UNIVERSITY OF AEGEAN

DOCTORAL THESIS

Computational open-set identification of the Web-genres

Author:
Dimitrios A. PRITSOS

Supervisor:
Dr. Efstathios STAMATATOS

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy*

in the

August 21, 2018

UNIVERSITY OF AEGEAN

Abstract

Doctor of Philosophy

Computational open-set identification of the Web-genres

by Dimitrios A. PRITSOS

Web genre detection is a task that can enhance information retrieval systems by providing rich descriptions of documents and enabling more specialized queries. Most of previous studies in this field adopt the closed-set scenario where a given palette comprises all available genre labels. However this is not a realistic setup since web genres are constantly enriched with new labels and existing web genres are evolving in time. Open-set classification, where some pages used in the evaluation phase do not belong to any of the known genres, is a more realistic setup for this task. In this case, all pages not belonging to known genres can be seen as noise. This paper focuses on systematic evaluation of open-set web genre identification when the noise is either structured or unstructured. Two open-set methods combined with alternative text representation schemes and similarity measures are tested based on two benchmark corpora. Moreover, we adopt the openness test for web genre identification that enables the observation of effectiveness for a varying number of known/unknown labels.

Contents

Introduction

Web Genre Identification (WGI) concerns the association of web pages with labels that correspond to their form, communicative purpose and style rather than their content. The ability to automatically recognize the genre of web documents can enhance modern information retrieval systems by enabling genre-based grouping/filtering of search results or building intuitive hierarchies of web page collections combining topic and genre information (Braslavski, 2007; Rosso, 2008; De Assis et al., 2009). For example, a search engine can provide its users with the option to define complex queries (e.g., blogs about machine learning or eshops about sports equipment) as well as the option to navigate through results based on genre labels (e.g. social media pages, web shops, discussion forum, blogs, etc). The recognition of web genre can also enhance the effectiveness of processing the content of web pages in information extraction applications. For example, given that a set of web pages has to be part-of-speech tagged, appropriate models can be applied to each web page according to their genre (Nooralahzadeh, Brun, and Roux, 2014). However, research in WGI is relatively limited due to fundamental difficulties emanating from the genre notion itself.

The most significant difficulties in the WGI domain are: (1) There is not a consensus on the exact definition of genre (Crowston, Kwaśnik, and Rubleske, 2011); (2) There is not a common genre palette that comprises all available genres and sub-genres (Santini, 2011; Mehler, Sharoff, and Santini, 2010; Mason, Shepherd, and Duffy, 2009b; Sharoff, Wu, and Markert, 2010a), moreover, genres are evolving in time since new genres are born or existing genres are modified (Boese and Howe, 2005); (3) It is not clear whether a whole web page should belong to a genre or sections of the same web page can belong to different genres (Jebari, 2015; Madjarov et al., 2015); (4) Style of documents is affected by both genre-related choices and author-related choices (Petrenz and Webber, 2011; Sharoff, Wu, and Markert, 2010b). As a result, it is hard to accurately distinguish between personal style characteristics and genre properties when style is quantified.

Most previous studies in WGI consider the case where all web pages should belong to a predefined taxonomy of genres (Lim, 2005; Santini, 2007; Kanaris and Stamatatos, 2009; Jebari, 2014). However, this naive assumption is not appropriate for most applications related with WGI. Since it is not possible to construct a universal genre palette, there should always exist web pages that would not fall into any of the predefined genre labels. We call such web pages *noise* which also includes web documents where multiple genres (predefined or not) co-exist (Santini, 2011; Levering, Cutler, and Yu, 2008). The vast majority of previous work in WGI avoid to examine the problems arising from the presence of noise and as a result it is not possible to estimate the effectiveness of most existing WGI approaches in realistic conditions.

To handle noise in WGI there are two options. First, to adopt the closed-set classification setup having one predefined category devoted to noise. Since this category would comprise all web pages not belonging to the known genre labels, it would not be homogeneous. Moreover, this noise class would be much more greater with respect to the other genres causing class imbalance problems. The second option is to adopt the open-set classification setting where it is possible for some web pages not to be classified into any of the predefined genre categories (Pritsos and Stamatatos, 2013). This setup avoids the problem of class imbalance caused by numerous noisy pages and also avoids the problem of handling a diverse and highly heterogeneous class. On the other hand, open-set classification requires strong generalization with respect to the closed-set setup (Scheirer et al., 2013).

A great variety of features to quantify the stylistic choices related to genre have been proposed in previous work. These are mainly based on textual content (e.g., character and word n-grams) (Mason, Shepherd, and Duffy, 2009a; Sharoff, Wu, and Markert, 2010b) and form or structure of the web page (html tags, image count, links count, etc.) (Lim, 2005; Levering, Cutler, and Yu, 2008). Both sources of information are useful and usually their combination enhances a WGI model (Kanaris and Stamatatos, 2009). However, features extracted from textual content are more robust since they do not depend on technology or format used to create a web page and therefore they are more likely to remain stable in time.

In this paper, we focus on the evaluation of WGI in realistic conditions where we assume that the given genre palette covers only a subset of existing genres. Any web documents that does not fall into the predefined genre categories is considered as noise. To be able to handle noise, we adopt the open-set classification setup. In particular, we are testing two open-set classification models, one based on *One-Class Support Vector Machines* (OCSVM) and another based on *Random Feature Subspacing Ensembles* (RFSE). Several text representation schemes based on textual content are examined and we focus on the appropriate selection of parameter settings for each model. Using two benchmark corpora we perform a systematic evaluation of WGI models when noise is either unstructured (the true genre of noisy pages is not available) or structured (the true genre of noisy pages is available). In order to handle the latter case, we employ the openness test in WGI that provides a detailed view of performance for a varying number of known/unknown labels. This test has already been used in visual object recognition (Scheirer et al., 2013) and it perfectly fits the WGI task.

The rest of the paper is organized as follows. In section **??**, previous work on WGI is described. Section **??** analytically presents the open-set classification models used in this study. In section **??**, the benchmark corpora and the setup of the conducted experiments are described while in section **??** the results of the conducted experiments are presented. Finally, in section **??** the main conclusions drawn from this study are summarized and future work directions are discussed.

Chapter 1

Web Genre Identification: A Survey

1.1 Web Genre Temporal Property

1.2 Introduction

Most previous work in WGI follows a typical closed-set text categorization approach where, first, features are extracted from documents and, then, a classifier is built to distinguish between classes. Attention is paid to the appropriate definition of features that are able to capture genre characteristics and should not be affected by topic shifts or personal style choices. To this end, several document representation features have been proposed and are related with textual content, e.g. character n-grams, word n-grams, part-of-speech histograms etc. (Kumari, Reddy, and Fatima, 2014; Petrenz and Webber, 2011; Mason, Shepherd, and Duffy, 2009b; Sharoff, Wu, and Markert, 2010a) as well as the form, structure, and visual appearance of web documents, e.g., html tags, number of images, scripts etc. (Lim, 2005; Levering, Cutler, and Yu, 2008). Usually, the combination of features from different sources enhances the robustness of WGI approaches (Levering, Cutler, and Yu, 2008; Kanaris and Stamatatos, 2009). Another useful source of information is the URL of web documents (Abramson and Aha, 2012; Jebari, 2014; Priyatam et al., 2013).

An alternative approach to WGI exploits the connection of the web pages via hyperlinks. A ranking algorithm called GenreSim has used this information in combination to textual and structural information for improving WGI performance (Zhu, Zhou, and Fung, 2011). Another study is based on the web-graph and the implicit genre relation among web pages assuming that neighbouring web pages are more likely to belong to the same genre, a property called *homophily*. Then, the content of neighboring pages is used to enhance the representation of a given web page in a semi-supervised learning framework (Asheghi, Markert, and Sharoff, 2014).

The majority of previous studies in WGI disregard the presence of noise. Santini (Santini, 2011) defines *structured noise* as the collection of web pages belonging to several genres. Such structured noise can be used as a negative class for training a binary classifier (Vidulin, Luštrek, and Gams, 2007). However, it is highly unlikely that such a collection represents the real distribution of pages on the web. On the other hand, *unstructured noise* is a random collection of pages (Santini, 2011). The effect of noise in WGI was first studied in (Shepherd, Watters, and Kennedy, 2004; Kennedy and Shepherd, 2005) where predefined genres were personal, organizational, and corporate home pages while noise consisted of non-home pages. However, the distribution of pages into these four categories was practically balanced, hence it was not realistic. Dong et al.(Dong et al., 2006) uses noise as the majority class in an experiment where 190 instances from personal homepage, FAQ, and e-shop categories were used in combination with 600 noise pages. Similarly, Levering et al (Levering, Cutler, and Yu, 2008) uses about 200 instances for the predefined genres of store homepages, product lists, and product descriptions in combination with about 800 other pages (noise).

Concerning the classification models involved in WGI studies, when a given genre taxonomy is utilized and there is no noise, then well-known machine learning models, like SVMs, decision trees, neural networks, naive Bayes, Random Forests, etc. are used (Lim, 2005; Santini, 2007; Kanaris and Stamatatos, 2009; Jebari, 2015; Sharoff, Wu, and Markert, 2010a). In case of presence of noise, in a clustering framework described in (Kennedy and Shepherd, 2005) one cluster is built for each predefined class and another cluster is built for the noise. However, the most common approach to handle noise is to build binary classifiers where the positive class is based on a certain predefined category and the negative class is based on the concatenation of all other predefined categories plus the noise (Kennedy and Shepherd, 2005; Dong et al., 2006; Levering, Cutler, and Yu, 2008). Such a combination of binary classifiers can also be seen as a multi-label and open-set classification model where a web page can belong to different genres and it is possible for one page not to belong to any of the predefined genres. More concrete open-set classification models for WGI were presented in (Stubbe, Ringlstetter, and Schulz, 2007; Pritsos and Stamatatos, 2013). However, these models were only tested in noise-free corpora (Pritsos and Stamatatos, 2015). More recently, Asheghi (Asheghi, 2015) showed that it is much more challenging to perform WGI in the noisy web in comparison to noise-free corpora.

Chapter 2

Open-set and Closed-Set Classification for WGI

2.1 Introduction

Chapter 3

Evaluation Methodology for Text Categorization and for WGI

3.1 Introduction

Chapter 4

Handling the Noise of the Web-Genres

4.1 Introduction

Chapter 5

Distributional Features: "Deep Learning" of the text’s terms relations

5.1 Introduction

Chapter 6

Semi-supervised Clustering: Handling the Genres' Indeterminate Negative Samples

6.1 Introduction

Bibliography

Abramson, Myriam and David W Aha (2012). "What's in a URL? Genre Classification from URLs". In: *Intelligent techniques for web personalization and recommender systems. aaai technical report. Association for the Advancement of Artificial Intelligence*.

Asheghi, Noushin Rezapour (2015). "Human Annotation and Automatic Detection of Web Genres". PhD thesis. University of Leeds.

Asheghi, Noushin Rezapour, Katja Markert, and Serge Sharoff (2014). "Semi-supervised Graph-based Genre Classification for Web Pages". In: *TextGraphs-9*, p. 39.

Boese, Elizabeth Sugar and Adele E Howe (2005). "Effects of web document evolution on genre classification". In: *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, pp. 632–639.

Braslavski, P. (2007). "Combining relevance and genre-related rankings: An exploratory study". In: *In Proceedings of the international workshop towards genreenabled search engines: The impact of NLP*, pp. 1–4.

Crowston, Kevin, Barbara Kwaśnik, and Joseph Rubleske (2011). "Problems in the use-centered development of a taxonomy of web genres". In: *Genres on the Web*. Springer, pp. 69–84.

De Assis, Guilherme T et al. (2009). "A genre-aware approach to focused crawling". In: *World Wide Web* 12.3, pp. 285–319.

Dong, L. et al. (2006). "Binary cybergenre classification using theoretic feature measures". In:

Jebari, Chaker (2014). "A Pure URL-Based Genre Classification of Web Pages". In: *Database and Expert Systems Applications (DEXA), 2014 25th International Workshop on*. IEEE, pp. 233–237.

— (2015). "A Combination based on OWA Operators for Multi-label Genre Classification of web pages". In: *Procesamiento del Lenguaje Natural* 54, pp. 13–20.

Kanaris, I. and E. Stamatatos (2009). "Learning to recognize webpage genres". In: *Information Processing & Management* 45.5, pp. 499–512. ISSN: 0306-4573.

Kennedy, Alistair and Michael Shepherd (2005). "Automatic identification of home pages on the web". In: *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, pp. 99c–99c.

Kumari, K Pranitha, A Venugopal Reddy, and S Sameen Fatima (2014). "Web page genre classification: Impact of n-gram lengths". In: *International Journal of Computer Applications* 88.13.

Levering, Ryan, Michal Cutler, and Lei Yu (2008). "Using visual features for fine-grained genre classification of web pages". In: *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*. IEEE, pp. 131–131.

Lim C. S., Lee, K. J. Kim, G. C. (2005). "Multiple sets of features for automatic genre classification of web documents". In: *Information Processing and Management* 41.5, pp. 1263–1276.

Madjarov, Gjorgji et al. (2015). "Web Genre Classification via Hierarchical Multi-label Classification". In: *Intelligent Data Engineering and Automated Learning–IDEAL 2015*. Springer, pp. 9–17.

Mason, J., M. Shepherd, and J. Duffy (2009a). "Classifying web pages by genre: A distance function approach". In: *Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*.

Mason, J.E., M. Shepherd, and J. Duffy (2009b). "An n-gram based approach to automatically identifying web page genre". In: *hicss*. IEEE Computer Society, pp. 1–10.

Mehler, A., S. Sharoff, and M. Santini (2010). *Genres on the Web: Computational Models and Empirical Studies*. Text, Speech and Language Technology. Springer. ISBN: 9789048191789.

Nooralahzadeh, Farhad, Caroline Brun, and Claude Roux (2014). "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.

Petrenz, Philipp and Bonnie Webber (2011). "Stable classification of text genres". In: *Computational Linguistics* 37.2, pp. 385–393.

Pritsos, Dimitrios and Efstathios Stamatatos (2015). "The Impact of Noise in Web Genre Identification". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, pp. 268–273.

Pritsos, Dimitrios A and Efstathios Stamatatos (2013). "Open-Set classification for automated genre identification". In: *Advances in Information Retrieval*. Springer, pp. 207–217.

Priyatam, Pattisapu Nikhil et al. (2013). "Don't Use a Lot When Little Will Do: Genre Identification Using URLs". In: *Research in Computing Science* 70, pp. 207–218.

Rosso, Mark A. (2008). "User-based identification of Web genres". In: *Journal of the American Society for Information Science and Technology* 59.7, pp. 1053–1072. ISSN: 1532-2890. DOI: 10.1002/asi.20798. URL: http://dx.doi.org/10.1002/asi.20798.

Santini, M. (2007). "Automatic identification of genre in web pages". PhD thesis. University of Brighton.

Santini, Marina (2011). "Cross-testing a genre classification model for the web". In: *Genres on the Web*. Springer, pp. 87–128.

Scheirer, Walter J et al. (2013). "Toward open set recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.7, pp. 1757–1772.

Sharoff, S., Z. Wu, and K. Markert (2010a). "The Web library of Babel: evaluating genre collections". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pp. 3063–3070.

Sharoff, Serge, Zhili Wu, and Katja Markert (2010b). "The Web Library of Babel: evaluating genre collections." In: *LREC*. Citeseer.

Shepherd, Michael A, Carolyn R Watters, and Alistair Kennedy (2004). "Cybergenre: Automatic Identification of Home Pages on the Web." In: *J. Web Eng.* 3.3-4, pp. 236–251.

Stubbe, Andrea, Christoph Ringlstetter, and Klaus U Schulz (2007). "Genre as noise: Noise in genre". In: *International Journal of Document Analysis and Recognition (IJDAR)* 10.3-4, pp. 199–209.

Vidulin, Vedrana, Mitja Luštrek, and Matjaž Gams (2007). "Using genres to improve search engines". In: *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pp. 45–51.

Zhu, Jia, Xiaofang Zhou, and Gabriel Fung (2011). "Enhance web pages genre identification using neighboring pages". In: *Web Information System Engineering–WISE 2011*. Springer, pp. 282–289.