The Impact of Noise in Web Genre Identification

Dimitrios Pritsos and Efstathios Stamatatos

University of the Aegean
Karlovassi, Samos – 83200, Greece.
{dpritsos, stamatatos}@aegean.gr

Abstract. Genre detection of web documents fits an open-set classification task. The web documents not belonging to any predefined genre or where multiple genres co-exist is considered as noise. In this work we study the impact of noise on automated genre identification within an open-set classification framework. We examine alternative classification models and document representation schemes based on two corpora, one without noise and one with noise showing that the recently proposed RFSE model can remain robust with noise. Moreover, we show how that the identification of certain genres is not practically affected by the presence of noise.

1 Introduction

The genre of web documents refer to their form, communicative purpose and it is associated with style rather than content. The ability to automatically recognize genre of web documents can enhance modern information retrieval systems by providing genre-based grouping/filtering of search results or intuitive hierarchies of web page collections. However, research in web genre identification (WGI), a.k.a automated genre identification (AGI), is limited mainly due to an inherent difficulty of defining the notion of genre and how many different genres (and sub-genres) exist [11, 10, 17, 5].

Traditionally, WGI has been viewed as a closed-set classification problem. Recently, it has been suggested that WGI better fits an open-set classification task since in any practical application it would not be easy to predefine the whole set of possible genres [13]. All web documents not belonging to a predefined genre taxonomy or documents where multiple (known or unknown) genres co-exist can be viewed as noise in WGI [11]. It is necessary to study in detail how such noise affects the effectiveness of WGI in an open-set scenario [1].

In this paper we focus on measuring and analysing the impact of noise in open-set WGI. In particular, similar to [13], we are testing two open-set models *Random Feature Subspacing Ensembles* (RFSE) and *One-Class Support Vector Machines* (OC-SVM). We are applying these models to a corpus without noise and another corpus with noise and we are examining differences in performance. The experiments indicate that both models are affected, RFSE still outperforms OC-SVM while the extracted results are more realistic. Other contributions of this paper are the examination of alternative text representation schemes for

both WGI models and the use of MinMax similarity in RFSE that seems to be helpful to improve performance on certain genres.

2 Previous Work

Most of the previous work on WGI view this problem as a closed-set classification task [11, 10, 6, 17, 5]. There is still lack of consensus about the definition of the genre itself and the web genre palette. This is due to the core characteristics of the genre notion, i.e. form, function, purpose, which are very abstract and even in the user agreement level the results are discouraging [14].

However, there is significant amount of work on several aspects of WGI, including *document representation* (e.g. character n-grams, words, part-of-speech features etc.), *term weighting schemas* (e.g. TF, TF-IDF, Binary, etc.) *feature selection methods* (e.g. frequency-based, chi-square, information gain, mutual information) and *classification models* (e.g., SVM, decision trees, aNN, etc.). Additionally, the contribution of the textual and/or the structural information has been investigated where textual information proven to be mostly useful [17, 5, 2, 15, 9, 3, 10]. As an exception, in [12], the structural information was yielding excellent results in blog/non-blog classification.

Santini in [11] defines *noise-set* as a collection of web-pages having *no genre* or *multiple genres* same as the non-noise genres of the corpus. Similarly, noise is defined as the set of web pages not belonging to any of the known genres of the corpus in [6, 2]. In these works noise was used as negative examples for training binary classifiers or as an additional ("Don't know") class rather than examining the robustness of classification models to deal with noise.

There are a couple of published studies that apply WGI on an open-set classification framework [13, 18]. However, noise-free corpora were used in their evaluation. Recently, Asheghi showed that WGI on the noisy web is more challenging as compared to noise-free corpora[1].

3 Experiments

In this paper, we use two corpora already used in previous work in WGI:

1. *7-GENRE* [15]: This is a collection of 1,400 English web pages evenly distributed into 7 genres (blogs, e-shops, FAQs, on-line front pages, listing, personal home pages, search pages).
2. *SANTINIS* [11]: This is a corpus comprising 1,400 English web pages evenly distributed into 7 genres (blogs, e-shops, FAQs, online front pages, listing, personal home pages, search pages), 80 documents evenly categorized to 4 additional genres taken from BBC web pages (DIY, editorial, bio, features) and a random selection of 1,000 English web pages taken from the SPIRIT corpus [4]. The latter can be viewed as noise in this corpus.

We are using only textual information from web pages excluding any structural information, URLs, etc. Based on the good results reported in [17, 13] as well as some preliminary experiments, the following document representation schemes are examined: *Character 4-grams*, *Words uni-grams*.

In our experiments, we do not use the noisy pages at all in the training phase. We only use them in evaluation phase. To obtain results comparable with previous studies, we followed the practice of performing 10-fold cross-validation with these corpora. In all cases, we use the Term-Frequency (TF) weighting scheme and the vocabulary only comprises the terms of the training set. Together with the RFSE model's random feature selection characteristic and the parameters selection (as explained later), the over-fitting has been prevented for the RFSE.

As concerns OC-SVM, two parameters have to be tuned: the number of features fs and ν. For the former, we used fs ={*1k, 5k, 10k, 50k, 90k*}, of most frequent terms of the vocabulary. Following the reports of previous studies [16] and some preliminary experiments, we examined ν={*0.05, 0.07, 0.1, 0.15, 0.17, 0.3, 0.5, 0.7, 0.9*}. In comparison to [13], this set of parameter values is more extended.

With respect to RFSE, four parameters should be set: the vocabulary size V, the number of feature used in each iteration f, the number of iterations I, and the threshold σ. We examined V ={*5k, 10k, 50k, 100k*}, f={*1k, 5k, 10k, 50k, 90k*}, I={*10, 50, 100*} (following the suggestion in [7] that more than 100 iterations does not improve significantly the results) and σ_s ={*0.5, 0.7, 0.9*} (based on some preliminary tests). Additionally, in this work we are testing two document similarity measures: cosine similarity (similar to [13]) and MinMax similarity (used also in a similar task by [8]).

Based on suggestions from previous work [7] and some preliminary experiments we used the following parameter values for RFSE: 100k *available Vocabulary*, 5k *Random Features per Iteration*, 0.5 σ threshold and 100 as *Iterations parameter*. It should be noted that these settings do not optimize the performance of RFSE models. They can be viewed as general settings to test the performance of RFSE in any given corpus.

On the contrary, we selected the parameters that optimize the performance of *OC-SVM* to be used as baseline in the following experiments. The optimal performance was achieved for character 4-grams in both corpora and parameter values: *50000 Features*, $\nu = 0.1$ for 7Genres and *5000 Features*, $\nu = 0.5$ for SANTINIS. The performance of these models in the following figures are referred as *baseline*.

We first applied the WGI models to noise-free 7Genres corpus. Figure 1 shows the precision-recall curves based on the parameters sets as explained above. It is evident that RFSE models are more effective than the *baseline*, although the later is optimized exactly on the 7Genre corpus. Another important observation is that all models seem to lose their effectiveness for high levels of recall. The results based on this corpus seems particularly encouraging since very high precision can be achieved for most of the standard recall values. Character n-grams seem to be more effective than word unigrams for this corpus.

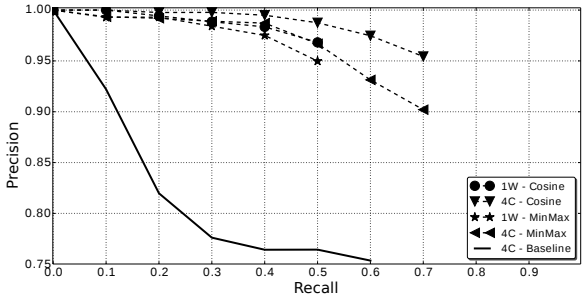

Fig. 1. Precision-Recall Curves of RFSE ensemble based on most occurred parameters found in preliminary cross-validation experiments, i.e. Vocabulary size 100k, Feature set 5k, simga threshold 0.5, Iterations 100. Corpus: 7Genres

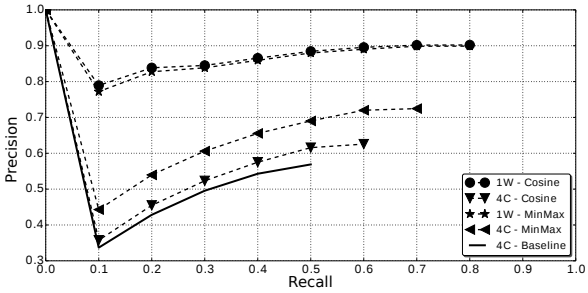

Fig. 2. Precision-Recall Curves of RFSE ensemble based on most occurred parameters found in preliminary cross-validation experiments, i.e. Vocabulary size 100k, Feature set 5k, simga threshold 0.5, Iterations 100. Corpus: SANTINIS

Next, we applied the WGI models to the SANTINIS corpus which comprises a big part of pages belonging to unknown genres (noise). Again, we show the precision-recall curves of the best OC-SVM model (baseline) and the RFSE models on the SANTINIS corpus in figure 2. As can be seen, both WGI approaches are heavily affected by the introduction of noise. Precision suddenly falls at low recall levels and then it increases quasi-linearly. This sudden fall is caused by the noisy pages and their incorrect classification to some of the known genres. It should be underlined that after that point, at standard recall level of 0.10, models with word unigrams are quite robust and achieve to maintain very high precision at high recall levels which indicates that the examined models are generally tolerant to noise. On the other hand, character n-gram models seem to be much more affected by the presence of noise. Again, RFSE is generally better than the baseline approach.

One important parameter for RFSE is the similarity measure. In figure 1 RFSE with *Cosine similarity* gives in general higher precision compared to Min-

	1W Cos		4C Cos		1W MinMax		4C MinMax		BASELINE	
	P	R	P	R	P	R	P	R	P	R
OTHER	0.93	0.95	0.96	0.60	0.93	0.95	0.95	0.73	0.90	0.60
Blog	0.32	0.96	0.17	0.98	0.33	0.96	0.18	0.99	0.28	0.50
Eshop	0.93	0.32	0.56	0.78	0.94	0.15	0.74	0.49	0.30	0.43
FAQs	1.00	0.64	1.00	0.65	0.99	0.89	0.95	0.99	1.00	0.35
Front Page	0.96	0.96	0.76	1.00	0.98	0.92	0.21	1.00	1.00	0.10
Listing	0.77	0.05	0.04	0.60	0.58	0.04	0.07	0.26	0.03	0.47
Per. Home P.	0.56	0.14	0.48	0.47	0.93	0.06	0.56	0.23	0.30	0.34
Search Page	0.76	0.54	0.74	0.82	0.56	0.51	0.64	0.79	0.86	0.41
DIY Guides	1.00	1.00	1.00	1.00	1.00	1.00	0.34	1.00	0.26	0.50
Editorial	0.72	0.90	0.53	1.00	0.35	1.00	0.09	1.00	1.00	0.25
Features	1.00	1.00	1.00	1.00	1.00	1.00	0.87	1.00	1.00	0.25
Short Bio	1.00	1.00	1.00	1.00	0.45	1.00	0.23	1.00	1.00	0.20
	F1 = .0.76		F1 = .75		F1 = .73		F1 = .60		F1 = .47	

Table 1. Precision-Recall table of *SANTINI'S corpus*, F1 has been calculated by macro-precision and macro-recall. The baseline precision-recalls is for character 4-grams with parameters $\nu = 0.5$ and 5k features RFSE models have been calculated with parameters: Vocabulary size 100k, Feature set 5k, σ threshold 0.5, Iterations 100.

Max. On the contrary, when noise is included in the corpus MinMax helps character n-gram models to improve. Word unigram models do not seem to be affected so much by the similarity measure.

Table 1 provides a closer look to precision and recall per genre of the SANTINIS corpus. As can be viewed, the identification of the *OTHER* class, corresponding to noise, is effective, especially when using word unigrams. Many genres (e.g., *Front Page*, *DIY Guides*, *Editorial*, *Features*, *Short Bio*) are not affected by the presence of noise. On the other hand, we observe that for *Blogs* and *Listing* genres precision is significantly low for character 4-grams and Cosine similarity. This is justified from the qualitative analysis reported in [11] where it is shown that a significant amount of web pages in this corpus could be assigned to both *Blog* and *Listing*, in the Spirit1000 (noise) part.

4 Conclusion

In this paper we focused on the impact of noise in WGI. This is necessary from a practical point of view since in any given application of WGI, it is impossible to predefine a complete genre palette. There will always be some web pages not belonging to the predefined genres. To test the robustness of WGI models, we used a corpus where a significant number of web pages does not belong to any of the known genres. Moreover, we examine appropriate classification models in an open-set scenario which is more realistic taking into account the lack of a consensus on genre palette and the constantly evolving web genres. Experimental results show that the precision of both RFSE and OC-SVM models are affected by noise, especially in low levels of recall, but in general RFSE based on word unigrams remains robust. MinMax seems to significantly improve the performance of character n-gram models in the presence of noise. Moreover, cer-

tain genres are not affected by the introduction of noise and their identification remains relatively easy.

References

1. Asheghi, N.R.: Human Annotation and Automatic Detection of Web Genres. Ph.D. thesis, University of Leeds (2015)
2. Dong, L., Watters, C., Duffy, J., Shepherd, M.: Binary cybergenre classification using theoretic feature measures (2006)
3. Meyer zu Eissen, S., Stein, B.: Genre classification of web pages. KI 2004: Advances in Artificial Intelligence pp. 256–269 (2004)
4. Joho, H., Sanderson, M.: The spirit collection: an overview of a large web collection. In: ACM SIGIR Forum. vol. 38, pp. 57–61. ACM (2004)
5. Kanaris, I., Stamatatos, E.: Learning to recognize webpage genres. Information Processing & Management 45(5), 499–512 (2009)
6. Kennedy, A., Shepherd, M.: Automatic identification of home pages on the web. In: System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on. pp. 99c–99c. IEEE (2005)
7. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation 45(1), 83–94 (2011)
8. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology 65(1), 178–187 (2014)
9. Lim, C. S., Lee, .K.J.Kim, .G.C.: Multiple sets of features for automatic genre classification of web documents. Information Processing and Management 41(5), 1263–1276 (2005)
10. Mason, J., Shepherd, M., Duffy, J.: An n-gram based approach to automatically identifying web page genre. In: hicss. pp. 1–10. IEEE Computer Society (2009)
11. Mehler, A., Sharoff, S., Santini, M.: Genres on the Web: Computational Models and Empirical Studies. Text, Speech and Language Technology, Springer (2010)
12. Pardo, F.M.R., Padilla, A.P.: Detecting blogs independently from the language and content. In: 1st International Workshop on Mining Social Media (MSM09-CAEPIA09). Citeseer (2009)
13. Pritsos, D.A., Stamatatos, E.: Open-set classification for automated genre identification. In: Advances in Information Retrieval, pp. 207–217. Springer (2013)
14. Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., Liu, X.: Genre based navigation on the web. In: System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on. pp. 10–pp. IEEE (2001)
15. Santini, M.: Automatic identification of genre in web pages. Ph.D. thesis, University of Brighton (2007)
16. Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87 (1999)
17. Sharoff, S., Wu, Z., Markert, K.: The web library of babel: evaluating genre collections. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation. pp. 3063–3070 (2010)
18. Stubbe, A., Ringlstetter, C., Schulz, K.U.: Genre as noise: Noise in genre. International Journal of Document Analysis and Recognition (IJDAR) 10(3-4), 199–209 (2007)