Learning to Recognize Webpage Genres

IOANNIS KANARIS & EFSTATHIOS STAMATATOS

Dept. of Information and Communication Systems Eng.
University of the Aegean
Karlovassi, Samos – 83200, Greece
{kanaris.i; stamatatos}@aegean.gr

Abstract

Webpages are mainly distinguished by their topic (e.g., politics, sports etc.) and genre (e.g., blogs, homepages, e-shops, etc.). Automatic detection of webpage genre could considerably enhance the ability of modern search engines to focus on the requirements of the user's information need. In this paper, we present an approach to webpage genre detection based on a fully-automated extraction of the feature set that represents the style of webpages. The features we propose (character *n*-grams of variable length and HTML tags) are language-independent and easily-extracted while they can be adapted to the properties of the still evolving web genres and the noisy environment of the web. Experiments based on two publicly-available corpora show that the performance of the proposed approach is superior in comparison to previously reported results. It is also shown that character *n*-grams are better features than words when the dimensionality increases while the binary representation is more effective than the term-frequency representation for both feature types. Moreover, we perform a series of cross-check experiments (e.g., training using a genre palette and testing using a different genre palette as well as using the features extracted from one corpus to discriminate the genres of the other corpus) to illustrate the robustness of our approach and its ability to capture the general stylistic properties of genre categories even when the feature set is not optimized for the given corpus.

Keywords: Web genre classification; Webpage representation; Character *n*-grams

1. INTRODUCTION

Topic and genre are the main factors of characterizing a document. The genre of documents relates closely to a distinctive type of discourse. That is, documents of the same genre share a set of communicative purposes (purpose, form of transmission, etc.) (Swales, 1990). As a result, documents of the same genre have stylistic (rather than thematic) similarities. Unfortunately, there is no agreed upon definition of genre, there is no consensus about the complete set of different genres, and there is no established hierarchy of genres. Another confusing factor is the definition of appropriate genre labels since people coming from different communities tend to describe a document in different terms.

Nowadays, webpages form new kinds of genre, taking advantage of a new communication medium and different type of interaction with the receiver. Some of the webpage genres are variants of traditional genres (e.g., online newspapers) while others have no antecedents in paper documents (e.g., personal home pages, blogs) (Shepherd & Watters, 1998). So far, search engines support topic-based queries. There is only limited control over the genre of the results (e.g., Google allows different searches for the web, discussion forums, and news). Automatic webpage genre identification could significantly improve the ability of information retrieval systems by suggesting queries that better describe the user's information need or by facilitating intuitive navigation through search results. In more detail, the contribution of genre-specific information would be twofold:

- In addition to topic-related terms, users could be able to specify the type of webpage they are looking for (e.g., blog, homepage, e-shop, etc.) (Rosso, 2005).
- Alternatively, the search results from a topic-related set of keywords could be grouped according to their genre in addition or in parallel to thematic clustering (Bekkerman, 2008), or filtered according to genre (Meyer zu Eissen, 2007), or ranked based on topic and genre relevance (Braslavski, 2007; Vidulin, *et al.*, 2007), so that the users can easily navigate through them and find the type of information that matches their interests.

Additionally, genre detection could also assist metadata extraction from XML repositories to enhance information management in digital libraries (Clark & Watt, 2007).

Usually, webpage genre identification is viewed as a single-label classification task (Meyer zu Eissen & Stein, 2004; Lim, *et al*., 2005; Kennedy and Shepherd, 2005). However, it is not rare for multiple genres to co-exist in a single page. That is, a webpage can be decomposed into parts, each of which belongs to a different genre (e.g., a part of an e-shop page may display search results or news). This fact raised questions about considering webpage genre detection as a multi-label classification task, so that a webpage can be an e-shop and a search page at the same time (Santini, 2008). However, since the parts of webpages belonging to different genres are visually separated rather than arbitrarily mixed, a webpage segmentation tool could detect the main areas of the webpages based on stylistic coherence and, then, a single-label webpage identification tool could be applied to each part separately. Another difficulty is that the same webpage (or webpage section) can be described with different labels (e.g., home page, informational, entry page, etc.). This situation could also be handled by a set of single-label genre detectors trained on different genre palettes to provide these multiple labels.

Studies on webpage genre identification focus on the definition of appropriate textual features that quantify the stylistic properties of genres. Following the practice of similar works on text-genre detection (Kessler, *et al.*, 1997; Stamatatos, *et al.*, 2000), topic-neutral features can be defined on the word level (e.g., ‘function’ words, closed-class words, etc.) or be extracted after appropriate text processing, e.g., part-of-speech frequencies (Dewdney, *et al.*, 2001). 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 (Meyer zu Eissen & Stein, 2004; Boese & Howe, 2005; Santini, 2007). In most cases, the combination of content and structure features improves the results. The vast majority of previous approaches to webpage genre detection are based on feature sets of small size comprising features selected manually or by applying a feature selection algorithm (Meyer zu Eissen & Stein, 2004; Lim, *et al.*, 2005; Kennedy and Shepherd, 2005; Santini, 2007, Levering, *et al.*, 2008). However, since the web is a rapidly changing and noisy environment, the features of an efficient webpage genre classifier should be able to adapt to the current conditions of webpages. In this way, the properties of the still evolving web genres would be better captured. A manually-defined feature set requires effort to be adapted to new genres or to capture modified genre properties. On the other hand, features extracted by a feature selection algorithm are optimized for specific corpora or genre palettes. Hence they are not exportable to other corpora or different genres. Moreover, many features used to represent textual information (e.g., part-of-speech frequencies) require the application of natural language processing tools, hence they are language and domain dependent.

In this paper, we propose a fully-automated approach that is able to extract the most appropriate low-level features from a given corpus. To this end, we explore the use of 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. They are language-independent, easily-extracted from any text, and robust to noisy texts. Note that certain webpage genres (e.g. blogs) are very noisy and include irregular use of punctuation and spelling errors. The character *n*-gram representation has also been applied to authorship identification (Keselj, *et al.*, 2003; Stamatatos, 2006), another type of style-based classification task, with promising results. Based on an existing approach for *n*-gram feature selection (Houvardas & Stamatatos, 2006), we produce a variable-length character *n*-gram representation suitable for the task of webpage genre identification. Moreover, we compare the character *n*-gram model with a traditional word-based approach of comparable dimensionality examining both binary and term-frequency (TF) features. In addition to the textual information, we propose a method for automating the extraction of structural information based on HTML tags.

A crucial factor for estimating the applicability of the proposed approaches is 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 corpora for this task. Another reason is that each study focuses on a different set of genres (Meyer zu Eissen & Stein, 2004; Lim, *et al.*, 2005; Finn & Kushmerick, 2006) or a specific genre and its sub-genres (e.g., home pages) (Kennedy & Shepherd, 2005). The proposed models are objectively evaluated in two corpora already used for webpage genre identification. Hence, we are able to compare our models with previous work based on two different genre palettes and similar evaluation techniques. A series of experiments show that our approach is quite effective and significantly improves the best reported results. Moreover, given that our approach requires high dimensionality and the evaluation corpora are small, we introduce a series of cross-check experiments to exhibit the credibility and robustness of our approach in cases where the feature set is not optimized for a particular corpus or, more interestingly, a model trained on a specific genre palette is applied to another genre palette.

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 character *n*-gram features as well as the extraction of structural information. The corpora used in this study and the results of the experiments are presented in Section 4 while Section 5 discusses the results and provides comparison to previous work on the same corpora. Finally, Section 6 summarizes the main points of this paper and proposes future work directions.

2. PREVIOUS WORK

The following review of previous work is presented in chronological order. Moreover, this review deals with research related to webpage genre classification, rather than text genre classification. Each study can be described in terms of three factors, namely:

- The genre palette (the set of different genres).
- The feature set used to represent the content and structure of webpages, and
- The classification algorithm used to identify the most likely genre.

Usually, the latter is performed by well-known statistical and machine learning algorithms (e.g., discriminant analysis, neural networks, memory-based learners, support vector machines, etc.) that have been thoroughly studied in the field of text categorization research (Sebastiani, 2002). Hence, in the rest of this section we will focus on the first two factors.

Lee and Myaeng (2002) were the first to use web documents in genre detection. However, their aim was to investigate whether genre-related features could help topical text classification, rather than studying the webpage genre detection problem. As a result, their genre palette was a mix of text genres and webpage genres (reportage, editorial, technical paper, critical review, personal homepage, Q&A, and product specification). The web genre palettes used so far were either general, so that to cover a large part of the web, or specific, so that to examine the properties of certain genres. Meyer zu Eissen and Stein (2004) provided a general palette of eight webpage genres (see KI-04 in Table 1) following a user study on web genre usefulness. Lim, *et al*. (2005) used 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). Santini (2007) reports another general palette comprising seven genres (see 7genre in Table 1). Kim and Ross (2007) enriched this collection with 24 general document genres (populated by PDF web documents). In another study, Vidulin, *et al*., (2007) describes a collection of 20 genre labels (see Table 13) allowing several labels to be assigned to one webpage. Although there are certain similarities in such general genre palettes (e.g. personal home page) it is not clear how each genre of one palette is associated to the genres of another palette.

Concerning more specific genre palettes, Kennedy and Shepherd (2005) emphasized the discrimination between home pages and non-home pages. On a second level, they classified home pages into three categories (personal, corporate, and organization) while the identification of personal home pages was the most effective. Finn and Kushmerick (2006) investigated 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. Levering, *et al.* (2008) used four related genres (store home pages, store product lists, store product descriptions, and other store pages). Kim and Ross (2008) focused on PDF web documents and a collection of six genres (academic monograph, business report, book of fiction, minutes, periodicals, and thesis). Such focused genre palettes provide a well-defined testing ground for genre identification but in a limited scope. Moreover, the conclusions drawn from experiments on a focused genre palette may be misleading since features found to be useful in some genres are not equally effective in other genres (Boese and Howe, 2005).

Usually, the feature set of a webpage genre detection study is a combination of different features. To represent the plain text information, a wide variety of features have been proposed, following the practice of other text categorization tasks, including term frequencies (or a bag-of-words representation) (Boese and Howe, 2005; Finn and Kushmerick, 2006), common (or 'function') word frequencies (Kennedy and Shepherd, 2005; Santini, 2007; Levering, *et al.*, 2008), genre-prolific words (Kim and Ross, 2008), and frequencies of certain symbols, e.g., punctuation marks (Meyer zu Eissen and Stein, 2004; Santini, 2007). Such features are quite easy to compute given the availability of a tokenizer and perhaps a stemmer (Dong, *et al.*, 2008). However, for many natural languages (e.g., Chinese, Japanese) the tokenization task is especially difficult. Additionally, more elaborate features requiring some kind of text analysis include sentence length (Finn and Kushmerick, 2006), part-of-speech frequencies (Meyer zu Eissen and Stein, 2004; Lim, *et al.*, 2005; Boese and Howe, 2005; Finn and Kushmerick, 2006; Santini, 2007; Levering, *et al*., 2008), part-of-speech *n*-grams (Santini, 2007), named-entity (e.g., names, dates) frequencies (Meyer zu Eissen and Stein, 2004; Lim, *et al*., 2005), phrase length (Lim, *et al.*, 2005), sentence types (Lim, *et al.*, 2005), etc. These syntax-related features depend on the availability of appropriate natural language processing tools for a specific language or domain. Such tools introduce noise in the estimation of the feature values, because they are still imperfect. On the other hand, useful stylistic features may become available as a side-effect of automated text analysis. For example, Lim, *et al.* (2005) propose the average number of syntactic trees per sentence produced by a syntactic analyzer as a means to quantify the syntactic ambiguities.

A great variety of features has been proposed also to represent the structure (i.e. presentation, or layout) of webpages such as HTML tag frequencies, metatag frequencies, script features, e.g., use of JavaScript, link analysis features, e.g., number of external links, number of email links, URL-related features, e.g. URL depth, form analysis features, e.g., number of input forms, image counts (Meyer zu Eissen and Stein, 2004; Lim, *et al*., 2005; Kennedy and Shepherd, 2005; Boese and Howe, 2005; Santini, 2007; Levering, *et al.*, 2008, Dong, *et al.*, 2008). In general, such measures can be easily extracted from webpages and provide useful information on their stylistic properties. However, they cannot be applied to a significant part of non-HTML web documents (e.g., PDF, DOC, PPT files). Moreover, so far these features have to be defined manually.

Another source of information comes from the various webpage objects and their position. Lim, *et al.* (2005) focused on the usefulness of information in different parts of the webpages (title, body, anchor text, etc.) and they found that the main body and anchor text parts provide the most effective features. Kim and Ross, (2008) propose a transformation of the first page of PDF documents to a bit map in order to identify the text regions of the page. In another recent study, Levering *et al*. (2008) propose the use of visual features that capture the layout characteristics of the genres. The visual features attempt to represent the rendered location and size of webpage objects. Their experiments showed that visual features improve the classification results.

Some approaches to webpage genre detection use an arbitrarily-defined feature set, usually of small size comprising a few dozens of features (Meyer zu Eissen & Stein, 2004; Santini, 2007). Therefore, such approaches require a manually update of the feature set to be able to adapt to current conditions of webpage genres. Since the web is a rapidly changing environment, this procedure is of vital importance. Alternatively, other approaches begin with a large feature set comprising hundreds or thousands of features and then a feature selection algorithm is used to extract the most useful features (Lim, *et al*., 2005; Boese & Howe, 2005; Kennedy and Shepherd, 2005; Levering, *et al.*, 2008). In that case, the resulting feature set is optimized for a particular corpus and a particular genre palette since features are selected according to their discriminatory ability. Hence, it is not efficient to use a feature set extracted from one corpus to another corpus. What is needed is a general and fully-automated methodology to extract stylistic features that do not depend on certain corpora, genres, or natural languages. As a result, such a feature set would be exportable, meaning that it could be extracted from one corpus of a specific genre palette and effectively be used for another corpus possibly of a different genre palette. The methodology presented in the next section aims towards this direction.

3. OUR APPROACH

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

3.1. Character *n*-grams of Variable Length

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 (2006) 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 do not 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 (more frequent, thus more probable) than the 'glue' of the *n*-gram |thea|[1] (less probable).

It has to be underlined that the method described in this section for extracting dominant character *n*-grams is frequency-based. Class information is not used for selecting the features (i.e., the discriminatory ability of character *n*-grams is not taken into account). This is in contrast to many feature selection algorithms, such as *information gain* or *odds-ratio* (Forman, 2003). As a result, the features extracted by the proposed algorithm from a given corpus can be portable to other corpora based on a different genre palette.

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 (1999) 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 successor 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:

$$\begin{array}{l} if\left(C.length > 3\right) \\ g\left(C\right) \geq g\left(ant\left(C\right)\right) \wedge g\left(C\right) > g\left(succ\left(C\right)\right),\ \forall ant\left(C\right), succ\left(C\right) \\ if\left(C.length = 3\right) \\ g\left(C\right) > g\left(succ\left(C\right)\right),\ \forall succ\left(C\right) \end{array} \qquad (1)$$

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*-

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

grams. So, it is expected that the proposed algorithm will favor 3-grams and 5-grams against 4-grams.

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.* (1999). The SCP of a bigram |xy| is the product of the conditional probabilities of each given the other:

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

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 * denotes 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:

$$SCP((c_1 \ldots c_{n-1}), c_n) = \frac{p(c_1 \ldots c_n)^2}{p(c_1 \ldots c_{n-1}) \cdot p(c_n)} \tag{3}$$

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 to account for any possible dispersion point (since this measure is based on fair dispersion point normalization, it will be called *fairSCP*). Hence, the *fairSCP* of the *n*-gram $|c_1 \ldots c_n|$ is as follows:

$$fairSCP(c_1 \ldots c_n) = \frac{p(c_1 \ldots c_n)^2}{\frac{1}{n-1} \sum_{i=1}^{i=n-1} p(c_1 \ldots c_i) \cdot p(c_{i+1} \ldots c_n)} \tag{4}$$

3.2. Structural Information

The structural information of a webpage is encoded into the HTML tags used in the 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 tags in that page.

Then, the ReliefF algorithm for feature selection (Robnik-Sikonja & Kononenko, 2003) is applied to the produced dataset to reduce the dimensionality. This algorithm is able to detect conditional dependencies between attributes and is noise tolerant. In this way, we avoid any manual definition of useful HTML tags and the whole procedure is fully-automated. In addition, the extracted HTML tags are adapted to the particular properties of a specific corpus.

4. EXPERIMENTS

4.1 Webpage Genre Corpora

Although there is not yet a large reference corpus covering a wide variety of web genres, several small webpage corpora appropriate for evaluating genre identification approaches have become available (Rehm, *et al.*, 2008). Despite the fact that there are significant differences in the methodologies used to define the genres and populating the genre categories with webpages, such publicly-available corpora provide an objective testing ground for comparing different approaches. In this paper, we make use of two corpora, already used in 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) (Santini, 2007). 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 (Meyer zu Eissen & Stein, 2004). 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 to PORTRAYAL-PRIVATE and SHOP of KI-04, respectively). However, the exact relationships between all the genres defined in these two corpora are not clearly defined.

As already mentioned, several researchers used these corpora to evaluate their methods (Meyer zu Eissen & Stein, 2004; Boese & Howe, 2005; Santini, 2007; Kim & Ross, 2007; Mason, et al., 2009). Dong, *et al.*, (2008) used a part of 3 genres from 7genre. The so far reported results based on cross-validation are summarized in Table 2.

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

4.2 Words vs. Character *n*-grams

In the first experiment, we compare the performance of the variable-length *n*-gram approach with a traditional bag-of-words method. In more detail, for the character *n*-gram approach, 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). 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 (Joachims, 1998), was used for the classification. As concerns the word-based approach, the most-frequent words of the corpus in question formed the set of features. Then, a SVM model was applied to build the webpage genre classifier.

We examined binary as well as TF features for both schemes (i.e., character *n*-grams and words). Figures 1 and 2 show the classification accuracy results (based on stratified 10-fold cross-validation) on the 7genre and KI-04 corpora, respectively, for various feature set sizes. As can be seen, the binary features are better than the corresponding TF features in all cases. Moreover, the word features are better when the dimensionality remains low. This can be explained since word features are more semantically loaded in comparison to character *n*-grams. However, the best results are obtained by binary character *n*-grams when the dimensionality exceeds 4,000 features. In other words, character *n*-grams require higher dimensionality in comparison to word features but provide better accuracy. Tables 3 and 4 present the confusion matrices for the 7genre corpus based on binary character *n*-grams of variable length and the word-based model, respectively. Similarly, Tables 5 and 6 present the corresponding confusion matrices for KI-04.

4.3 Incorporating Structural Information

So far, only textual information was used. In the next experiment, the structural information extracted as described in Section 3.2 was incorporated to the character *n*-gram model (binary features). The evaluation results based on (microaverage) classification accuracy of stratified 10-fold cross-validation for 7genre and KI-04 are presented in Figures 3 and 4, respectively. In addition to the performance of the combined model of textual and structural information, the performance of the textual only model (i.e., binary character *n*-grams) as well as the word-based model (i.e. binary features) are also depicted for comparative purposes.

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). The structural information assists the character *n*-gram model to perform better than the word-based model even in low dimensional spaces. 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. Tables 7 and 8 present the confusion matrices for 7genre and KI-04 based on the character *n*-gram plus structural information model, respectively.

4.4 Cross-check Experiments

The previous experiments exhibited the effectiveness of our approach. However, the resulting models are based on high dimensional spaces (several thousands of features) to achieve the reported results. Moreover, the features used in our method are not pre-defined but they are extracted from a given corpus, therefore, they are optimized for the specific experiments. These facts in combination with the small size of the corpora used for evaluation and the specific evaluation method that was followed (cross-validation) could raise questions on the robustness and credibility of our approach. To examine the generality of the proposed method and the extracted features, we performed a series of cross-check experiments based on the two publicly-available corpora described in Section 4.1. In more detail, given two corpora A and B, each one having its own genre palette, the following types of cross-check experiments can be defined:

- Type 1: Training the webpage genre detection model using the genre palette of corpus A and applying the model to the corpus B, and vice versa. This experiment would reveal the robustness of the examined method since similarities and differences between genres belonging to different genre palettes should be identified. In more detail, a model trained on one genre palette so that to optimize its performance on a specific corpus is likely to overfit properties of that corpus and lose part of its generality. By testing such a model on a different genre palette, we can check whether it maintains its generality. In such a case, the model will be able to identify genres with certain similarities (e.g., the genres E-SHOP of 7genre and SHOP of KI-04).

- Type 2: Training the webpage genre detection model for the corpus A using the features extracted by the corpus B and testing its performance in corpus A by cross validation, and vice versa. This experiment would evaluate the credibility of the proposed method since the detection model should perform equally (or comparably) well based on the feature set extracted by another corpus, that is not optimized for the specific experiment. Taking into consideration the high dimensionality of the character *n*-gram representation, this type of experiment will show whether these features are general enough so that to be extracted from a reference corpus and applied to a new corpus.

Therefore, we first trained a model using the KI-04 corpus and then applied it to detect the genres of the 7genre corpus based on binary character *n*-gram features. The results are shown in the confusion matrix of Table 9. Note that this is not a typical confusion matrix since the row and column labels are different. Apparently, most of the genre similarities between the palettes of these two corpora are correctly identified: BLOG pages are classified as DISCUSSION and HELP pages; ESHOPs of 7genre are mainly classified as SHOP and PORTRAYAL-NON-PRIVATE pages of KI-04; DOWNLOAD pages are found similar to HELP pages; ONLINE NEWSPAPER FRONTPAGE of 7genre is divided into LINK COLLECTION, PERSONAL-NON-PRIVATE and SHOP of KI-04; LISTING pages are found more similar to LINK COLLECTION pages; PERSONAL HOME PAGES of KI-04 are mainly classified as PORTRAYAL-PRIVATE of 7genre; finally, SEARCH pages of 7genre are found similar with LINK COLLECTIONS and PORTRAYAL-NON-PRIVATE of KI-04. In other words, the main stylistic similarities of the genres used for training and test are indicated.

Similarly, the 7genre was used as training corpus and then the extracted model was applied to the KI-04 corpus based on binary character *n*-gram features. The results are shown in Table 10. Here, it is obvious that the LISTING genre overshadowed most of the genre similarities. Recall from Section 4.1 that this category can be decomposed into many sub-genres. It is considered, therefore, a super-genre and as can be seen in the results of this experiment it confused the genre detection model. On the other hand, the basic genre similarities between the two palettes are still indicated: SHOP pages of KI-04 are found similar to E-SHOP pages of 7genre; PORTRAYAL-PRIVATE pages of KI-04 are mainly classified as PERSONAL HOME PAGES of 7genre.

For the second type of cross-check experiment, we used the character *n*-gram features extracted from KI-04, following the procedure described in 3.1, to train a webpage detection model based on the 7genre corpus. The microaverage classification accuracy results are now 95.2%, as opposed to 96.2% when the character *n*-gram features are extracted from 7genre (see

Table 2). Hence, the performance remains high although the features are extracted from a different corpus and, more significantly, with a different genre palette. Similarly, when training the proposed webpage detection model on KI-04 using the character *n*-gram features extracted from the 7genre corpus, the classification accuracy is 82.4%, as opposed to 82.8% when using character *n*-gram features extracted from KI-04. Again, it is confirmed that the proposed approach is able to achieve high performance even in cases where the feature set has not been optimized for a specific corpus. Tables 11 and 12 show the confusion matrices for these cross-check experiments and can be directly compared with Tables 3 and 6 for 7genre and KI-04, respectively. As can be seen, there are no significant differences in the performance of the genre detection models when features extracted from a different corpus are used. Approximately, 2/3 of the character *n*-gram features extracted from 7genre and KI-04 overlap. Recall from Section 3.1 that the procedure of extracting the character *n*-gram features is based on their frequency rather than their discriminatory ability.

To have additional evidence on this, we performed another experiment using the publicly-available 20-Genre corpus (Vidulin, *et al*., 2007) which comprises 1,539 webpages labeled by 20 genre categories. In contrast to 7genre and KI-04, this corpus is multi-labeled (each webpage can belong to more than one categories). Following the method of Vidulin, *et al*. (2007) we built 20 binary classifiers to discriminate each genre from all the remaining genres. Each genre detection model was based on character *n*-gram features only. Two features sets were tested, namely: the (binary) character *n*-grams extracted from 20-Genre corpus and the character *n*-grams extracted from KI-04 corpus. The cross-validation results in terms of precision, recall, and F-measure are shown in Table 13 and can be directly compared to the corresponding results reported by Vidulin, *et al*. (2007). As expected, the feature set extracted from 20-Genre itself achieves better results in comparison to the features extracted by KI-04. However, the KI-04 features achieve comparable average performance verifying that it is possible to use a general set of character *n*-grams. Moreover, both feature sets produce better (average) results than those reported by Vidulin, *et al*. (2007).

5. DISCUSSION

The webgenre genre detection models presented in the previous section were evaluated in terms of 10-fold cross-validation. Table 2 summarizes the best accuracy results achieved by the textual models (words or character *n*-grams only) and the textual plus structural models together with the best reported results on the same corpora (based in cross-validation as well) by other researchers. It is clear that the character *n*-grams in combination with structural features (textual plus structural model) achieve the best performance, significantly better than previously reported results. However, the model based only on binary character *n*-gram features also achieved good results.

In order to have a more detailed picture of the classification derived by the produced models, we examine the confusion matrices of Tables 3, 4, and 7 for 7genre and Tables 5, 6, and 8 for KI-04. For comparative purposes, similar confusion matrices of previously reported results on the same corpora can be found in Meyer zu Eissen & Stein (2004) and Santini (2007). 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 by Meyer zu Eissen & Stein (2004). 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 Meyer zu Eissen & Stein (2004), that is 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 Meyer zu Eissen & Stein (2004). Apparently, the textual+structure model achieves a slightly increased performance in many cases. Most notably, it solves the confusion between certain pairs (PHP – SHOP) and reduces the cells with non-zero values (excluding the diagonal that corresponds to correct answers).

Concerning 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 by Santini (2007) the best cases were DOWNLOAD and ARTICLE and the worst cases were (again) 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 Santini (2007). However, the pair PORTRAYAL-PRIVATE – LINK COLLECTION that was also a confusing factor according to the results of Santini (2007), does not contribute significantly to the misclassifications of our approach.

Moreover, the comparison of the performance of the textual only model and the textual plus 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.

The cross-check experiments (Section 4.4) were performed to verify that the effectiveness of the produced models is not corpus dependent (optimized by cross-validation and the large number of features). More specifically, the success of the cross-check experiment of type 1 (training on and testing on another genre palette) shows that the knowledge acquired by the genre detection model trained on one genre palette is able to identify genres with similar properties included in another genre palette. Although this is only intuitive evidence, it shows that despite the large number of features and the small number of texts in the used corpora, the produced models acquired general knowledge on webpage genre properties.

In addition, the type 2 cross-check experiments show that even when we use a character *n*-gram feature set extracted from another corpus, the webpage genre detection results are not significantly decreased. This can be explained since the feature selection algorithm used to extract the character *n*-grams of variable–length is based on the frequency rather than the discriminatory ability of features. As a consequence, it is possible to have a common set of character *n*-grams extracted from one corpus and then use this feature set in any other corpus (possibly of a different genre palette) of the same natural language without a significant decrease in performance. This would simplify the application of the proposed approach to a new corpus and would considerably decrease the training time cost.

6. CONCLUSIONS

In this paper we presented a model based on low-level information, that is, character *n*-grams of variable-length and HTML tags, for the task of webpage genre identification. In comparison to previous work, 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 webpages. Therefore, it can be adapted to the specific properties of a given (either general or focused) collection of genres. The training time cost of extracting variable-length character *n*-gram features is analogous to the size of training corpus and the size of the initial set of *n*-grams (as described in Section 3.1). On the other hand, the success of the type 2 cross-check experiment indicated that it is possible to use a standard character *n*-gram feature set extracted from a reference corpus in any arbitrarily selected genre collection. This would also significantly decrease the training time cost of our approach.

The proposed features are very effective in capturing the stylistic properties of webpages since they can deal with noisy text and can be adapted to the needs of the still evolving web genres. An important advantage of the proposed features is that they do not require any complicated language processing avoiding problems with the availability of accurate natural language processing tools for any text domain. Moreover, they are language-independent and can be directly applied in eastern languages too (where the tokenization task is extremely difficult). Experiments on two corpora show that character *n*-gram features are better genre discriminators than words, although they increase considerably the dimensionality of the problem. Moreover, binary features are found to perform better than TF features for both schemes in all the experiments we conducted. This fact reduces the complexity of an automatic genre identification tool that could deal with millions of pages.

The combination of character *n*-gram features with structural features further improves the classification accuracy. However, the models based on character *n*-gram features alone are quite effective. This is particularly important given the large amount of non-HTML documents in the web, indicating that textual information alone could identify the genre of webpages. However, this remains to be verified in larger and more heterogeneous corpora.

The proposed approach was also evaluated in a novel way. In addition to cross-validation results that are in accordance with previous work in the area, a series of cross-check experiments were performed, each one requiring at least two different corpora. Those experiments exhibited the robustness and credibility of our method since the performance remains on a high level even when the feature set is not optimized for the specific corpus used for the evaluation. In addition, when a certain genre palette is used to train the proposed method and the produced model is applied to another genre palette, the most important similarities and differences between the genres were identified. The latter procedure could offer clues for a deeper understanding of the properties of webpage genres and the relationships among genres taken from different genre palettes. Another promising future work direction is the application of semi-supervised learning techniques to the webpage genre identification task since there is plethora of unlabeled data available.

ACKNOWLEDGEMENT

The authors wish to thank the anonymous IPM reviewers for their valuable and insightful comments that significantly improved an earlier draft of this paper.

REFERENCES

Bekkerman, R. (2008). *Combinatorial Markov Random Fields and their Applications to Information Organization*. Ph.D. Thesis, University of Massachusetts Amherst.

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

Braslavski, P. (2007). "Combining Relevance and Genre-Related Rankings: An Exploratory Study". In *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines: The Impact of NLP*, pp.1-4.

Clark, M. and S. Watt (2007). "Classifying XML Documents by Using Genre Features". In *Proc. of the 4th Int. Workshop on Text-based Information Retrieval*.

Dewdney, N., C. VanEss-Dykema, and R. MacMillan (2001). "The Form is the Substance: Classification of Genres in Text", *Workshop on Human Language Technology and Knowledge Management* (39th ACL-10th EACL), pp. 53-61.

Dong L., Watters C., Duffy J. and Shepherd M. (2008). "An Examination of Genre Attributes for Web Page Classification". In *Proc. of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008).*

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

Forman, G. (2003). "An Extensive Empirical Study of Feature Selection Metrics for Text Classification". *Journal of Machine Learning Research*, 3, 1289-1305.

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

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

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

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

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

Kim Y. and Ross S. (2008). "Examining Variations of Prominent Features in Genre Classification. In *Proc. of the 41st Annual Hawaiian International Conference on System Sciences (HICSS 2008).*

Kim Y. and Ross S. (2007). "Variations of Word Frequencies in Genre Classification Tasks". In *Proc. of the DELOS conference on Digital Libraries.*

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

Levering, R., M. Cutler, and L. Yu (2008). “Using Visual Features for Fine-Grained Genre Classification of Web Pages”. In *Proc. of the 41st Hawaii International Conference on System Sciences*.

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

Mason, J.E., M. Shepherd, and J. Duffy (2009). “An N-gram Based Approach to Automatically Identifying Web Page Genre”. In *Proc. of the 42nd Hawaii International Conference on System Sciences*.

Meyer zu Eissen, S. (2007). *On Information Need and Categorizing Search*. Ph.D. Thesis, Department of Computer Science, University of Paderborn.

Meyer zu Eissen, S. and B. Stein (2004). “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, pp. 256-269.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R, Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). “Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems”. In *Proc. of the 6th Int. Conference on Language Resources and Evaluation (LREC 2008)*.

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

Rosso, M. (2005). *Using Genre to Improve Web Search,* Ph.D. Thesis, University of North Carolina at Chapel Hill.

Sanitni, M. (2007). *Automatic Identification of Genre in Webpages*. Ph.D. Thesis, University of Brighton.

Santini, M., (2008). “Zero, Single, or Multi? Genre of Web Pages through the Users’ Perspective”, *Information Processing and Management,* 44(2), pp. 702-737.

Sebastiani, F. (2002). “Machine Learning in Automated Text Categorization”. ACM Computing Surveys, 34(1).

Shepherd M. and Watters C. (1998). “The Evolution of Cybergenres”. In *Proc. of the 31st Hawaii International Conference on System Sciences*.

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

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

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

Stamatatos, E. (2006). “Ensemble-based Author Identification Using Character N-grams”. *Proc. of 3rd Int. Workshop on Text-based Information Retrieval,* pp. 41-46.

Swales J. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge University Press.

Vidulin, V., M. Lustrek, and M. Gams (2007). “Using Genres to Improve Search Engines”. In *Proc. of the Int. Workshop Towards Genre-enable Search Engines: The Impact of Natural Language Processing*.

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

Table 1. The webpage genre corpora used in this study.

Approach	**7Genre**	**KI-04**
(Meyer zu Eissen & Stein, 2004)	-	70.0%
(Boese & Howe, 2005)	-	74.8%
(Santini, 2007)	90.6%	68.9%
(Kim & Ross, 2007)	92.7%	-
(Mason, *et al*., 2009)	94.6%	-
Proposed in this paper		
Character *n*-grams – Binary	96.2%	82.8%
Character *n*-grams – TF	92.5%	79.6%
Words – Binary	95.5%	82.0%
Words – TF	95.1%	81.8%
Textual + structural	**96.5%**	**84.1%**

Table 2. Previously reported results for the corpora together with the best results found in this study. The results from (Boese, 2005) refer to a subcorpus of KI-04.

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 3. Confusion matrix for the classification of 7genre corpus using character *n*-gram binary features (overall accuracy: 96.2% and F-measure: 0.96).

Actual	**Classified as**						
	BLOG	**ESHOP**	**FAQ**	**ONF**	**LIST**	**PHP**	**SEARCH**
BLOG	97.5%				0.5%	1.5%	0.5%
ESHOP		95.0%			1.5%	0.5%	3.0%
DOWN			99.0%		0.5%	0.5%	
ONF				100.0%			
LIST	1.0%	3.0%	0.5%		88.0%	4.0%	3.5%
PHP	1.0%	0.5%			3.5%	94.5%	0.5%
SEARCH	0.5%	1.5%		0.5%	2.5%	0.5%	94.5%

Table 4. Confusion matrix for the classification of 7genre corpus using word binary features (overall accuracy: 95.5% and F-measure: 0.95).

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 5. Confusion matrix for the classification of KI-04 corpus using character *n*-gram binary features (overall accuracy: 82.8% and F-measure: 0.83).

Actual	**Classified as**							
	ART	**DOWN**	**LINK**	**P-P**	**DISC**	**HELP**	**P-NP**	**SHOP**
ART	82.7%		2.4%	3.9%	0.8%	4.7%	5.5%	
DOWN		94.7%		1.3%	0.7%	0.7%	2.0%	0.7%
LINK	1.5%	2.0%	82.4%	2.0%	0.5%	4.9%	4.9%	2.0%
P-P	5.6%	0.8%	0.8%	86.5%			5.6%	0.8%
DISC		0.8%	3.1%	1.6%	91.3%	1.6%	1.6%	
HELP	6.5%	2.9%	7.9%	2.2%	2.2%	71.2%	4.3%	2.9%
P-NP	2.5%	1.8%	6.7%	9.2%	1.8%	3.1%	66.9%	8.0%
SHOP		2.4%	3.6%	3.0%		1.8%	6.6%	82.6%

Table 6. Confusion matrix for the classification of KI-04 corpus using word binary features (overall accuracy: 82.0% and F-measure: 0.82).

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 7. Confusion matrix for the classification of 7genre corpus using a combination of character *n*-gram and structural features (overall accuracy: 96.6% and F-measure: 0.97).

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%

Table 8. Confusion matrix for the classification of KI-04 corpus using a combination of character *n*-gram and structural features (overall accuracy: 84.1% and F-measure: 0.84).

Actual	Classified as							
	ART	**DOWN**	**LINK**	**P-P**	**DISC**	**HELP**	**P-NP**	**SHOP**
BLOG	5.0%		1.0%	1.5%	69.0%	21.0%	1.5%	1.0%
ESHOP		0.5%		2.5%	0.5%	0.5%	14.0%	82.0%
DOWN	6.0%	0.5%	4.0%	1.0%		82.0%	6.5%	
ONF			39.5%			8.0%	35.0%	17.5%
LIST	7.5%	1.0%	49.5%	3.5%	1.0%	17.5%	15.5%	4.5%
PHP	1.5%	1.0%	4.5%	71.0%	2.0%	1.0%	12.0%	7.0%
SEARCH		0.5%	57.5%	6.5%	0.5%	1.5%	22.5%	11.0%

Table 9. Confusion matrix for the classification of 7genre corpus using a model trained on KI-04 corpus.

Actual	Classified as						
	BLOG	**ESHOP**	**FAQ**	**ONF**	**LIST**	**PHP**	**SEARCH**
ART	3.1%	0.8%			78.0%	17.3%	0.8%
DOWN		32.5%			47.0%	12.6%	7.9%
LINK	0.5%	3.9%		0.5%	77.6%	3.9%	13.7%
P-P	0.8%	0.8%			7.1%	88.9%	2.4%
DISC	13.4%	13.4%		0.8%	40.9%	22.0%	9.4%
HELP	8.6%	2.9%	2.2%		73.4%	5.0%	7.9%
P-NP		26.4%		2.5%	42.9%	17.2%	11.0%
SHOP	0.6%	71.9%		0.6%	15.0%	4.8%	7.2%

Table 10. Confusion matrix for the classification of KI-04 corpus using a model trained on 7genre corpus.

Actual	**Classified as**						
	BLOG	**ESHOP**	**FAQ**	**ONF**	**LIST**	**PHP**	**SEARCH**
BLOG	96.5%	0.5%			0.5%	2.5%	
ESHOP		95.5%			1.0%	2.0%	1.5%
DOWN			99.0%		1.0%		
ONF				100.0%			
LIST	1.0%	1.0%	0.5%		88.5%	5.5%	3.5%
PHP	1.5%				2.0%	96.0%	0.5%
SEARCH	0.5%	1.5%		0.5%	4.0%	1.0%	92.5%

Table 11. Confusion matrix for the classification of 7genre corpus using character *n*-gram features extracted by KI-04 corpus.

Actual	**Classified as**							
	ART	**DOWN**	**LINK**	**P-P**	**DISC**	**HELP**	**P-NP**	**SHOP**
ART	82.7%		2.4%	2.4%	0.8%	7.1%	4.7%	
DOWN		92.1%		0.7%		2.0%	3.3%	2.0%
LINK	0.5%	1.5%	85.4%	1.5%		3.4%	5.9%	2.0%
P-P	5.6%	0.8%	1.6%	88.9%			3.2%	
DISC	1.6%	0.8%	2.4%	0.8%	89.8%	3.1%	1.6%	
HELP	7.2%	2.9%	6.5%	2.2%	2.2%	71.2%	5.0%	2.9%
P-NP	1.2%	2.5%	6.1%	8.6%	1.8%	2.5%	70.6%	6.7%
SHOP	0.6%	2.4%	1.8%	2.4%		1.8%	10.8%	80.2%

Table 12. Confusion matrix for the classification of KI-04 corpus character *n*-gram features extracted by 7genre corpus.

	20-Genre Features			**KI-04 Features**		
Genre	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
adult	0.82	0.46	0.56	0.83	0.35	0.47
blog	0.88	0.41	0.52	0.85	0.35	0.47
childrens	0.74	0.47	0.56	0.68	0.50	0.57
commercial-promotional	0.28	0.11	0.15	0.23	0.11	0.15
community	0.88	0.42	0.56	0.89	0.43	0.57
content-delivery	0.43	0.28	0.32	0.41	0.30	0.33
entertainment	0.25	0.06	0.10	0.12	0.06	0.08
error-message	0.67	0.56	0.60	0.60	0.58	0.58
FAQ	0.96	0.60	0.73	0.95	0.61	0.73
gateway	0.49	0.22	0.29	0.50	0.20	0.27
index	0.36	0.21	0.26	0.42	0.24	0.30
informative	0.58	0.18	0.27	0.52	0.18	0.26
journalistic	0.79	0.52	0.62	0.75	0.51	0.61
official	0.78	0.45	0.56	0.83	0.44	0.56
personal	0.97	1.00	0.98	0.97	1.00	0.98
poetry	0.98	1.00	0.99	0.98	1.00	0.99
prose-fiction	0.98	1.00	0.99	0.98	1.00	0.99
scientific	0.98	1.00	0.99	0.98	1.00	0.99
shopping	0.98	1.00	0.99	0.98	1.00	0.99
user-input	0.97	0.99	0.98	0.98	0.99	0.98
average	*0.74*	*0.55*	*0.60*	*0.72*	*0.54*	*0.59*

Table 13. Performance results on the multi-labeled 20-Genre corpus. Precision, Recall, and F-measure for character *n*-gram features extracted from 20-Genre and KI-04 corpora.

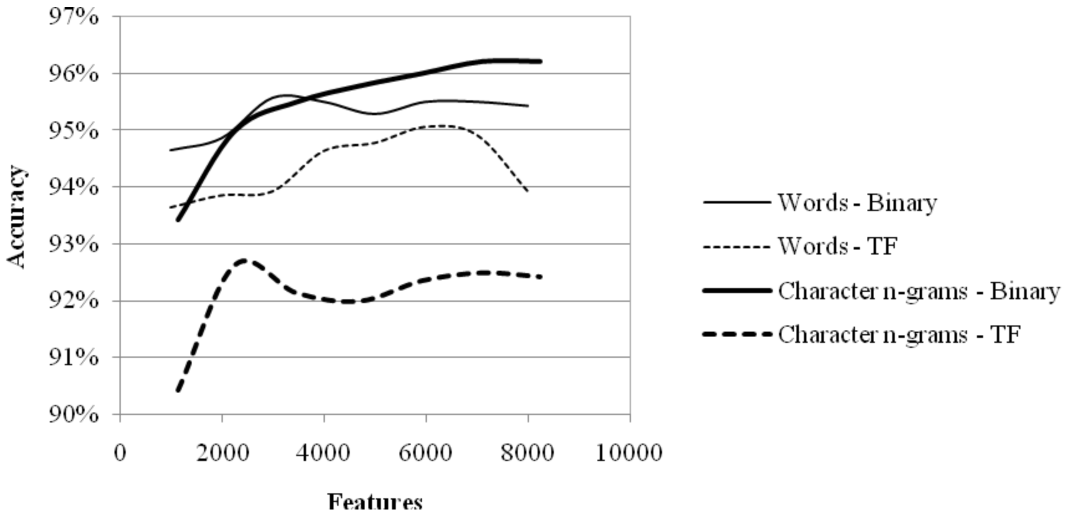

Figure 1. Webpage genre detection results on 7genre corpus based on word and character *n*-gram features (binary or TF) for different sizes of feature set.

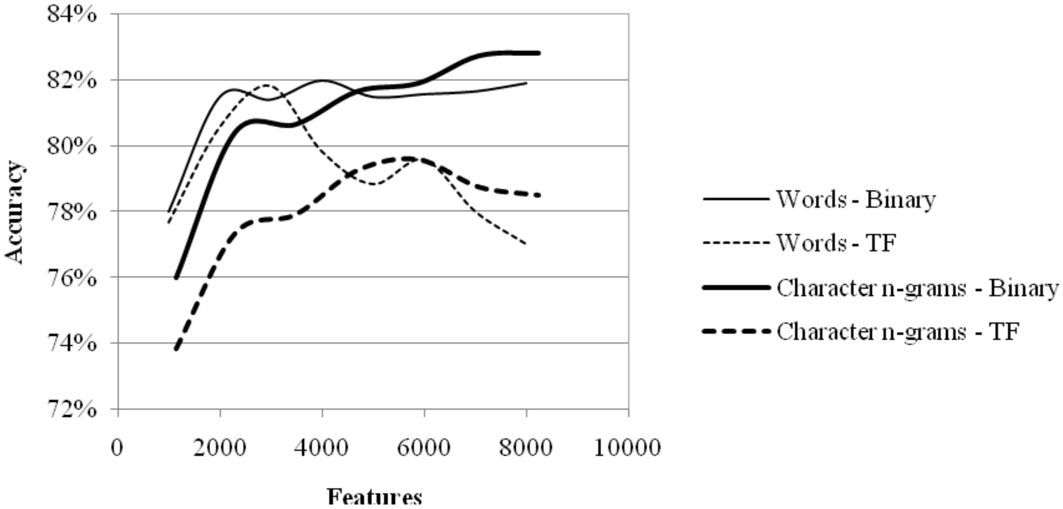

Figure 2. Webpage genre detection results on KI-04 corpus based on word and character *n*-gram features (binary or TF) for different sizes of feature set.

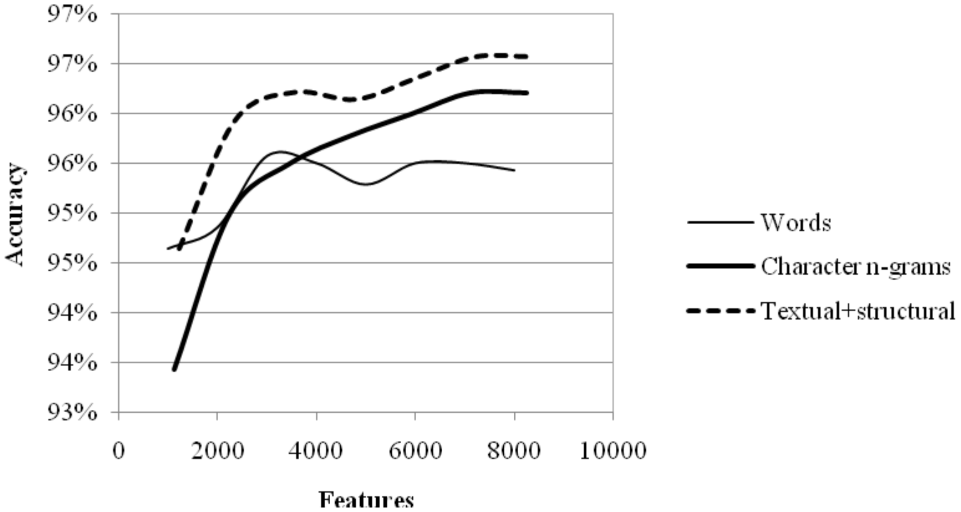

Figure 3. Webpage genre detection results on 7genre corpus based on word and character *n*-gram features as well as combination of character *n*-grams and structural features.

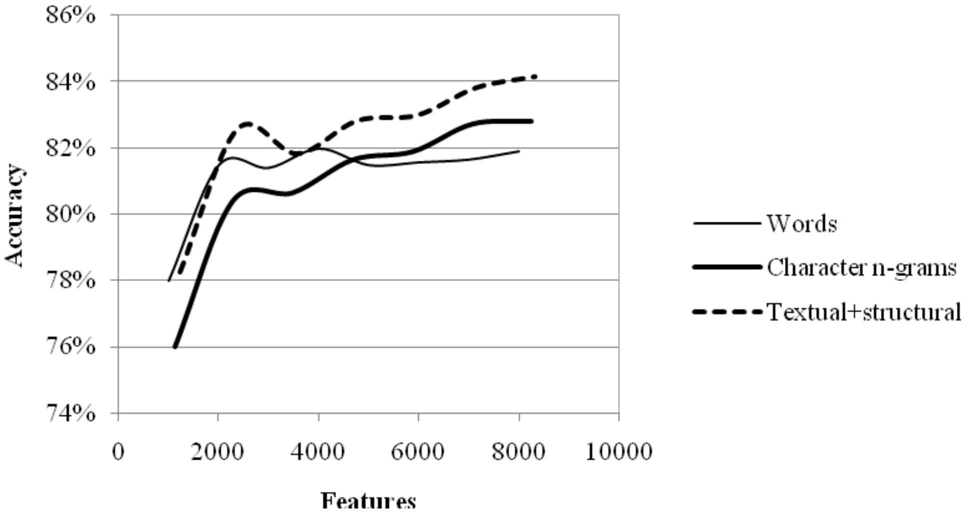

Figure 4. Webpage genre detection results on KI-04 corpus based on word and character *n*-gram features as well as combination of character *n*-grams and structural features.