Open-set Web Genre Identification Using Distributional Features and Nearest Neighbors Distance Ratio

Dimitrios Pritsos[1], Anderson Rocha[2], and Efstathios Stamatatos[1]

[1] University of the Aegean
Karlovassi, Samos – 83200, Greece.
{dpritsos,stamatatos}@aegean.gr
[2] Institute of Computing, University of Campinas (Unicamp),
Campinas, SP, Brazil

Abstract. Web genre identification can boost information retrieval systems by providing rich descriptions of documents and enabling more specialized queries. The open-set scenario is more realistic for this task as web genres evolve over time and it is not feasible to define a universally agreed genre palette. In this work, we bring to bear a novel approach to web genre identification underpinned by distributional features acquired by doc2vec and a recently-proposed open-set classification algorithm — the nearest neighbors distance ratio classifier. We present experimental results using a benchmark corpus and a strong baseline and demonstrate that the proposed approach is highly competitive, especially when emphasis is given on precision.

Keywords: Web genre identification · Open-set classification · Distributional features

1 Introduction

Web Genre Identification (WGI) is a multi-class text classification task aiming at the association of web pages to labels (e.g., blog, e-shop, personal home page, etc.) corresponding to their form, communicative purpose, and style rather than their content. WGI can enhance the potential of information retrieval (IR) systems by allowing more complex and informative queries, whereby topic-related keywords and genre labels are combined to better express the information need of users and grouping search results by genre [31, 16]. Moreover, WGI is specially useful to enhance performance of Natural Language Processing (NLP) methods, such as part-of-speech tagging (POS) [22] and text summarization [36] by empowering genre-specific model development.

In spite of WGI's immediate applications, there are certain fundamental difficulties hardening its deployment in practice. First, there is a lack of both a consensus on the exact definition of genre [5] and a genre palette that comprises all available genres and sub-genres [33, 18, 17, 34] to aim for. New web genres appear on-the-fly and existing genres evolve over time [4]. Furthermore, it is not clear whether a whole web page should belong to a single genre or sections of the same web page can belong to different genres [8, 15]. Finally, style of documents is affected by both genre-related choices and author-related choices [24, 34].

Instead of aiming to anticipate all possible web-genres possible to appear in a practical scenario, it would be wiser to consider a proper handling of genres of interest while properly handling ”unseen” genres. In this vein, WGI can be viewed as an open-set classification task to better deal with incomplete genre palettes [2, 28, 27, 26, 37]. This scheme requires strong generalization in comparison to the traditional closed-set setup — the one in which all genres of interest are known or defined a priori. One caveat, though, is that open-set classification methods tend to perform better while operating in not-so-high dimensional manifolds. However, to date, most common and effective stylometric features in prior art, e.g., word and character n-grams, yield high-dimensional spaces [10, 34].

Aiming at properly bringing to bear the powerful algorithm modeling of open-set classification to the WGI setup, in this paper, we apply a recently-proposed open-set classification algorithm, the *Nearest Neighbors Distance Ratio* (NNDR) [19], to WGI. To produce a compact representation of web pages — more amenable to the open-set modelling — we rely upon *Distributional Features* (DF) [39] in this paper. Finally, we are also using an evaluation methodology that is more appropriate for the open-set classification framework with unstructured noise [27].

We organize the remaining of this paper into four more sections. Sec. 2 presents previous work on WGI while Sec. 3 describes the proposed approach. Sec. 4 discusses the experimental setup and obtained results. Finally, Sec. 6 draws the main conclusions of this study and presents some future work directions.

2 Related Work

Most previous studies in WGI consider the case where all web pages should belong to a predefined taxonomy of genres [14, 32, 10, 7]. Putting this setup under the vantage point of machine learning, it is the same as assuming what is known as a closed-set problem definition. However, this naïve assumption is not appropriate for most applications related to WGI as it is not possible to construct a universal genre palette a priori nor force web pages to always fall into any of the predefined genre labels. Such web pages are considered *noise* and include web documents where multiple genres co-exist [33, 13].

Santini [33] defines *structured noise* as the collection of web pages belonging to several genres, unknown during training. Such structured noise can be used as a negative class for training a binary classifier [38]. However, it is highly unlikely that such a collection represents the real distribution of pages of the web at large. On the other hand, *unstructured noise* is a random collection of pages [33] for which no genre labels are available. The effect of noise in WGI was first studied in [35, 11, 6, 13].

Open-set classification models for WGI were first described in [28, 37]. However, these models were only tested in noise-free corpora [26]. Asheghi [2] showed that it is much more challenging to perform WGI in the noisy web setup in comparison to noise-free corpora. Recently, *Ensemble Methods* were shown to achieve high effectiveness in open-set WGI setups [27].

Great attention historically on WGI has been given to the appropriate definition of features that are capable of capturing genre characteristics — which includes but are not limited to character n-grams or word n-grams, part-of-speech histograms, the frequency of the most discriminative words, etc. [10, 12–14, 17, 23, 24, 34]. Additionally, some additional useful features might come from exploiting HTML structure and/or the hyperlink functionality of web pages [1, 3, 7, 29, 40]. Recently deep learning methods have also been tested in genre detection setups with promising results [39].

3 Proposed Approach — Open-Set Web Genre Identification

3.1 Distributional Features Learning

In this study, we rely upon a Doc2Vec text representation to provide distributional features for the WGI problem [30, 20, 21]. In particular, we have implemented a special module inside our package, named *Html2Vec* [3] where a whole corpus can be used as input and one *Bag-of-Words Paragraph Vector* (PV-BOW) is returned per web-page of the corpus. PV-BOW consists of a *Neural Network* (NNet) comprising a *softmax* multi-class classifier approximating max $\frac{1}{T}\sum_{T-k}^{a=k}\log p(t_a|t_{a-k},...,t_{a+k})$. PV-BOW is trained using *stochastic gradient-descent* where the gradient is obtained via *back-propagation*. Given a sequence of training n-grams (word or character) $t_1, t_2, t_3, ..., t_T$, the objective function of the NNet is the maximized *average log-probability* $p(t_a|t_{a-k},...,t_{a+k}) = \frac{e^{y_{t_a}}}{\sum_i e^{y_i}}$.

For training the PV-BOW in this study, for each iteration, of the stochastic gradient descent, a *text window* is sampled with size w_{size}. Then a random term (n-gram) is sampled from the text window and form a classification task given the paragraph vector. Thus $y = b + s(t_1, t_2, t_3, ..., t_{w_{size}})$, where $s()$ is the sequence of word-n-grams or character-n-grams of the sampled window. Each type of n-grams is used separately as suggested in [25]. This model provides us with a representation of web pages of pre-defined dimensionality DF_{dim}.

3.2 Nearest Neighbors Distance Ratio Classifier

The Nearest Neighbors Distance Ratio (NNRD) classifier is an open-set classification algorithm introduced in [19], which in turn, is an extension upon the *Nearest Neighbors* (NN) algorithm. NNRD calculates the distance of a new sample s to its nearest neighbor t and to the closest training sample u belonging to a different class with respect to t. Then, if the ratio $d(s,t)/d(s,u)$ is higher than a threshold, the new sample is classified to the class of s. Otherwise, it is left unclassified.

It is remarkable that, in contrast to other open-set classifiers, training of NNDR requires both known samples (belonging to classes known during training) and unknown examples (belonging to other/unknown classes) of interest. In more details, the *Distance Ratio Threshold* (DRT) used to classify new samples is adjusted by maximizing the *Normalized Accuracy* (NA) $NA = \lambda A_{KS} + (1-\lambda)A_{US}$, where A_{KS} is the accuracy on known samples and A_{US} is the accuracy on unknown samples. The parameter λ regulates the mistakes trade-off on the known and unknown samples prediction. Since

[3] https://github.com/dpritsos/html2vec

usually in training phase only known classes are available, Mendes et al. [19] propose an approach to repeatedly split available training classes into two sets (known and "simulated" unknown). In our implementation of NNDR, we use cosine distance rather than the Euclidean distance because previous work found this type of distance more suitable for WGI [27].[4]

4 Experiments

4.1 Corpus

Our experiments are based on *SANTINIS*, a benchmark corpus already used in previous work in WGI [18, 27, 32]. This dataset comprises 1,400 English web-pages evenly distributed into seven genres (blog, eshop, FAQ, frontpage, listing, personal home page, search page) as well as 80 BBC web-pages evenly categorized into four additional genres (DIY mini-guide, editorial, features, short-bio). In addition, the dataset comprises a random selection of 1,000 English web-pages taken from the SPIRIT corpus [9]. The latter can be viewed as *unstructured noise* since genre labels are missing.

4.2 Experimental Setup

To represent web-pages, we use features exclusively related to textual information, excluding any structural information, URLs, etc. The following representation schemes are examined: Character 4-grams (C4G), Word unigrams (W1G), and Word 3-grams (W3G). For each of these schemes, we use either Term-Frequency (TF) weights or DF features. The feature space for TF is defined by a vocabulary V_{TF}, which is extracted based on the most frequent terms of the training set — we consider $V_{TF} = \{5k, 10k, 50k, 100k\}$. The DF space is pre-defined in the PV-BOW model — we consider $DF_{dim} = \{50, 100, 250, 500, 1000\}$.

In PV-BOW, the terms with very low-frequency in the training set are discarded. In this study, we examine $TF_{min} = \{3, 10\}$ as cutoff frequency threshold. The text window size is selected from $W_{size} = \{3, 8, 20\}$. The remaining parameters of PV-BOW are set as follows: $\alpha = 0.025$, $epochs = \{1, 3, 10\}$ and $decay = \{0.002, 0.02\}$.

Regarding the NNRD open-set classifier, there are two parameters, *lambda* and DRT, and their considered values are: $\lambda = \{0.2, 0.5, 0.7\}$, *DRT*=\{*0.4, 0.6, 0.8, 0.9*\}. All aforementioned parameters are adjusted based on grid-search using only the training part of the corpus.

For a proper comparison with prior art, we use two open-set WGI approaches with good previously reported results as baselines: Random Feature Subset Ensemble (RFSE) and one-class SVM (OCSVM) [28, 27]. All parameters of these methods have been been adjusted as suggested in [27] (based on the same corpus).

We follow the open-set evaluation framework with unstructured noise introduced in [27]. In particular, the open-set F1 score [19] is calculated over the known classes (the noisy class is excluded). The reported evaluation results are obtained by performing 10-fold cross-validation and, in each fold, we include the full set of 1,000 pages of noise. This setup is comparable to previous studies [27].

[4] https://github.com/dpritsos/OpenNNDR

Table 1: Performance of baselines and NNDR on the SANTINIS coprus. All evaluation scores are macro-averaged.

Model	Features	Dim.	Precision	Recall	AUC	F1
RFSE	TF-C4G	50k	0.739	**0.780**	0.652	0.759
RFSE	TF-W1G	50k	0.776	0.758	**0.657**	**0.767**
RFSE	TF-W3G	50k	0.797	0.722	0.615	0.758
OCSVM	TF-C4G	5k	0.662	0.367	0.210	0.472
OCSVM	TF-W1G	5k	0.332	0.344	0.150	0.338
OCSVM	TF-W3G	10k	0.631	0.654	0.536	0.643
NNDR	TF-C4G	5k	0.664	0.403	0.291	0.502
NNDR	TF-W1G	5k	0.691	0.439	0.348	0.537
NNDR	TF-W3G	10k	0.720	0.664	0.486	0.691
NNDR	DF-C4G	50	**0.829**	0.600	0.455	0.696
NNDR	DF-W1G	50	0.733	0.670	0.541	0.700
NNDR	DF-W3G	100	0.827	0.615	0.564	0.706

5 Results

We apply the baselines and NNDR in the SANTINIS corpus. In the training phase, we use only the 11 known genre classes while in test phase, we also consider an additional class (unstructured noise). Table 1 shows the performance of tested methods when either TF or DF representation schemes, based on C4G, W1G, or W4G features, are used.

First, we compare NNDR using TF features with baselines, also using this kind of features. In this case, NNDR outperforms OCSVM. On the other hand, RFSE performed better than NNDR for Macro-F1 and Macro-AUC. This is consistent for any kind of features (C4G, W1G, or W3G). There is notable difference in the dimensionality of representation used by the examined approaches though. RFSE relies upon a 50k-D manifold while NNDR and OCSVM are based on much lower dimensional spaces. It has to be noted that RFSE builds an ensemble by iteratively selecting a subset of the available features (randomly). That way, it internally reduces the dimensionality for each constituent base classifier. On the other hand, NNDR seems to be confused when thousands of features are considered as it is based on distance calculations.

Next, we compare NNDR models using either TF or DF features. There is a notable improvement when DFs are used in associated with the open-set NNDR classifier. The dimensionality of DF is much lower than TF and this seems to be crucial to improve the performance of NNDR. This is consistent for all three feature types (C4G, W1G, and W3G). NNDR with TF scheme is competitive only when W3G features are used. It has also to be noted that in all cases the selected value of parameter DRT is 0.8. This indicates that NNDR is a very robust algorithm.

Finally, the proposed approach using NNDR and DF outperforms OCSVM but it is outperformed by the strong baseline RFSE in both macro-AUC and macro F1. However, when precision is concerned, NNDR is much better. A closer look at the comparison of the two methods is provided in Fig. 1, where precision curves in 11-standard recall levels are depicted. The precision value at r_j level is interpolated as follows: $P(r_j) = max_{r_j \leq r \leq r+j+1}(P(r))$.

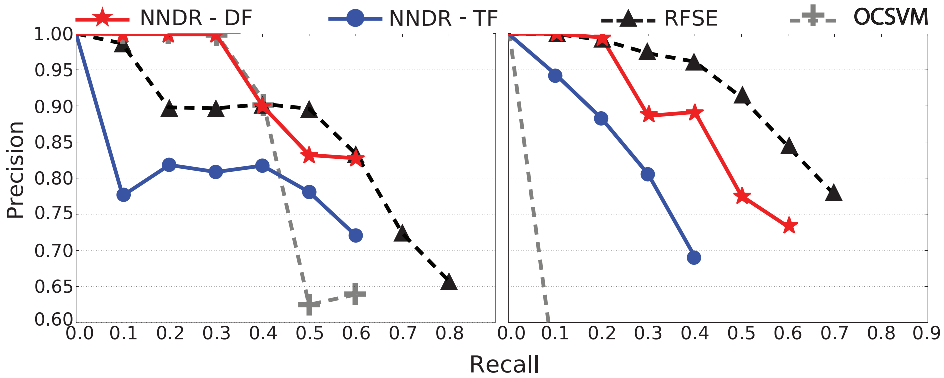

Fig. 1: Precision curves in 11-standard recall levels of the examined open-set classifiers using either W3G features (left) or W1G features (right).

The NNDR-DF model maintains very high precision scores for low levels of recall. The difference between NNDR-DF and RFSE at that point is clearer when W3G features are used. NNDR-TF is clearly worse than both NNDR-DF and RFSE. In addition, OCSVM is competitive in terms of precision only when W3G features are used but its performance drops abruptly in comparison to that of NNDR-DF. Note that the point where the curves end indicates the percentage of corpus that is left unclassified (assigned to unknown class). RFSE manages to recognize correctly larger part of the corpus, more than 70%, with respect to NNDR-DF that reaches 60%.

6 Conclusions

It seems that distributional features provide a significant enhancement to the NNDR open-set method. The low-dimensionality of DF is crucial to boost the performance of NNDR. Yet, RFSE proves to be a hard-to-beat baseline at the expense of relying upon a much higher representation space (usually in the thousands of features). However, with respect to precision, the proposed approach is much more conservative and it prefers to leave web-pages unclassified rather than guessing an inaccurate genre label. Depending on the application of WGI, precision can be considered much more important than recall and this is where our proposed algorithm shines.

Further research could focus on more appropriate distance measures within NNDR specially with recent data-driven features obtained with powerful NLP convolutional and recurrent deep networks. Moreover, alternative types of distributional features could be used (e.g., topic modeling). Finally, a combination of NNDR with RFSE models could be studied as they seem to exploit complementary views of the same problem.

Acknowledgement

Prof. Rocha thanks the financial support of FAPESP DéjàVu (Grant #2017/12646-3) and CAPES DeepEyes Grant.

References

1. Abramson, M., Aha, D.W.: What's in a url? genre classification from urls. Intelligent techniques for web personalization and recommender systems. aaai technical report. Association for the Advancement of Artificial Intelligence (2012)
2. Asheghi, N.R.: Human Annotation and Automatic Detection of Web Genres. Ph.D. thesis, University of Leeds (2015)
3. Asheghi, N.R., Markert, K., Sharoff, S.: Semi-supervised graph-based genre classification for web pages. TextGraphs-9 p. 39 (2014)
4. Boese, E.S., Howe, A.E.: Effects of web document evolution on genre classification. In: Proceedings of the 14th ACM international conference on Information and knowledge management. pp. 632–639. ACM (2005)
5. Crowston, K., Kwaśnik, B., Rubleske, J.: Problems in the use-centered development of a taxonomy of web genres. In: Genres on the Web, pp. 69–84. Springer (2011)
6. Dong, L., Watters, C., Duffy, J., Shepherd, M.: Binary cybergenre classification using theoretic feature measures. In: 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06). pp. 313–316 (2006)
7. Jebari, C.: A pure url-based genre classification of web pages. In: Database and Expert Systems Applications (DEXA), 2014 25th International Workshop on. pp. 233–237. IEEE (2014)
8. Jebari, C.: A combination based on owa operators for multi-label genre classification of web pages. Procesamiento del Lenguaje Natural **54**, 13–20 (2015)
9. Joho, H., Sanderson, M.: The spirit collection: An overview of a large web collection. SIGIR Forum **38**(2), 57–61 (2004)
10. Kanaris, I., Stamatatos, E.: Learning to recognize webpage genres. Information Processing & Management **45**(5), 499–512 (2009)
11. 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)
12. Kumari, K.P., Reddy, A.V., Fatima, S.S.: Web page genre classification: Impact of n-gram lengths. International Journal of Computer Applications **88**(13) (2014)
13. Levering, R., Cutler, M., Yu, L.: Using visual features for fine-grained genre classification of web pages. In: Hawaii International Conference on System Sciences, Proceedings of the 41st Annual. pp. 131–131. IEEE (2008)
14. 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)
15. Madjarov, G., Vidulin, V., Dimitrovski, I., Kocev, D.: Web genre classification via hierarchical multi-label classification. In: Intelligent Data Engineering and Automated Learning–IDEAL 2015, pp. 9–17. Springer (2015)
16. Malhotra, R., Sharma, A.: Quantitative evaluation of web metrics for automatic genre classification of web pages. International Journal of System Assurance Engineering and Management **8**(2), 1567–1579 (Nov 2017)
17. 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)
18. Mehler, A., Sharoff, S., Santini, M.: Genres on the Web: Computational Models and Empirical Studies. Text, Speech and Language Technology, Springer (2010)
19. Mendes Júnior, P.R., de Souza, R.M., Werneck, R.d.O., Stein, B.V., Pazinato, D.V., de Almeida, W.R., Penatti, O.A., Torres, R.d.S., Rocha, A.: Nearest neighbors distance ratio open-set classifier. Machine Learning pp. 1–28 (2016)

20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
22. Nooralahzadeh, F., Brun, C., Roux, C.: Part of speech tagging for french social media data. In: COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland. pp. 1764–1772 (2014)
23. Onan, A.: An ensemble scheme based on language function analysis and feature engineering for text genre classification. Journal of Information Science **44**(1), 28–47 (2018)
24. Petrenz, P., Webber, B.: Stable classification of text genres. Computational Linguistics **37**(2), 385–393 (2011)
25. Posadas-Durán, J.P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., Chanona-Hernández, L.: Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing **21**(3), 627–639 (2017)
26. Pritsos, D., Stamatatos, E.: The impact of noise in web genre identification. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 268–273. Springer (2015)
27. Pritsos, D., Stamatatos, E.: Open set evaluation of web genre identification. Language Resources and Evaluation pp. 1–20 (2018)
28. Pritsos, D.A., Stamatatos, E.: Open-set classification for automated genre identification. In: Advances in Information Retrieval, pp. 207–217. Springer (2013)
29. Priyatam, P.N., Iyengar, S., Perumal, K., Varma, V.: Don't use a lot when little will do: Genre identification using urls. Research in Computing Science **70**, 207–218 (2013)
30. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
31. Rosso, M.A.: User-based identification of web genres. Journal of the American Society for Information Science and Technology **59**(7), 1053–1072 (2008). https://doi.org/10.1002/asi.20798, http://dx.doi.org/10.1002/asi.20798
32. Santini, M.: Automatic identification of genre in web pages. Ph.D. thesis, University of Brighton (2007)
33. Santini, M.: Cross-testing a genre classification model for the web. In: Genres on the Web, pp. 87–128. Springer (2011)
34. 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)
35. Shepherd, M.A., Watters, C.R., Kennedy, A.: Cybergenre: Automatic identification of home pages on the web. J. Web Eng. **3**(3-4), 236–251 (2004)
36. Stewart, J.G.: Genre oriented summarization. Ph.D. thesis, Carnegie Mellon University (2009)
37. 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)
38. Vidulin, V., Luštrek, M., Gams, M.: Using genres to improve search engines. In: Proc. of the Int. Workshop Towards Genre-Enabled Search Engines. pp. 45–51 (2007)
39. Worsham, J., Kalita, J.: Genre identification and the compositional effect of genre in literature. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1963–1973 (2018)
40. Zhu, J., Zhou, X., Fung, G.: Enhance web pages genre identification using neighboring pages. In: Web Information System Engineering–WISE 2011, pp. 282–289. Springer (2011)