Feature Selection for an n-gram Approach to Web Page Genre Classification

Jane E. Mason
Michael Shepherd
Jack Duffy

Technical Report CS-2009-04

June 22, 2009

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

Feature Selection for an n-gram Approach to Web Page Genre Classification

Jane E. Mason[1], Michael Shepherd[2], and Jack Duffy[3]

[1] Dalhousie University, Halifax, NS Canada jmason@cs.dal.ca
[2] Dalhousie University, Halifax, NS Canada shepherd@cs.dal.ca
[3] Dalhousie University, Halifax, NS Canada jack.duffy@dal.ca

Abstract. Web page genre classification is a potentially powerful tool in filtering the results of online searches. In this paper, we describe a set of experiments investigating the automatic classification of Web pages by their genres using n-gram representations of the Web pages and Web page genres, and a distance function classification model. The experiments in this study examine the effect of three feature selection measures on the accuracy of Web page classification with this model. The feature selection measures which are investigated include frequency, the Chi-square statistic, and Information Gain. The experiments are run on two well-known data sets, 7-Genre and KI-04, for which published results are available. Our results compare very favorably with those of other researchers.

1 Introduction

The research reported in this paper focuses on the automatic classification of Web pages by their genres, using n-gram representations of the Web pages and Web page genres, and a distance function classification model. The goal of the research in this study is not only to determine the usefulness of n-gram Web page representations, but also to compare and assess the effectiveness of three feature selection measures in choosing the n-grams with which to represent each Web page. The feature selection measures to be tested are frequency, the Chi-square statistic, and Information Gain. We run our experiments on the popular 7-Genre and KI-04 data sets. We also examine a range of n-gram lengths and a range of Web page profile sizes to determine what combination(s) of feature selection measure, n-gram length, and Web page profile size give the best classification accuracy. Although other authors have looked at these factors individually, this study has looked at them in combination.

The remainder of the paper proceeds as follows. Section 2 gives a brief overview of related work, and Section 3 describes our classification model, discusses the feature selection measures to be tested, and provides details about the data sets and experiments. Section 4 focuses on presenting and analyzing the results of the experiments, while Section 5 summarizes our conclusions and describes the direction of our future work.

2 Related Work

For the task of classifying Web pages by genre, Web pages can be represented much like documents in text classification, however the Web page representations may also include information unique to Web pages, such as URL information or HTML tags. For example, Lee and Myaeng [8] represent Web pages using terms from a genre-specific core vocabulary, whereas Rehm [12] uses linguistic features combined with HTML meta data and presentation related tags. Meyer zu Eissen and Stein [11] combine genre-specific vocabulary and closed-class word sets with text statistics, part-of-speech information, and HTML tags, while Boese and Howe [1] use a bag of words representation augmented with other information that includes text statistics, readability scales, part-of-speech information, HTML tags, and URL information. Jebari [5] combines two centroid-based classifiers, one of which uses structural information from the document, such as the title, heading, and anchors, while the other uses URL information. Stein and Meyer zu Eissen [17] give a detailed chronological overview of the document representations used for genre classification on Web-based corpora.

The research discussed in this paper represents Web pages using fixed-length byte n-grams. These n-grams can be thought of as the contents of a fixed-size sliding window moved through the text. The use of n-grams has been common in language modeling since at least 1948 when Claude Shannon, considered the father of information theory, investigated the question of determining the likelihood of the next letter in a given sequence of characters [15]. Since that time, n-grams have been widely used in natural language processing and statistical analysis.

Closely related to the n-gram technique used in this paper is the work on n-gram based text classification by Cavnar and Trenkle [2], and on authorship attribution by Kešelj et al. [7]. In those cases, each document is represented by a profile of the L most frequent character n-grams. As with our classification model, the distance between two documents is determined using a distance function, although the methods for representing genres and authors differ. Houvardas and Stamatatos [4], also working on the problem of authorship attribution, proposed a new selection technique for variable-length character n-grams in which each n-gram is compared with similar n-grams (either longer or shorter) in the feature set and the most important of them is kept. Houvardas and Stamatatos adapted this selection technique from an existing approach, proposed by Silva and Lopes [16], for extracting multiword terms (i.e., word n-grams of variable length) from texts. Once the feature selection has been made, a support vector machine (SVM) is then trained on this reduced feature set, and the SVM model is applied to the test set. Their experimental results indicated that the proposed feature selection technique resulted in higher classification accuracy than when Information Gain was used as a feature selection measure. Kanaris and Stamatatos [6] applied the new feature selection technique for variable-length character n-grams to the problem of Web page genre identification. In their experiments, they tested two models: the first model uses only 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. The HTML tag information is constructed by first creating a list of all HTML tags that appear three or more times in the entire data set. Each Web page is represented by a vector of the HTML tag frequencies; the ReliefF feature selection algorithm [13] is then applied to reduce the dimensionality of the vectors. As with Houvardas and Stamatatos [4], after the feature sets are reduced, classification is performed using an SVM. Kanaris and Stamatatos report that the accuracy of the Web page genre classification using their technique is higher than that previously reported by researchers on the same data sets.

Once the type of representation for the Web page has been determined, it is often necessary to select a subset of features in order to reduce the dimensionality of the search space, and thus reduce the computational complexity of the problem. In practice, it is often necessary to use feature selection techniques to select a subset of relevant features for building robust learning models, because most standard machine learning techniques cannot be directly applied when the dimensionality is very high. Yang and Pedersen [18] provide a comparative study of the traditional feature selection techniques in text classification. Yang and Pedersen evaluated five feature selection measures: term selection based on document frequency, Information Gain, Mutual Information, the Chi-square test, and term strength. In their experiments, they found that Information Gain and the Chi-square test were the most effective measures for feature selection.

Focusing specifically on the classification of Web pages by genre, Dong et al. [3] evaluated three measures for feature selection. They compared the performance of Information Gain, Mutual Information, and the Chi-square statistic for selecting features for the binary classification of Web page genres. They found that although all three selection measures were capable of detecting small sets of discriminating features, when feature sets were as small as 5, only Information Gain and the Chi-square statistic were able to successfully select features that gave good performance. In this paper, we investigate the use of frequency, Information Gain, and the Chi-square statistic as feature selection techniques.

3 Methodology

3.1 Classification Model

This classification model represents Web pages and Web page genres as byte n-gram profiles, and uses a distance function to determine the similarity between two profiles. For the experiments reported in this paper, the n-grams of which the profiles are comprised are chosen based on one of three feature selection measures: frequency, the Chi-square statistic, or Information Gain. Each data set is partitioned into a training and test set. For each Web page in the test set, profiles are constructed that consist of the L top ranked fixed-length byte n-grams and their normalized frequencies, Chi-square statistics, or Information Gain values, depending on the feature selection measure being tested. The ini-

tial n-gram profiles are produced using the Perl package `Text:Ngrams`[4], which normalizes the n-gram frequencies by dividing by the total number of n-grams of the given length. These byte n-grams are raw character n-grams in which no bytes are ignored, including the whitespace characters, thus byte n-grams capture some of the structure of a document.

Each genre in a data set is also represented using a profile of fixed-length byte n-grams and their corresponding frequency, Chi-square statistic, or Information Gain value. 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 centroid 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 their frequency (or Chi-square statistic, or Information Gain) values, and the genre profiles are then truncated to the size of the smallest of the genre profiles. See Tables 4 and 5 for a comparison of the sizes of the smallest and largest centroids.

The profiles for each Web page in the test set are compared with each centroid genre profile from the training set. The Web page is assigned the label of the genre profile to which it is closest (most similar), according to a distance measure. Based on extensive experimentation with a variety of distance functions, we compute the distance between two n-gram profiles using the formula suggested by Kešelj et al. [7]. The distance between two profiles is defined as

$$d(P_1, P_2) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{2 \cdot (f_1(m) - f_2(m))}{f_1(m) + f_2(m)} \right)^2, \quad (1)$$

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 Feature Selection Measures

Although Mason et al. [10] achieve high classification accuracy using frequency as a feature selection method for Web page classification by genre, our hypothesis is that a more theoretically sound feature selection measure could be more effective. Thus, we investigate not only frequency, but also the Chi-square statistic, and Information Gain as feature selection measures. The Chi-square statistic is a statistically based measure, while Information Gain is probability based. See, for example, Yang and Pedersen [18] for further details.

3.3 Data Sets

For the experiments reported in this paper, we use two established data sets for which published results are available. These data sets, known as the 7-Genre

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

and KI-04 data sets, are available online[5]. In each case, the unit of analysis is the individual Web page, and each Web page is labeled with one, and only one, genre label.

The 7-Genre data set, constructed by Marina Santini and described in Santini [14], 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.

The KI-04 data set, constructed by Meyer zu Eissen and Stein [11] has eight genres suggested by participants in a user study on the usefulness of Web page genres. These genres are ARTICLE, DISCUSSION, DOWNLOAD, HELP, LINK COLLECTION, PORTRAIT (NON-PRIVATE), PORTRAIT (PRIVATE), AND SHOP. The original corpus includes some empty Web pages, and so we follow the lead of Santini [14] and Jebari [5] in using the 1205 non-empty pages. The number of Web pages per genre ranges from 126 to 205. The Web pages in this data set include supplementary tagged information, such as the title, genre, and a plain text summary, all of which was removed prior to processing.

3.4 Experiments

The experiments reported in this paper are performed on the 7-Genre and KI-04 data sets, using 10-fold cross validation in each case. This helps provide robustness against overfitting and gives additional strength to the statistical analysis. We use classification accuracy as the evaluation metric. Because we perform 10-fold cross validation for each experiment, our results for each experiment are a micro-average of the classification accuracy.

The objective of these experiments is not only to determine the usefulness of n-gram Web page representations, but also to investigate the effect of three feature selection measures. We also examine the effect of the length of n-gram used, and the effect of the Web page profile size, and determine with which combination(s) of these parameters our Web genre classification model achieves the best accuracy. Each experiment uses fixed-length byte n-grams, however we vary the length of the n-gram from 2 to 10. For each n-gram length, the number of n-grams used to create the Web page profiles is varied from 5 to 500, in increments of 5 from 5 to 50, and in increments of 25 from 50 to 500.

Because previous research on the use of n-gram representations of Web pages and Web page genres indicates that when using small Web page profile sizes, it is beneficial to preprocess the Web pages by removing the HTML tags and Javascript code [9], each Web page in each data set is preprocessed to remove all HTML tags and JavaScript code. The remaining textual content of each Web page is then used to form a byte n-gram representation of the Web page. No stemming of terms or stopword removal is performed.

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

4 Results and Discussion

The results of these experiments are reported in Tables 1–8. Analysis of variance (ANOVA) tests were performed to determine the effect on classification accuracy of the feature selection measure, the n-gram length, the Web Page profile size, and the combination of these parameters. A summary of the results follows.

4.1 Overall Results

As shown in Table 1, each of the feature selection measures allows the Web page genre classification model to achieve a high mean classification accuracy for n-grams of length 2 to 10 and profile sizes of 5 to 500. This gives very strong support for the use of n-gram representations of Web pages and Web page genres. Using the Chi-square statistic as a feature selection measure gives the best performance on each of the data sets, and an ANOVA test shows that this performance is significantly better than using either frequency or Information Gain as feature selection measures ($p < 0.001$). For this reason, the remainder of our discussion of the results focuses on the use of the Chi-square statistic as the feature selection measure.

Table 1. Mean accuracy for feature section measures averaged over n-gram lengths of 2 to 10 and Web page profile sizes of 5 to 500. Standard error in parenthesis.

Feature Selection Measure	7-Genre	KI-04
Frequency	0.828 (< 0.001)	0.900 (0.001)
Information Gain	0.887 (< 0.001)	0.918 (0.001)
Chi-square Statistic	0.955 (< 0.001)	0.969 (0.001)

It is interesting to note that the feature selection measure that results in the best performance for the classifier also results in the use of the smallest centroid genre profiles. Tables 4 and 5 compare the sizes of the smallest and largest centroids for each measure, for each data set. For each feature selection measure, the size of the centroid increases for each n-gram length as the size of the Web page profile increases from 5 to 500, thus the smallest centroids at each n-gram length are for Web page profiles of size 5, whereas the largest centroids at each n-gram length are for Web page profiles of size 500. The tables show that using frequency as the feature selection measure results in the use of the largest centroid genre profiles, whereas the use of the Chi-square statistic results in the use of the smallest centroid genre profiles of the three feature selection measures.

4.2 Effect of n-gram Length

In these experiments, the n-gram length ranges from 2 to 10 in increments of 1. As shown in Table 2, there is a significant impact on the classification accuracy depending on which n-gram length is used; this effect on the accuracy for each data set is significant at $p < 0.001$. The partial Eta squared for the n-gram length was 0.334 for the 7-Genre data set and 0.077 for the KI-04 data set. These results indicate that the proportion of total variability in the classification accuracy for the 7-Genre data set is moderately influenced by the n-gram length. Although the partial Eta squared is lower for the KI-04 data set, as the accuracy rate approaches 100%, which it does with this data set, we could expect a ceiling effect for any variance overlap. Table 2 compares the mean accuracy, averaged over profile sizes of 5 to 500, for each data set when the Chi-square statistic is used as a feature selection measure. The high mean classification accuracy for both data sets is an indication not only of the soundness of the Chi-square statistic as a feature selection measure, but also of the robustness of the classification model and the effectiveness of the use of n-gram profile representations of the Web pages.

Table 2. Mean accuracy for n-grams of length 2 to 10 for the Chi-square statistic as the feature selection measure, averaged over all Web page profile sizes. Standard error in parenthesis.

n-gram Length	7-Genre	KI-04
2	0.961 (0.001)	0.976 (0.002)
3	0.935 (0.001)	0.950 (0.002)
4	0.954 (0.001)	0.958 (0.002)
5	0.966 (0.001)	0.967 (0.002)
6	0.969 (0.001)	0.971 (0.002)
7	0.962 (0.001)	0.974 (0.002)
8	0.956 (0.001)	0.973 (0.002)
9	0.948 (0.001)	0.975 (0.002)
10	0.943 (0.001)	0.977 (0.002)

4.3 Effect of Web Page Profile Size

In these experiments, the Web page profile size ranges from 5 to 50 in increments of 5, and from 50 to 500 in increments of 25. The effect of profile size on the classification accuracy for each data set is significant at $p < 0.001$. The partial Eta squared for the Web page profile size was 0.687 for the 7-Genre data set and 0.197 for the KI-04 data set. These results indicate that the proportion of total variability in the classification accuracy for the 7-Genre data set is quite

highly influenced by the profile size, whereas that of the KI-04 data set is less so. As with the effect of n-gram length, the partial Eta squared is lower for the KI-04 data set, however this could be due to a ceiling effect, as the classification accuracy rate reaches 100%. Table 8 compares the mean accuracy, averaged over n-gram lengths of 2 to 10, for each data set when the Chi-square statistic is used as a feature selection measure. We note again that the high mean classification accuracy for both data sets indicates the effectiveness of the n-gram Web page representations, as well as the strength of the Chi-square statistic as a feature selection measure and the robustness of the classification model.

4.4 Best Results

The best classification accuracy achieved by our model is 99.1% for the 7-Genre data set, and 100% for the KI-04 data set. These results are a strong indication of the effectiveness of this model. For the 7-Genre data set, the best accuracy was obtained with four combinations of n-gram length and Web page profile size. In each of these cases, the n-gram length was 2; the profile sizes were 10, 20, 25, and 40. For the KI-04 data set, the best accuracy was achieved with three different combinations of n-gram length and Web page profile size. As with the 7-Genre data set, the n-gram length was 2 in each case; the profile sizes were 15, 20, and 500. Tables 6 and 7 give the best and worst combinations of n-gram length and Web page profile size for each data set.

Table 3 gives a comparison of these results with the results of other researchers on the same data sets, as published in the literature. On the 7-Genre data set, Santini [14], 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 [6] 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%. Using the KI-04 data set, Santini [14] achieves an accuracy of 68.9% while Meyer zu Eissen and Stein [11] 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 [1] 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 [6] report a best accuracy of 84.1%, whereas Jebari [5] achieves an accuracy of 96.0% using a centroid-based classifier.

5 Summary and Conclusions

The research reported in this paper is part of a project on the automatic classification of Web pages by their genres, using n-gram representations of the Web pages and Web page genres, and a distance function classification model. The goals of this research were to determine the usefulness of n-gram Web page representations, and to investigate the effect of three feature selection measures. We also examined the effect of the length of the n-gram used and the effect of

Table 3. A comparison of the best accuracy results for several researchers.

Researchers	7-Genre	KI-04
Santini [14]	0.906	0.689
Meyer zu Eissen and Stein [11]		0.700
Boese and Howe [1]		0.748
Kanaris and Stamatatos [6]	0.965	0.841
Jebari [5]		0.960
Our results	0.991	1.00

the Web page profile size, and determined with which combination(s) of these parameters our Web genre classification model achieves the best accuracy. Although other authors have looked at these factors individually, this study has looked at them in combination.

This examination of the effect on the automatic classification of Web pages by genre has shown that very high classification accuracy can be achieved using n-gram representations of Web pages. We also determined that the type of feature selection method, the n-gram length, and the Web page profile size have significant effects on the performance of the classification model, measured in terms of mean classification accuracy. The Chi-square statistic was found to be the most effective feature selection method, achieving very high accuracy with Web page profiles as small as 5. Although high classification accuracy was achieved with each length of n-gram, from 2 to 10, the best accuracy on both the 7-Genre and KI-04 data sets was found using n-grams of length 2, indicating that this n-gram length could be a good first choice for use with other data sets. The best accuracy for the 7-Genre data set was with profile sizes of 10, 20, 25, and 40; for the KI-04 data set, the best accuracy was with profile sizes of 15, 20, and 500. These results indicate that small profile sizes of between 10 and 40 are excellent choices for use with this model, particularly when coupled with the use of n-grams of length 2.

The results of this work indicate that n-gram Web page representations are very effective for the task of classifying Web pages by genre, and that the use of the Chi-square statistic aids identifying small feature sets of n-grams to represent Web pages and Web page genres in this distance function classification model. Although all three feature selection measures (frequency, the Chi-square statistic, and Information Gain) performed well, the Chi-square statistic significantly outperformed the other measures on each data set.

This research is continuing and experimentation is ongoing. We recognize the need to expand this work to a larger scale in which we work with more genres, noise, and varying levels of genre granularity, as well as with multi-labeled and unbalanced data sets. A comparison of the distance function classification model to popular machine learning methods such as SVM and k-nearest neighbours is also planned.

Table 4. The range of centroid genre profile sizes for Web page profile sizes of 5 to 500 on the 7-Genre data set for n-grams of length 2 to 10.

7-Genre	Frequency		Information Gain		Chi-square Statistic	
n-gram Length	Smallest Centroid (Web Page Profile Size 5)	Largest Centroid (Web Page Profile Size 500)	Smallest Centroid (Web Page Profile Size 5)	Largest Centroid (Web Page Profile Size 500)	Smallest Centroid (Web Page Profile Size 5)	Largest Centroid (Web Page Profile Size 500)
2	25	2113	272	2541	16	1411
3	51	4644	138	11953	15	1614
4	46	12231	44	28953	12	2667
5	111	20564	25	31419	11	3762
6	167	27360	32	20421	13	3861
7	160	33494	35	13327	15	3478
8	165	36753	38	10426	16	3195
9	173	39290	41	9048	17	2980
10	181	41138	42	8455	25	2875

Table 5. The range of centroid genre profile sizes for Web page profile sizes of 5 to 500 on the KI-04 data set for n-grams of length 2 to 10.

KI-04	Frequency		Information Gain		Chi-square Statistic	
n-gram Length	Smallest Centroid (Web Page Profile Size 5)	Largest Centroid (Web Page Profile Size 500)	Smallest Centroid (Web Page Profile Size 5)	Largest Centroid (Web Page Profile Size 500)	Smallest Centroid (Web Page Profile Size 5)	Largest Centroid (Web Page Profile Size 500)
2	53	2688	430	2892	29	2626
3	102	6941	262	15815	20	2894
4	148	14219	192	34468	19	4143
5	259	21065	188	37273	19	5832
6	328	27158	178	33414	20	7530
7	346	33296	176	31186	24	9821
8	358	38366	188	30007	31	12866
9	375	42458	194	29432	50	16154
10	379	44075	195	29021	86	19923

Table 6. Best combinations of n-gram length and Web page profile size, based on mean accuracy, for the 7-Genre and KI-04 data sets. Standard error in parenthesis.

Data Set	Best Accuracy	n-gram Length	Web Page Profile Size
7-Genre	0.991 (0.005)	2	10, 20, 25, 40
		3	20, 25
		4	15, 20
		5	25
KI-04	1.00 (0.010)	2	15, 20, 500

Table 7. Worst combinations of n-gram length and Web page profile size, based on mean accuracy, for the 7-Genre and KI-04 data sets. Standard error in parenthesis.

Data Set	Worst Accuracy	n-gram Length	Web Page Profile Size
7-Genre	0.841 (0.005)	3	500
KI-04	0.872 (0.010)	3	500

Table 8. Mean accuracy using Chi-square statistic as the feature selection measure on the 7-Genre and KI-04 data sets, for Web page profile sizes of 5 to 500. Standard error in parenthesis.

Web Page Profile Size	7-Genre	KI-04	Web Page Profile Size	7-Genre	KI-04
5	0.932 (0.002)	0.949 (0.003)	175	0.960 (0.002)	0.972 (0.003)
10	0.967 (0.002)	0.975 (0.003)	200	0.955 (0.002)	0.967 (0.003)
15	0.977 (0.002)	0.983 (0.003)	225	0.951 (0.002)	0.965 (0.003)
20	0.980 (0.002)	0.985 (0.003)	250	0.948 (0.002)	0.961 (0.003)
25	0.980 (0.002)	0.987 (0.003)	275	0.945 (0.002)	0.960 (0.003)
30	0.980 (0.002)	0.987 (0.003)	300	0.940 (0.002)	0.959 (0.003)
35	0.980 (0.002)	0.987 (0.003)	325	0.938 (0.002)	0.957 (0.003)
40	0.981 (0.002)	0.988 (0.003)	350	0.933 (0.002)	0.957 (0.003)
45	0.981 (0.002)	0.988 (0.003)	375	0.929 (0.002)	0.955 (0.003)
50	0.980 (0.002)	0.988 (0.003)	400	0.925 (0.002)	0.954 (0.003)
75	0.978 (0.002)	0.985 (0.003)	425	0.925 (0.002)	0.951 (0.003)
100	0.974 (0.002)	0.983 (0.003)	450	0.925 (0.002)	0.950 (0.003)
125	0.968 (0.002)	0.978 (0.003)	475	0.923 (0.002)	0.948 (0.003)
150	0.965 (0.002)	0.975 (0.003)	500	0.920 (0.002)	0.945 (0.003)

References

1. E. Boese and A. Howe. Effects of Web Document Evolution on Genre Classification. In *Proc. of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, pages 632–639, New York, NY, USA, 2005. ACM Press.
2. W. Cavnar and J. Trenkle. N-gram-based text categorization. *Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.
3. L. Dong, C. Watters, J. Duffy, and M. Shepherd. Binary cybergenre classification using theoretic feature measures. *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006)*, pages 313–316, 2006.
4. J. Houvardas and E. Stamatatos. N-gram Feature Selection for Authorship Identification. *Proc. of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, pages 77–86, 2006.
5. C. Jebari. Refined and incremental centroid-based approach for genre categorization of web pages. In *Proc. of the 17th International World Wide Web Conference (WWW2008), NLPIX Workshop*, 2008.
6. I. Kanaris and E. Stamatatos. Webpage Genre Identification Using Variable-Length Character n-Grams. *Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, 2:3–10, 2007.
7. V. Kešelj, F. Peng, N. Cercone, and T. Thomas. N-gram-based author profiles for authorship attribution. In *Proc. of the Conference of the Pacific Association for Computational Linguistics (PACLING'03)*, pages 255–264, 2003.
8. Y. Lee and S. Myaeng. Text Genre Classification with Genre-revealing and Subject-revealing Features. In *Proc. of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*, pages 145–150, New York, NY, USA, 2002. ACM Press.
9. J. Mason, M. Shepherd, and J. Duffy. Classifying web pages by genre: A distance function approach. *Proc. of the 5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*, 2009.
10. J. Mason, M. Shepherd, and J. Duffy. An n-gram based approach to automatically identifying web page genre. *Proc. of the 41st Annual Hawaii International Conference on System Sciences (HICSS-42)*, 2009.
11. S. Meyer zu Eissen and B. Stein. Genre classification of web pages. *Proc. of the 27th German Conference on Artificial Intelligence (KI-2004)*, 2004.
12. G. Rehm. Towards Automatic Web Genre Identification. *Proc. of the 37th Hawaii International Conference on System Sciences (HICSS-37)*, 04, 2002.
13. M. Robnik-Šikonja and I. Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53(1):23–69, 2003.
14. M. Santini. *Automatic identification of genre in web pages.* PhD thesis, University of Brighton, U.K., 2007.
15. C. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379 – 423 and 623 – 656, July and October 1948.
16. J. Silva and G. Lopes. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. *Proc. of the 6th Meeting on the Mathematics of Language*, 369, 1999.
17. B. Stein and S. Meyer zu Eissen. Retrieval models for genre classification. *Scandinavian Journal of Information Systems*, 20(1):93–119, 2008.
18. Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the 14th International Conference on Machine Learning (ICML '97)*, pages 412–420, 1997.