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.