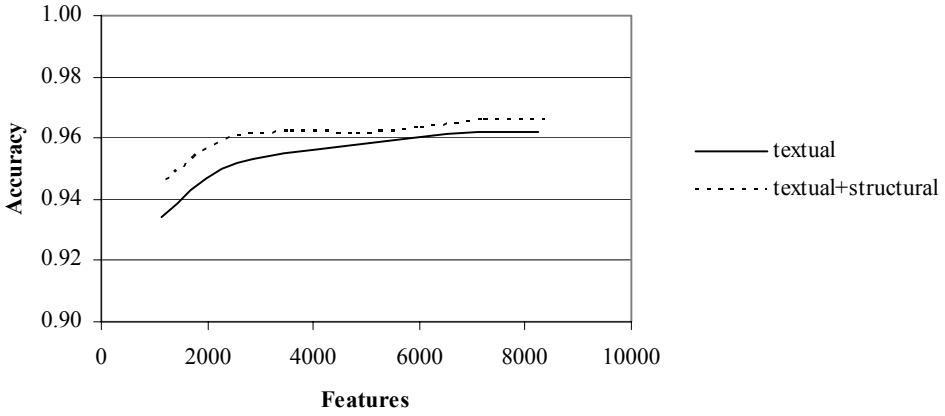

Figure 1. Microaverage accuracy results for the 7genre corpus using textual only and textual+structural models and variable feature set size.

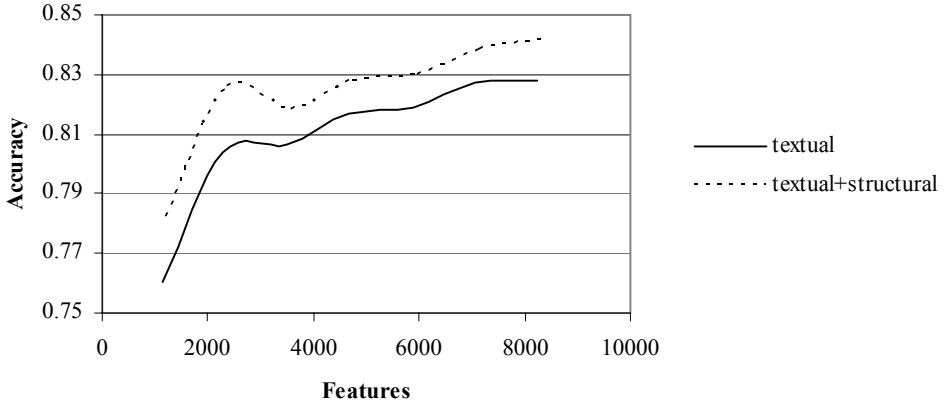

Figure 2. Microaverage accuracy results for the KI-04 corpus using textual only and textual+structural models and variable feature set size.

Specifically, in the case of 7genre corpus the structural only classifier achieved an accuracy of 76% while for the KI-04 corpus the best results were only 42%.

In the third set of experiments, the textual information was enriched with structural information. The classification accuracy results (stratified cross-validation) for the enriched feature sets and the SVM model applied to the 7genre and KI-04 corpora are presented in Figures 1 and 2, respectively. Apparently, the structural information assists the model to achieve even higher performance. The most notable increase in the performance with respect to the textual model is for the relatively low dimensional feature sets (<4,000 features). Moreover, the increase in performance is greater for the KI-04 corpus in comparison to the 7genre corpus. This can be explained by the fact that the textual information already achieved quite high classification accuracy for the 7 genre corpus.

The best classification results achieved by the textual models and the textual+structural models can be seen in Table 2 together with the best reported results on the same corpora (based in cross-validation as well). In order to have a more detailed picture of the classification produced by these models, Tables 3 and 4 present the confusion matrices for the 7genre corpus based on the textual only model and the textual+structural model, respectively. Similarly, Tables 5 and 6 present the corresponding confusion matrices for KI-04. For comparative purposes, similar confusion matrices of previously reported results on the same corpora can be found in [12] and [10].

Regarding the 7genre corpus, the most successful identification results are for the genres ONLINE NEWSPAPER FRONTPAGE, FAQs, and BLOG. On the other hand, the most difficult case is the genre LISTING. This is in accordance with the results reported in [12].

Table 3. Confusion matrix for the 7genre corpus and the textual only model. Main misclassification pairs are in boldface.

Actual	Classified as						
	BLOG	**ESHOP**	**FAQ**	**ONF**	**LIST**	**PHP**	**SEARCH**
BLOG	98.0%				0.5%	**1.5%**	
ESHOP		96.5%			1.0%	1.5%	**1.0%**
DOWN			99.0%		1.0%		
ONF				100.0%			
LIST	1.0%	1.0%			91.0%	**4.5%**	**2.5%**
PHP	**1.0%**				**2.5%**	96.0%	0.5%
SEARCH	0.5%	**2.0%**		0.5%	**3.5%**	0.5%	93.0%

Table 4. Confusion matrix for the 7genre corpus and the textual+structural model. Main misclassification pairs are in boldface.

Actual	Classified as						
	BLOG	**ESHOP**	**FAQ**	**ONF**	**LIST**	**PHP**	**SEARCH**
BLOG	98.5%					**1.5%**	
ESHOP		97.5%			1.5%		**1.0%**
DOWN			99.0%		1.0%		
ONF				100.0%			
LIST	1.0%	0.5%			91.5%	**4.5%**	**2.5%**
PHP	**1.0%**				**2.5%**	96.5%	
SEARCH	**0.5%**	2.0%		0.5%	**3.5%**	0.5%	93.0%

Table 5. Confusion matrix for the KI-04 corpus and the textual only model. Main misclassification pairs are in boldface.

Actual	Classified as							
	ART	**DOWN**	**LINK**	**P-P**	**DISC**	**HELP**	**P-NP**	**SHOP**
ART	81.1%		2.4%	3.1%	0.8%	**5.5%**	6.3%	0.8%
DOWN		94.0%		0.7%		2.0%	1.3%	2.0%
LINK	1.5%	1.0%	84.4%	2.0%		2.4%	**6.3%**	2.4%
P-P	4.0%	1.6%	0.8%	89.7%			4.0%	
DISC	2.4%		1.6%	0.8%	90.6%	2.4%	2.4%	
HELP	**6.5%**	2.9%	5.0%	1.4%	3.6%	69.8%	7.9%	2.9%
P-NP	0.6%	1.8%	**6.7%**	6.1%	2.5%	1.8%	71.8%	**8.6%**
SHOP		3.0%	1.2%	1.8%		1.8%	**9.6%**	82.6%

Table 6. Confusion matrix for the KI-04 corpus and the textual+structural model. Main misclassification pairs are in boldface.

Actual	Classified as							
	ART	**DOWN**	**LINK**	**P-P**	**DISC**	**HELP**	**P-NP**	**SHOP**
ART	85.8%		2.4%	3.1%		**3.9%**	4.7%	
DOWN		94.0%		0.7%		2.6%	0.7%	2.0%
LINK	1.0%	1.0%	84.9%	2.4%		2.9%	**5.9%**	2.0%
P-P	3.2%	1.6%	0.8%	92.1%			2.4%	
DISC	2.4%		1.6%	0.8%	90.6%	2.4%	2.4%	
HELP	**7.2%**	2.9%	5.0%	0.7%	2.9%	69.8%	8.6%	2.9%
P-NP	0.6%	1.8%	**5.5%**	4.3%	2.5%	2.5%	74.2%	**8.6%**
SHOP		3.0%	1.2%			1.8%	**10.2%**	83.8%

Recall that this class is considered a super-genre and can be decomposed into several subgenres. However, the identification results for this genre (91.5%) are much better than the ones reported in [12] (74.5%). The main misclassifications take place in the pairs SEARCH – LISTING, BLOG – PERSONAL HOME PAGE, ESHOP – SEARCH, and LIST – PERSONAL HOME PAGE. Again, this is in accordance with the results presented in [12]. Apparently, the textual+structure model achieves to slightly increase the performance in many cases. Most notably, it solves the confusion between certain pairs (PHP – SHOP) and reduces the cells (excluding the diagonal that corresponds to correct answers) with non-zero values.

As concerns the KI-04 corpus, the most successful cases are the genres DOWNLOAD, DISCUSSION, and PORTRAYAL-PRIVATE. On the other hand, the HELP and PORTRAYAL-NON-PRIVATE genres are the most difficult to be discriminated. Note that in the results reported in [10] the best cases were DOWNLOAD and ARTICLE and the worst cases were (also) HELP and PORTRAYAL-NON-PRIVATE. The main misclassifications are between the pairs SHOP – PERSONAL-NON-PRIVATE, PORTRAYAL-NON-PRIVATE – LINK COLLECTION, and ARTICLE – HELP. Again, this is in accordance with the results of [10]. However, the pair PORTRAYAL-PRIVATE – LINK COLLECTION that was also a confusing factor according to the results of [10], do not contribute significantly to the misclassifications of our approach.

Moreover, the comparison of the performance of the textual only model and the textual+structural model shows that the latter is better able to discriminate the ARTICLE, PORTRAYAL-PRIVATE, and PORTRAYAL-NON-PRIVATE genres. Similarly to 7genre, the number of cells with non-zero value (excluding the diagonal) is reduced by taking into account the structural information. On the other hand, the textual only model is slightly better in discriminating among the pair HELP – PORTRAYAL-NON-PRIVATE. This indicates structural similarities between these webpage genres.

5. Conclusions

In this paper, we proposed the use of character *n*-grams for the task of webpage genre identification. This type of features is very effective in representing the stylistic properties of textual data. The application of the proposed approach based on character *n*-grams of variable-length to two benchmark corpora improved the best reported results. Moreover, we examined a methodology for extracting structural information from the webpages using html tag frequencies. The enriched models using both textual and structural information improved the results even more.

In comparison with previous work on webpage genre identification, the proposed approach is fully-automated, in the sense that it does not require any manual selection of features that best capture the stylistic properties of text or structure of webpage. In contrast to the majority of the previous studies, the size of the proposed feature set is high (several thousands of features).

Regarding the evaluation procedure, we chose to perform stratified cross-validation in order to directly compare the produced results with previously reported results on the same datasets. Due to lack of data it is not easy to perform experiments on a higher scale. Another crucial issue is a deeper understanding of the properties of webpage genres and the relationships among genres taken from different genre palettes. To this end, since there is not yet a consensus about the notion of webpage genre, cross-check experiments (i.e., models trained on a given genre palette and tested on a different, and possibly unknown, genre palette) would provide useful clues about genre similarities/differences.

References

[1] E. Boese, A. Howe, "Effects of Web Document Evolution on Genre Classification". *Proc. of the ACM 14th Conference on Information and Knowledge Management*, 2005.

[2] Finn A. and Kushmerick N. "Learning to Classify Documents According to Genre", *Journal of the American Society for Information Science and Technology*, 7(5), 2006, pp. 1506-1518.

[3] J. Houvardas, E. Stamatatos, "N-gram Feature Selection for Authorship Identification", *Proc. of the 12th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications*, LNCS, 4183, Springer, 2006, pp. 77-86.

[4] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". *Proc. of the 10th European Conference on Machine Learning*, 1998, pp. 137-142.

[5] A. Kennedy, M. Shepherd, "Automatic Identification of Home Pages on the Web". *Proc. of the 38th Hawaii International Conference on System Sciences*, 2005.

[6] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based Author Profiles for Authorship Attribution". *Proc. of the Conf. of Pacific Association for Computational Linguistics*, 2003.

[7] B. Kessler, G. Nunberg, H. Shütze, "Automatic Detection of Text Genre", *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997, pp. 32-38.

[8] Y.B. Lee, S.H. Myaeng, "Text Genre Classification with Genre-revealing and Subject-revealing Features". *Proc. of the 25th ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2002, pp. 145-150.

[9] C.S. Lim, K.J. Lee, G.C. Kim, "Multiple Sets of Features for Automatic Genre Classification of Web Documents", *Information Processing and Management,* 41(5), 2005, pp. 1263-1276.

[10] S. Meyer zu Eissen, B. Stein, "Genre Classification of Web Pages: User Study and Feasibility Analysis". In Biundo S., Fruhwirth T. and Palm G. (eds.). *KI 2004: Advances in Artificial Intelligence,* Springer, 2004, pp. 256-269.

[11] M. Robnik-Sikonja, I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF", *Machine Learning,* 53(1-2), 2003, pp. 23-69.

[12] M. Sanitni, *Automatic Identification of Genre in Webpages*, Ph.D. Thesis, University of Brighton, 2007.

[13] J. Silva, G. Dias, S. Guilloré, and G. Lopes, "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units", *Lecture Notes on Artificial Intelligence*, 1695, Springer, 1999, pp. 113-132.

[14] J. Silva and G. Lopes, "A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units", In *Proc. of the 6th Meeting on the Mathematics of Language,* 1999, pp. 369-381.

[15] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text Genre Detection Using Common Word Frequencies" In *Proc. of the 18th Int. Conf. on Computational Linguistics,* 2000, pp. 808-814.

[16] E. Stamatatos, "Ensemble-based Author Identification Using Character N-grams". *Proc. of 3rd Int. Workshop on Text-based Information Retrieval*, 2006, pp. 41-46.

[17] J. Swales, *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press, 1990.