UNIVERSITY OF THE AEGEAN

DOCTORAL THESIS

Open-set Web Genre Identification

Author:
Dimitrios A. PRITSOS

Supervisor:
Efstathios STAMATATOS

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

at the

Dept. of Information and Communication Systems Eng.

November 14, 2019

UNIVERSITY OF THE AEGEAN

Abstract

Doctor of Philosophy

Open-set Web Genre Identification

by Dimitrios A. PRITSOS

World wide web is constantly increasing and people use information in web-pages for everyday activities. There is an emerging need for facilitating access in this huge repository in a seamless way that is in accordance with users' understanding. Genre is an important factor to characterize the properties of web-pages. Web genres (e.g., blogs, e-shop, FAQs, etc.) refer to the form, structure, and communicative purpose of web-pages rather than their topic. Web Genre Identification (WGI) provides a means to improve effectiveness of information retrieval systems by allowing sophisticated queries combining topic and genre information and ranking/grouping search results according to genre. Specialized document collections can be compiled by adopting genre-aware focused crawling. The credibility assessment of web-pages can be significantly enhanced given that information about their genre is available. Cyber-security applications like anti-phishing can also be enhanced by incorporating genre of web-pages. In case natural language technology tools should be applied to the textual part of web-pages, knowing their genre allows the selection of appropriate tools that have been trained to handle similar documents.

Existing work in WGI largely follows the closed-set classification scenario where given a genre palette and training examples for each known genre the task is to assign every new web-page to one of the known genres. However, this does not fit most of applications related to WGI. There is no consensus about the definition of a large genre palette covering most of the Web. It should be expected that large volumes of web-pages will not belong to any of the pre-defined genre labels. This could be viewed as noise in WGI. In addition, genres evolve in time, new genres emerge and existing genres are modified (e.g., blogs and micro-blogs). It seems reasonable to adopt the open-set scenario to better deal with WGI tasks. The very few existing studies focusing on open-set WGI lack an objective evaluation that will reveal their true potential.

In this thesis, we develop three open-set WGI methods. One follows the one-class classification paradigm (OCSVM) where only positive examples of a target class are used during training. Another follows the ensemble learning paradigm (RFSE) and applies random subspacing to avoid the curse of dimensionality. The third approach is a modification of k-Nearest Neighbor classifier (NNDR) that attempts to regulate the open-space risk (i.e., the area that lies away of positive examples of a class could be occupied by another, unknown, class). In addition, we examine several text representation methods including low-level and language-independent features like character n-grams and word n-grams and syntactic features like part-of-speech n-grams. We also introduce the use of distributed representations obtained by neural network language models in WGI.

Another major contribution of this thesis is the evaluation framework we propose for open-set WGI methods. In contrast to previous approaches in this field, we focus on both unstructured and structured noise. The former means that noise is composed by a random collection of web-pages without any information about their genre. The latter assumes that noise consists of web-pages of certain genres. We adopt open-set evaluation measures, variants of the well-known precision, recall, and F_1 measures, excluding true positives of the unknown class. In addition, we use graphical evaluation measures that depict the performance of the examined methods in varying conditions. We also introduce the use of the openness test in WGI studies allowing to control the homogeneity of noise and the difficulty of the task.

A series of experiments is conducted to evaluate the proposed WGI methods using the open-set evaluation framework when both unstructured and structured noise is available. The ensemble-based approach (RFSE) achieved the best overall results demonstrating its ability to handle high-dimensional and sparse representations. NNDR is significantly improved when coupled with distributed representations that provide compact and dense vectors. This method is quite competitive especially when special emphasis is put on precision rather than recall. This is important given that several WGI applications (e.g., ranking of search results) prefer to optimize precision. The one-class learning approach (OCSVM) in general is not competitive. However, it surpasses RFSE for high openness scores, that is when very few known genres are available and noise is quite heterogeneous. Several ideas for further improving the obtained results are discussed.