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.

Περίληψη

Ο παγκόσμιος Ιστός (World Wide Web) αναπτύσσεται συνεχώς και οι άνθρωποι χρησιμοποιούν πληροφορίες από ιστοσελίδες για να πραγματοποιήσουν καθημερινές δραστηριότητες. Υπάρχει επιτακτική ανάγκη να διευκολυνθεί η πρόσβαση σε αυτό το τεράστιο απόθεμα πληροφοριών με τρόπο που να συμφωνεί με τον τρόπο σκέψης των χρηστών. Το είδος (genre) των ιστοσελίδων είναι ένας σημαντικός παράγοντας για να διακρίνουμε της ιδιότητές τους. Τα είδη του Ιστού (π.χ. blogs, e-shop, FAQs, κτλ.) αναφέρονται στην μορφή, την δομή και το επικοινωνιακό σκοπό των ιστοσελίδων παρά στο θέμα τους. Η Αυτόματη Αναγνώριση Είδους Ιστοσελίδων (AAEI) παρέχει δυνατότητα βελτίωσης της επίδοσης των συστημάτων ανάκτησης πληροφορίας επιτρέποντας την δημιουργία περίπλοκων ερωτήσεων που συνδυάζουν πληροφορία θέματος και είδους καθώς και την κατάταξη και ομαδοποίηση των αποτελεσμάτων αναζήτησης με βάση το είδος τους. Εξειδικευμένες συλλογές εγγράφων μπορούν να συλλεχθούν υιοθετώντας την εστιασμένη ανίχνευση (focused crawling) με βάση το είδος. Η αξιοπιστία της πληροφορίας των ιστοσελίδων μπορεί να βελτιωθεί σημαντικά αν υπάρχει διαθέσιμη πληροφορία για το είδος τους. Εφαρμογές κυβερνο-ασφάλειας, όπως το anti-phishing, μπορούν επίσης να ενισχυθούν συμπεριλαμβάνοντας πληροφορία για το είδος των ιστοσελίδων. Σε περίπτωση που εργαλεία επεξεργασίας φυσικής γλώσσας πρέπει να εφαρμοστούν στο κειμενικό μέρος των ιστοσελίδων, η γνώση του είδους τους επιτρέπει την επιλογή κατάλληλων μοντέλων που έχουν εκπαιδευτεί να χειρίζονται αξιόπιστα παρόμοια κείμενα.

Η υπάρχουσες έρευνες στην AAEI κυρίως ακολουθούν το σενάριο της ταξινόμησης κλειστού συνόλου όπου δεδομένου ενός προκαθορισμένου συνόλου ειδών και παραδειγμάτων εκπαίδευσης για καθένα από τα είδη αυτά, ο στόχος είναι να ανατεθεί οποιαδήποτε νέα ιστοσελίδα σε ένα από τα γνωστά είδη. Όμως, αυτό δεν ταιριάζει με τις περισσότερες από τις εφαρμογές που σχετίζονται με την AAEI. Καταρχάς, δεν υπάρχει γενική συμφωνία ως προς τον ορισμό ενός μεγάλου συνόλου ειδών που θα καλύπτει το μεγαλύτερο κομμάτι του Ιστού. Θα πρέπει να αναμένεται ότι μεγάλος όγκος ιστοσελίδων δεν θα ανήκουν σε κανένα από τα προκαθορισμένα είδη. Αυτές οι ιστοσελίδες μπορούν να θεωρηθούν ως θόρυβος στην AAEI. Επιπλέον, τα είδη των ιστοσελίδων εξελίσσονται στον χρόνο, νέα είδη αναδύονται και υπάρχοντα είδη τροποποιούνται (π.χ. blogs και micro-blogs). Φαίνεται λοιπόν ότι είναι δικαιολογημένο να υιοθετηθεί το σενάριο ανοιχτού συνόλου για την AAEI. Στις πολύ λίγες υπάρχουσες μελέτες που εστιάζουν στην AAEI ανοιχτού συνόλου δεν έχει εφαρμοστεί αντικειμενική αξιολόγηση που θα αποκαλύψει τις πραγματικές δυνατότητές τους.

Στην παρούσα διατριβή, αναπτύσσουμε τρεις μεθόδους AAEI ανοιχτού συνόλου. Η πρώτη μέθοδος (OCSVM) ακολουθεί το παράδειγμα της ταξινόμησης μιας κλάσης όπου στη φάση της εκπαίδευσης χρησιμοποιούνται μόνο θετικά παραδείγματα από μία συγκεκριμένη κλάση κάθε φορά. Μια άλλη μέθοδος (RFSE) ακολουθεί την λογική της μάθησης συνόλων (ensemble learning) και εφαρμόζει τυχαία επιλογή χαρακτηριστικών για να αποφύγει την κατάρα της διαστασιμότητας. Η τρίτη

μέθοδος (NNDR) είναι τροποποίηση του ταξινομητή κ-κοντινότερων γειτόνων και προσπαθεί να εκτιμήσει το ρίσκο ανοιχτού χώρου (στην περιοχή που βρίσκεται μακριά από τα θετικά παραδείγματα εκπαίδευσης μιας γνωστής κλάσης μπορεί να βρίσκονται παραδείγματα μιας άλλης, άγνωστης, κλάσης). Επιπλέον, εξετάζουμε διάφορα σχήματα αναπαράστασης κειμένου περιλαμβάνοντας χαρακτηριστικά χαμηλού επιπέδου και ανεξάρτητα γλώσσας όπως τα ν-γράμματα λέξεων και χαρακτήρων καθώς και χαρακτηριστικά που απαιτούν συντακτική ανάλυση των κειμένων όπως τα ν-γράμματα μερών του λόγου. Επίσης, εισάγουμε στην ΑΑΕΙ την χρήση κατανεμημένων αναπαραστάσεων που εξάγονται από μοντέλα γλώσσας νευρωνικών δικτύων.

Μια άλλη κύρια συνεισφορά της παρούσας διατριβής είναι το πλαίσιο αξιολόγησης που προτείνουμε για μεθόδους ΑΑΕΙ ανοιχτού συνόλου. Σε αντίθεση με προηγούμενες εργασίες στην περιοχή αυτή, εστιάζουμε και σε αδόμητο θόρυβο και σε δομημένο θόρυβο. Το πρώτο αναφέρεται στην περίπτωση που ο θόρυβος αποτελείται από μία τυχαία συλλογή ιστοσελίδων χωρίς καμία πληροφορία για το είδος τους. Ο δομημένος θόρυβος, απ' την άλλη, αποτελείται από ιστοσελίδες συγκεκριμένων ειδών. Υιοθετούμε την χρήση μέτρων αξιολόγησης ειδικά για ταξινόμηση ανοιχτού συνόλου που είναι παραλλαγές των γνωστών μέτρων ακρίβειας, ανάκλησης και μέτρου F1. Τα μέτρα αυτά αποκλείουν τα αληθώς θετικά (true positives) παραδείγματα της άγνωστης κλάσης. Επιπλέον, χρησιμοποιούμε γραφικές μεθόδους αξιολόγησης που αναπαριστούν την επίδοσης των εξεταζόμενων μεθόδων υπό διάφορες συνθήκες. Επίσης, εισάγουμε την χρήση του ελέγχου ανοικτότητας (openness) στις μελέτες ΑΑΕΙ που επιτρέπει τον έλεγχο της ομογένειας του θορύβου και της δυσκολίας του προβλήματος.

Περιγράφονται τα πειράματα που εκτελέστηκαν για την αξιολόγηση των προτεινόμενων μεθόδων ΑΑΕΙ με την χρήση του πλαισίου αξιολόγησης ανοιχτού συνόλου όταν ο θόρυβος είναι είτε αδόμητος είτε δομημένος. Η μέθοδος βάσει συνόλων (RFSE) πέτυχε τα καλύτερα αποτελέσματα συνολικά αποδεικνύοντας την ικανότητά της να χειριστεί δεδομένα υψηλής διαστασιμότητας και αραιότητας (sparseness). Η μέθοδος NNDR βελτιώνεται σημαντικά όταν συνδυάζεται με κατανεμημένες αναπαραστάσεις που παρέχουν συμπαγή και πυκνά διανύσματα. Αυτή η μέθοδος είναι πολύ ανταγωνιστική ειδικά όταν δίνεται έμφαση στην ακρίβεια έναντι της ανάκλησης. Αυτό είναι σημαντικό δεδομένου ότι σε αρκετές εφαρμογές ΑΑΕΙ (π.χ. κατάταξη αποτελεσμάτων αναζήτησης) προτιμάται η βελτιστοποίηση της ακρίβειας. Η μέθοδος που βασίζεται στην μάθηση μιας κλάσης (OCSVM) γενικά δεν είναι ανταγωνιστική. Όμως, υπερέχει της RFSE για μεγάλες τιμές ανοικτότητας, δηλαδή όταν πολύ λίγα γνωστά είδη είναι διαθέσιμα και ο θόρυβος είναι εξαιρετικά ετερογενής. Διάφορες ιδέες για την επιπλέον βελτίωση των αποτελεσμάτων συζητούνται.