CLASSIFYING WEB PAGES BY GENRE
Dealing with Unbalanced Distributions, Multiple Labels and Noise

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

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

Abstract: Web page genre classification is a potentially powerful tool for filtering the results of online searches. The goal of this research is to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, and in the presence of noise, in order to better represent a real world environment. The approach is based on *n*-gram representations of the Web pages and centroid representations of the genre classes. Experimental results compare very favorably with those of other researchers.

1 INTRODUCTION

Recent research has shown genre to be a potentially powerful tool for filtering the results of online searches. Although most information retrieval searches are topic-based, users are typically looking for a specific type of information with regard to a particular query, and genre can provide a complementary dimension along which to categorize documents. The relevance of a particular document to a given query depends on the information need of the user who issues the query, and information retrieval systems could be enhanced by providing users with the ability to filter documents according to their genre (Finn and Kushmerick, 2006).

The research reported in this paper explores the classification of Web pages by genre using *n*-gram representations of the Web pages. Experiments are run on a multi-label corpus using both a centroid classification model and an SVM classifier, and on a highly unbalanced single-label corpus. In order to investigate the research question of whether the centroid classification model is effective when noise Web pages are included in the corpus, this study compares the classification performance of the centroid model with and without the addition of noise Web pages.

2 RELATED WORK

Most research on the classification of Web pages by genre has focused on labeling each Web page as belonging to a single genre, however the difficulty of assigning a single genre label to a Web page has been acknowledged by researchers who have conducted surveys and user studies (Crowston and Williams, 1997; Meyer zu Eissen and Stein, 2004; Santini, 2008). (Rosso, 2008) explored the use of genre to improve the effectiveness of Web searching, and found that the two factors which seemed to hamper participant agreement on Web page genres were that some of the Web pages seemed to fit into multiple genres, and that some of the genres seemed to have fuzzy boundaries. Despite these issues, 90% of the Web pages were classified into a single genre by the majority of the participants, leading Rosso to conclude that although a multi-genre classification scheme would be superior, a single genre classification scheme could still offer improvement in Web searching.

The representations of Web pages used in the genre classification task tend to be based on those used in text classification, often augmented with information such as HTML tags and URL information. For example (Meyer zu Eissen and Stein, 2004) combine genre-specific vocabulary and closed-class word sets with text statistics, part-of-speech information, and HTML tags, whereas (Jebari, 2008) combines two centroid-based classifiers, one of which uses structural information from the document, while the other uses URL information. (Kanaris and Stamatatos, 2009) use feature sets of variable-length character *n*-grams and information about the most frequent HTML tags to perform classification using a support vector machine. (Stein and Meyer zu Eissen,

E. Mason J., Shepherd M., Duffy J. and Kešelj V..
CLASSIFYING WEB PAGES BY GENRE - Dealing with Unbalanced Distributions, Multiple Labels and Noise.
DOI: 10.5220/0003435058905594
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 589-594
ISBN: 978-989-8425-51-5

2008) give a detailed overview of the document representations used for Web genre classification.

A shortcoming of the existing Web page classification research is that it tends to be carried out on corpora which do not contain any noise. (Shepherd et al., 2004), however, introduced noise pages in their classification of homepages, and found that the performance of the classifier deteriorated. (Levering et al., 2008) added 798 noise Web pages to a three genre corpus containing 501 single-label Web pages. They found that textual features alone performed very poorly in the presence of noise, but that the addition of HTML features dramatically improved performance.

3 METHODOLOGY

3.1 Classification Models

This study uses *n*-gram profile representations of Web pages in the automatic identification of the genre of the Web pages. Web pages are represented by fixed-length byte *n*-grams. Initially, Web page profiles containing the *n*-grams and their associated normalized frequencies are produced using the Perl package Text:Ngrams[1]. The byte *n*-grams are raw character *n*-grams in which no bytes are ignored, including the whitespace characters, thus some of the structure of a document is captured by using byte *n*-grams. Based on research by (Mason et al., 2009c), the Chi-square statistic is then used as a feature selection measure; the *n*-grams are ranked according to the Chi-square statistics, and profiles are constructed for the Web pages. Each profile contains the L top ranked *n*-grams in the Web page, and the corresponding frequency for each of these *n*-grams.

The centroid classification model has been shown to be an effective model for Web page genre classification (Mason et al., 2009a; Mason et al., 2009b). When using the centroid classification model, each Web page genre is represented by a profile that is constructed by combining the *n*-gram profiles for each Web page of that genre from the training set, forming a centroid profile for each Web page genre. These centroid profiles will contain a varying number of *n*-grams, therefore each of the centroid profiles is truncated to the size of the smallest centroid profile. Each Web page profile from the test set is compared with each genre centroid profile from the training set. The distance between two *n*-gram profiles is computed using the formula suggested by (Kešelj et al., 2003) in their paper on the use of *n*-gram profiles for authorship attribution. The distance (dissimilarity) between two *n*-gram profiles is defined as

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

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

In order to classify a Web page as belonging to more than one genre, or as not belonging to any known genre, the centroid classification model includes thresholds that are computed for each genre. If the distance between the Web page profile and a genre profile is less than or equal to the threshold, the Web page is labeled as belonging to that genre. If the distance is greater than the threshold, the Web page is deemed not to belong to that genre. Thus, Web pages that are labeled as noise Web pages exceed the threshold for every genre; Web pages that are labeled as belonging to more than one genre fall within the thresholds of multiple genres.

The method for setting each genre threshold is to first order all of the Web pages in the training set according to their distance from a particular genre profile, in ascending order. This ordered list of Web pages from the training set can then be stepped through one Web page at a time, such that at each step, the current Web page is labeled as belonging to the genre in question, and the accuracy of the classification thus far is computed. In this manner, the optimal threshold for each genre, based on the training data, can be determined. This process is then repeated for each genre in the corpus. This method of setting the genre thresholds is known as the *optimal threshold method* (Mason et al., 2010), because the method gives a set of fixed thresholds that give the optimal classification accuracy on the training set.

An alternative method of assigning more than one label to a Web page is to use the support vector machine (SVM) classification model, rather than the centroid classification model. For the experiments conducted for this paper, multiple binary SVM classifiers are trained individually and the outputs of the classifiers are combined for classification of multiple genres; thus, for a classification problem with twenty genres, twenty SVM classifiers are trained using the conventional *one-against-all* approach.

3.2 Corpora

The 20-Genre corpus was constructed by Mitja Luštrek and Andrej Bratko at the Jožef Stefan Institute, and is available online[2]. Of the 1539 Web pages

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

[2] http://dis.ijs.si/mitjal/genre/

in this collection, 1059 have one genre label, 438 have two genre labels, 39 have three labels, and 3 have four labels. The Syracuse corpus was assembled by a team of researchers led by Barbara Kwasnik and Kevin Crowston at Syracuse University. The collection contains a total of 2748 labeled Web pages, each of which has one, and only one, genre label. There are 118 different genres represented, with the number of Web pages in each genre ranging from 1 to 350. Of these, only 24 of the genres contain 30 or more Web pages each. These experiments use a subset of the Syracuse corpus consisting of the 24 genres that contain at least 30 Web pages; the remainder of the labeled Web pages from the corpus are reserved for use as noise Web pages. Within the 24 genre subset there are 1985 Web pages. Figure 1 shows the Zipfian-like log-log scale plot of these genre densities in which there are a few genres with many Web pages, and many genres with very few Web pages.

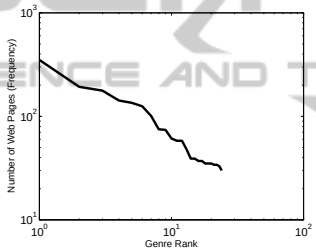

Figure 1: Log-log scale plot of the genre distributions for the Syracuse corpus.

3.3 Experiments

These experiments test the centroid classification model on the unbalanced, multi-label 20-Genre corpus, and on the highly unbalanced Syracuse corpus with and without the addition of 750 noise Web pages. For the purpose of this research, a noise Web page is a Web page that does not belong to any genre in the corpus in question. Because the Web pages in the Syracuse corpus have only one genre label, the centroid classification model is modified such that each Web page is assigned at most one label. The optimal threshold method is used to determine thresholds for each genre, and a Web page is assigned the label of the genre to which it is most similar, within these thresholds. If the Web page is not within the threshold for any genre, it is labeled as a noise Web page.

Based on previous research (Mason et al., 2009c), the *n*-gram length in these experiments is varied from 2 to 4, and for each *n*-gram length the Web page profile size is varied from 15 to 50. The Web pages are not preprocessed. The metrics used to evaluate the classification performance in these experiments are macro-precision and macro-recall.

4 RESULTS AND DISCUSSION

The results of the experiments in this study are reported in Tables 1 to 6. A summary of the experiments and results follows.

4.1 Effect of *n*-gram Length

Table 1 gives the mean classification results on the 20-Genre corpus for the centroid classification method and the SVM method, while Table 2 gives the results for the centroid classification method on the Syracuse corpus, with and without noise Web pages. The results are averaged over Web page profile sizes of 15 to 50, for each *n*-gram length from 2 to 4.

Table 1 shows that as the *n*-gram length is increased, the precision, recall and F1-measure values decrease, with the exception of the precision values for the SVM method. The results of Scheffé post hoc tests show that for each method overall, *n*-grams of length 2 are the best choice. The effect of the *n*-gram length on the precision, recall, and F1-measure for each method on the 20-Genre corpus is statistically significant ($p < 0.01$), however the partial η^2 is less than 0.10 in each case. This indicates that less than 10% of the total variability in the precision, recall, and F1-measure for each method is accounted for by the *n*-gram length. Based on precision, there is no statistically significant difference between the centroid classifier and the SVM method. In terms of recall and the F1-measure, the centroid classification method is significantly better than the SVM method ($p < 0.001$).

Table 2 indicates that the addition of the noise Web pages to the Syracuse corpus decreases the precision of the classification, but increases the recall; this means that the noise Web pages are less likely to be mislabeled as belonging to another genre than are the non-noise Web pages. In each case, the results of Scheffé post hoc multiple comparison tests confirm that, as with the 20-Genre corpus, an *n*-gram length of 2 gives the best results. We conclude that using an *n*-gram length of 2 is the best choice in terms of precision, recall, and the F1-measure. The use of a small *n*-gram size may also have the advantage of being less computationally intensive. The effect of the *n*-gram length on the precision, recall, and F1-measure for the classification performance of the centroid model on the Syracuse corpus, both with and without noise Web pages, is statistically significant ($p < 0.01$). For the precision, the partial η^2 in each case is less than 0.10, indicating that the proportion of total variability in the precision is only slightly influenced by the *n*-gram length. For the recall and F1-measure, however, the partial η^2 in each case is at least 0.38. This indicates that at least 38% of the total variability in the re-

Table 1: 20-Genre Corpus: results averaged over Web page profile sizes 15 to 50. Standard error ≤ 0.005.

20-Genre Corpus	Centroid Classification Method			Support Vector Machine		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.992	0.768	0.855	0.989	0.720	0.808
3	0.990	0.764	0.851	0.998	0.702	0.798
4	0.989	0.731	0.823	0.998	0.662	0.768
Average	0.990	0.754	0.843	0.995	0.695	0.791

Table 2: Syracuse Corpus: centroid classifier results averaged over Web page profile sizes 15 to 50. Standard error ≤ 0.005.

Syracuse Corpus	Without Noise Web Pages			With Noise Web Pages		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.998	0.947	0.970	0.990	0.949	0.966
3	0.993	0.821	0.876	0.980	0.828	0.875
4	0.963	0.737	0.810	0.954	0.747	0.812
Average	0.985	0.835	0.885	0.975	0.841	0.884

Table 3: 20-Genre Corpus: results averaged over *n*-gram lengths 2 to 4. Standard error ≤ 0.010.

20-Genre Corpus	Centroid Classification Method			Support Vector Machine		
Profile Size	Precision	Recall	F1	Precision	Recall	F1
15	0.987	0.725	0.816	0.995	0.691	0.789
20	0.989	0.725	0.816	0.996	0.695	0.791
25	0.991	0.764	0.852	0.995	0.694	0.790
30	0.991	0.765	0.852	0.996	0.696	0.793
35	0.994	0.767	0.855	0.994	0.696	0.792
40	0.991	0.765	0.852	0.994	0.697	0.793
45	0.991	0.764	0.851	0.996	0.696	0.792
50	0.988	0.761	0.848	0.994	0.696	0.792
Average	0.990	0.754	0.843	0.995	0.695	0.791

Table 4: Syracuse Corpus: centroid classifier results averaged over *n*-gram lengths 2 to 4. Standard error ≤ 0.009.

Syracuse Corpus	Without Noise Web Pages			With Noise Web Pages		
Profile Size	Precision	Recall	F1	Precision	Recall	F1
15	0.979	0.805	0.861	0.966	0.813	0.859
20	0.978	0.818	0.871	0.968	0.825	0.870
25	0.985	0.834	0.885	0.974	0.840	0.884
30	0.987	0.843	0.893	0.977	0.849	0.892
35	0.989	0.847	0.896	0.980	0.853	0.896
40	0.986	0.845	0.891	0.976	0.851	0.891
45	0.983	0.843	0.891	0.974	0.849	0.890
50	0.992	0.845	0.894	0.981	0.851	0.892
Average	0.985	0.835	0.885	0.975	0.841	0.884

call and F1-measure are accounted for by the *n*-gram length, thus *n*-gram length has a more pronounced effect on these measures than it has on the precision.

4.2 Effect of Web Page Profile Size

Table 3 gives the mean classification results for the centroid classification model and the SVM model on the 20-Genre corpus, while Table 4 gives the results for the centroid classification model on the Syracuse corpus. The results are averaged over *n*-gram lengths from 2 to 4. The number of *n*-grams used to represent each Web page ranges from 15 to 50 in increments of 5, and Tables 3 and 4 show that the classification performance of the centroid model is very stable over these Web page profile sizes. The effect of the Web page profile size was not statistically significant on the precision for either corpus. The effect of the Web page profile size was statistically significant on the recall and F1-measure for the centroid classification method ($p < 0.01$), however the partial η^2 was less than or equal to 0.05, indicating that the Web page profile size accounted for at most 5% of the overall variance of the recall and F1-measure.

4.3 Effect of Genre

Table 5 gives a comparison of the mean precision and recall for each genre in the 20-Genre corpus, using the centroid classifier. Table 5 also gives the best results of (Vidulin et al., 2007) and (Kanaris and Stamatatos, 2009) using the same corpus. The results show that

Table 5: Mean classification results by genre on the 20-Genre corpus. The results for the centroid classifier are averaged over Web page profile sizes of 15 to 50 and *n*-gram lengths of 2 to 4, with a standard error ≤ 0.016.

20-Genre Corpus		Vidulin et al., 2007a		Kanaris and Stamatatos, 2009		Centroid Classifier	
Genre	Size	Precision	Recall	Precision	Recall	Precision	Recall
OFFICIAL	55	0.73	0.27	0.78	0.45	0.988	0.924
SHOPPING	66	0.72	0.33	0.98	1.00	1.000	0.590
PROSE FICTION	67	0.69	0.30	0.98	1.00	0.966	0.751
ADULT	68	0.78	0.71	0.82	0.46	1.000	0.671
FAQ	70	0.98	0.73	0.96	0.60	0.998	0.971
POETRY	72	0.76	0.61	0.98	1.00	0.996	0.910
ENTERTAINMENT	76	0.69	0.27	0.25	0.06	0.969	0.653
SCIENTIFIC	76	0.85	0.51	0.98	1.00	0.989	0.955
BLOG	77	0.83	0.56	0.88	0.41	0.998	0.762
GATEWAY	77	0.45	0.12	0.49	0.22	0.986	0.587
ERROR MESSAGE	79	0.87	0.68	0.67	0.56	0.993	0.981
COMMUNITY	82	0.76	0.55	0.28	0.11	0.979	0.792
USER INPUT	84	0.83	0.57	0.97	0.99	0.995	0.756
CHILDREN'S	105	0.81	0.48	0.74	0.47	0.996	0.860
PERSONAL	113	0.72	0.16	0.97	1.00	0.999	0.572
COMM./PROMO.	121	0.40	0.04	0.28	0.11	0.978	0.787
CONTENT DELIVERY	138	0.64	0.23	0.43	0.28	0.989	0.450
JOURNALISTIC	186	0.62	0.36	0.79	0.52	0.994	0.903
INFORMATIVE	225	0.30	0.09	0.58	0.18	1.000	0.699
INDEX	227	0.63	0.37	0.36	0.21	0.991	0.514
Average	103	0.70	0.40	0.74	0.55	0.990	0.754

Table 6: Mean classification results by genre for the centroid classifier on the Syracuse corpus. The results are averaged over Web page profile sizes of 15 to 50 and *n*-gram lengths of 2 to 4. Standard error ≤ 0.015.

Syracuse Corpus		Without Noise Web Pages			With Noise Web Pages		
Genre	Size	Precision	Recall	F1	Precision	Recall	F1
INDEX TO MISC. RESOURCES	30	0.786	0.460	0.506	0.767	0.460	0.505
TABLE OF CONTENTS	33	0.993	0.828	0.888	0.993	0.828	0.888
BIBLIOGRAPHIC RECORD	34	1.000	0.833	0.901	0.965	0.833	0.885
BIOGRAPHY	34	1.000	0.707	0.814	1.000	0.707	0.814
ENCYCLOPEDIA ENTRY	35	0.999	0.803	0.887	0.979	0.803	0.878
TUTORIAL AND HOW-TO	35	0.930	0.677	0.762	0.930	0.677	0.762
DEFINITION/DESCRIPTION	35	1.000	0.577	0.698	0.995	0.577	0.697
OTHER BIOGRAPHY	37	1.000	0.873	0.927	1.000	0.873	0.927
LESSON PLAN	37	0.997	0.562	0.654	0.997	0.562	0.654
PRESS RELEASE	39	1.000	0.861	0.921	0.999	0.861	0.920
DIRECTORY OF COMPANIES	39	1.000	0.844	0.896	1.000	0.844	0.896
ABOUT A PROGRAM	48	0.983	0.741	0.823	0.980	0.741	0.822
ABOUT AN ORGANIZATION	58	0.999	0.924	0.957	0.999	0.924	0.957
FACTS-AND-FIGURES PAGE	58	0.998	0.868	0.922	0.998	0.868	0.922
DISCUSSION FORUM	61	0.997	0.917	0.953	0.996	0.917	0.952
COMPANY/ORG. HOMEPAGE	74	0.983	0.861	0.913	0.983	0.861	0.912
ADVERTISING	75	0.998	0.898	0.940	0.994	0.898	0.938
MAGAZINE ARTICLE	101	1.000	0.941	0.969	1.000	0.941	0.969
RECIPE	125	1.000	0.973	0.986	1.000	0.973	0.986
BLOG	135	0.999	0.981	0.990	0.996	0.981	0.988
DIRECTORY RESOURCES/LINKS	142	0.981	0.975	0.977	0.981	0.975	0.977
OTHER ARTICLE	177	1.000	0.971	0.985	1.000	0.971	0.985
NEWS STORY	193	0.996	0.975	0.985	0.996	0.975	0.985
PRODUCT/SERVICE PAGE	350	0.998	0.987	0.992	0.998	0.987	0.992
NOISE	750	n/a	n/a	n/a	0.821	0.996	0.898
Average	109	0.985	0.835	0.885	0.975	0.841	0.884

each method has a much higher precision than recall, averaged over all 20 genres. This means that the genre labels assigned by the classifiers are quite accurate, but that these machine learning classifiers are not assigning as many labels as did the human annotators when the corpus was constructed. Table 6 gives a comparison by genre of the mean precision, recall, and F1-measure of the centroid classifier on the Syracuse corpus, with and without noise Web pages. We are not aware of other published results for the corpus that could be used for comparison purposes.

In Tables 5 and Table 6 the results are averaged over *n*-gram lengths from 2 to 4 and Web page profile sizes of 15 to 50. The results indicate, not surprisingly, that some genres are easier to classify than others. ANOVA on the results of the centroid classi-

fication model indicates that genre is the leading factor (over *n*-gram length and Web page profile size) in predicting the outcome of the classification. Although genre is an influential factor in predicting the classification performance, a specific hypothesis about which genres can be better classified than others has not been developed. The variability between genres is likely to be caused by a factor that has not been explored as part of the current research, such as the length of the Web pages, or the homogeneity of each genre.

4.4 The Effect of Noise

As shown in Tables 2, 4, and 6, the addition of 750 noise Web pages to the Syracuse corpus resulted in a slight decrease in the precision of the centroid classifier, and a slight increase in the recall. Thus, the noise Web pages are less likely to be mislabeled as belonging to another genre than are the non-noise Web pages. Of the 1985 non-noise Web pages, an average of 170 pages (8.6%) were erroneously labeled as noise Web pages by the centroid model; of the 750 noise Web pages, an average of 3 pages (0.4%) were erroneously labeled as non-noise pages. The number of Web pages erroneously labeled as noise increases from 3.6% to 13.3% as the *n*-gram length was increased from 2 to 4, whereas the number of noise pages erroneously given genre labels decreases from 0.85% to 0.03% as the *n*-gram length was increased from 2 to 4. This suggests that the proportion of noise Web pages expected to appear in a corpus could influence the choice of the *n*-gram length to be used.

5 CONCLUSIONS

The major contribution of this study is to show that byte *n*-gram Web page representations can be used effectively, with more than one classification model, to classify Web pages by genre, even when the Web pages belong to more than one genre or to no known genre, and when the number of Web pages in each genre is quite variable. The results of these experiments also showed that in general, as the length of the *n*-grams used to represent the Web pages was increased, the classification performance for each model decreased. The results also indicated that over the range of 15 to 50, the number of *n*-grams used to represent each Web page has only a slight impact on the classification results.

REFERENCES

Crowston, K. and Williams, M. (1997). Reproduced and Emergent Genres of Communication on the World-Wide Web. In *Proc. 30th Hawaii Intl. Conf. on System Sciences*, pages 30–39.

Finn, A. and Kushmerick, N. (2006). Learning to Classify Documents According to Genre. *Journal of American Society for Information Science and Technology*, 57(11):1506–1518.

Jebari, C. (2008). Refined and Incremental Centroid-based Approach for Genre Categorization of Web Pages. In *Proc. 17th Intl. World Wide Web Conf.*

Kanaris, I. and Stamatatos, E. (2009). Learning to Recognize Webpage Genres. *Information Processing & Management*, 45(5):499–512.

Kešelj, V., Peng, F., Cercone, N., and Thomas, T. (2003). N-gram-based author profiles for authorship attribution. In *Proc. Conf. Pacific Association for Computational Linguistics*, pages 255–264.

Levering, R., Cutler, M., and Yu, L. (2008). Using Visual Features for Fine-Grained Genre Classification of Web Pages. In *Proc. 41st Hawaii Intl. Conf. on System Sciences*. IEEE Computer Society.

Mason, J., Shepherd, M., and Duffy, J. (2009a). An N-gram Based Approach to Automatically Identifying Web Page Genre. In *Proc. 42nd Hawaii Intl. Conf. on System Sciences*.

Mason, J., Shepherd, M., and Duffy, J. (2009b). Classifying Web Pages by Genre: A Distance Function Approach. In *Proc. 5th Intl. Conf. on Web Information Systems and Technologies*.

Mason, J., Shepherd, M., and Duffy, J. (2009c). Classifying Web Pages by Genre: An *n*-gram Based Approach. In *Proc. Intl. Conf. on Web Intelligence*.

Mason, J., Shepherd, M., Duffy, J., Kešelj, V., and Watters, C. (2010). An *n*-gram Based Approach to Multi-labeled Web Page Genre Classification. In *Proc. 43rd Hawaii Intl. Conf. on System Sciences*.

Meyer zu Eissen, S. and Stein, B. (2004). Genre Classification of Web Pages. In *Proc. 27th German Conf. on Artificial Intelligence*. Springer.

Rosso, M. (2008). User-based identification of Web genres. *Journal of the American Society for Information Science and Technology*, 59(7).

Santini, M. (2008). Zero, Single, or Multi? Genre of Web Pages Through the Users' Perspective. *Information Processing and Management*, 44(2):702–737.

Shepherd, M., Watters, C., and Kennedy, A. (2004). Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering*, 3(3&4):236–251.

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

Vidulin, V., Luštrek, M., and Gams, M. (2007). Training the Genre Classifier for Automatic Classification of Web Pages. In *Proc. 29th Intl. Conf. on Information Technology Interfaces*, pages 93–98.