Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Υπηρεσίες αναζήτησης πηγών και δεδομένων στον ιστό για υποστήριξη επιστημονικής καινοτομίας

Διδακτορική Διατριβή

του

Γεώργιου Γιαννόπουλου

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών Ε.Μ.Π. (2006)

Αθήνα, Δεκέμβριος 2013

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος 'Εκπαίδευση και Δια Βίου Μάθηση' του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων Και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Υπηρεσίες αναζήτησης πηγών και δεδομένων στον ιστό για υποστήριξη επιστημονικής καινοτομίας

Διδακτορική Διατριβή

του

Γεώργιου Γιαννόπουλου

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2006)

Συμβουλευτική Επιτροπή: Τ. Σελλής
Ι. Βασιλείου
Θ. Δαλαμάγκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 20η Δεκεμβρίου 2013.

Τ. Σελλής Καθ. ΕΜΠ	Ι. Βασιλείου Καθ. ΕΜΠ	Θ. Δαλαμάγκας Ερευνητής Β ΙΠΣΥ/ΕΚ Αθηνά
Ν. Κοζύρης Καθ. ΕΜΠ	Φώτω Αφράτη Καθ. ΕΜΠ	Α. Γ. Σταφυλοπάτης Καθ. ΕΜΠ

Χ. Παπαθεοδώρου
Αναπ. Καθ. Ιόνιο Πανεπιστήμιο

Αθήνα, Δεκέμβριος 2013

...

Γεώργιος Γιαννόπουλος
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Πρόλογος

Η παρούσα διατριβή εκπληρώνει τις απαιτήσεις για την απόκτηση διπλώματος του Διδάκτορα της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου. Η παρούσα δουλειά περιγράφει διάφορες μεθόδους αναταξινόμησης αποτελεσμάτων αναζήτησης με βάση εξατομίκευση ή διαφοροποίηση αποτελεσμάτων και με βάση σημασιολογικά δεδομένα και πραγματοποιήθηκε τα τελευταία επτά χρόνια, στο Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων.

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Τιμολέοντα Σελλή για την εμπιστοσύνη, την υποστήριξη και την καθοδήγησή του όλα αυτά τα χρόνια, καθώς και τον Ερευνητή Β΄ του ΙΠΣΥ, Θοδωρή Δαλαμάγκα για τη συνεργασία μας και την πολύτιμη βοήθειά του τα τελευταία οκτώ χρόνια. Χάρη στην πολύτιμη συνεισφορά και των δύο μπόρεσα να ασχοληθώ με ενδιαφέρουσες ερευνητικές περιοχές και να πραγματοποιήσω τις εργασίες που περιγράφονται στην παρούσα διδακτορική διατριβή.

Επιπλέον, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ιωάννη Βασιλείου, καθώς και τα μέλη του Εργαστηρίου Συστημάτων Βάσεων Γνώσεων και Δεδομένων και του Ινστιτούτου Πληροφοριακών Συστημάτων, και ειδικά τα παιδιά με τα οποία συνεργαστήκαμε κατά καιρούς, για τις ευχάριστες στιγμές που περάσαμε και τη συνεργασία τους. Τέλος, θα ήθελα να αφιερώσω την παρούσα διατριβή στην οικογένειά μου και στο Θοδωρή Δαλαμάγκα.

Γιώργος Γιαννόπουλος
Αθήνα, Δεκέμβριος 2013

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ . Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περίληψη

Η διατριβή πραγματεύεται ζητήματα αναταξινόμησης (εξατομίκευσης, διαφοροποίησης, συνδυασμού) αποτελεσμάτων αναζήτησης στον ιστό. Συγκεκριμένα, μελετώνται και προτείνονται μέθοδοι για αναταξινόμηση των αποτελεσμάτων μηχανών αναζήτησης, ώστε να ανταποκρίνονται στις εκάστοτε ανάγκες αναζήτησης πληροφορίας ενός χρήστη ή ομάδας χρηστών. Ως βάση χρησιμοποιούνται προσεγγίσεις που στηρίζονται στην εκπαίδευση συναρτήσεων αναταξινόμησης αποτελεσμάτων, χρησιμοποιώντας πληροφορία που εξάγεται από το ιστορικό αναζήτησης του χρήστη (ερωτήματα αναζήτησης, αποτελέσματα και επιλεγμένα αποτελέσματα). Επιπλέον, προτείνονται μέθοδοι για ημι-αυτόματη σημασιολογική επισημείωση εγγράφων με χρήση οντολογιών, για υβριδική αναζήτηση εγγράφων (με λέξεις κλειδιά και με έννοιες οντολογίας) και για εξατομίκευση αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα. Επίσης, εφαρμόζονται ευριστικές και ορίζονται κριτήρια για διαφοροποίηση σχολίων χρηστών σε κοινωνικά δίκτυα, καθώς και σημασιολογικών, δομημένων δεδομένων για αναζήτηση με λέξεις κλειδιά. Τέλος, εξετάζεται το πρόβλημα της αναταξινόμησης αποτελεσμάτων αναζήτησης σε οντότητες με αλλαγές στην ονοματολογία τους (βιολογικές οντότητες).

Στα πλαίσια της διατριβής μελετήθηκαν και υλοποιήθηκαν μέθοδοι για την αποτελεσματικότερη και αποδοτικότερη χρησιμοποίηση του ιστορικού αναζήτησης, μέσω της εκπαίδευσης εξειδικευμένων συναρτήσεων ταξινόμησης. Συγκεκριμένα, σε πρώτη φάση υλοποιήθηκε μία μέθοδος εμπλουτισμού της εξαγόμενης πληροφορίας από το ιστορικό του χρήστη, για ταχύτερη εκπαίδευση των συναρτήσεων. Στη συνέχεια, αναπτύχθηκαν μέθοδοι εκπαίδευσης πολλαπλών συναρτήσεων με βάση, είτε το περιεχόμενο αναζήτησης, είτε τη συμπεριφορά αναζήτησης του χρήστη. Η καινοτομία των μεθόδων έγκειται στη συγκέντρωση συνεργατικής πληροφορίας από το ιστορικό του συνόλου των χρηστών και στο διαχωρισμό αυτής της πληροφορίας σε συστάδες που αντιπροσωπεύουν διαφορετικό περιεχόμενο ή συμπεριφορά αναζήτησης. Η τελική αναταξινόμηση επιτυγχάνεται με το συνδυασμό των αποτελεσμάτων από τις συναρτήσεις που έχουν εκπαιδευτεί χρησιμοποιώντας τις παραπάνω συστάδες. Επιπλέον, στα πλαίσια της διατριβής μελετήθηκε η προσαρμογή μεθόδων διαφοροποίησης αποτελεσμάτων αναζήτησης, στο σενάριο διαφοροποίησης σχολίων χρηστών σε κοινωνικά δίκτυα. Ορίστηκαν εξειδικευμένα κριτήρια διαφοροποίησης και εφαρμόστηκαν διαφορετικοί ευριστικοί αλγόριθμοι διαφοροποίησης. Για να καταδειχθεί η αποτελεσματικότητα των προτεινόμενων προσεγγίσεων, ορίστηκαν ειδικές μετρικές αξιολόγησης της ετερογένειας συνόλων σχολίων χρηστών. Πέρα από το σενάριο διαφοροποίησης σχολίων, έγινε μία πρώτη προεργασία για τη διαφοροποίηση αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα, δηλαδή δομημένα δεδομένα

που ακολουθούν ορισμένο σχήμα και διασυνδέονται μέσω ιδιοτήτων. Επιπρόσθετα, προτάθηκαν μέθοδοι για τη βελτίωση της αναζήτησης εγγράφων μέσω σημασιολογικής επισημείωσής τους και μετέπειτα υβριδικής (με λέξεις κλειδιά και σημασιολογικής) αναζήτησής τους, καθώς και εξατομικευμένης αναζήτησης σημασιολογικών δεδομένων. Τέλος, εξετάστηκαν σχήματα ευρετηρίασης και αλγόριθμοι βαθμολόγησης οντοτήτων των οποίων η ονοματολογία μεταβάλλεται με το χρόνο, όπως συμβαίνει, για παράδειγμα, σε ορισμένες βιολογικές οντότητες.

Οι παραπάνω εργασίες αξιολογήθηκαν σε διαφορετικά σενάρια αναζήτησης, καθώς και σε ετερογενή σύνολα δεδομένων, όπως έγγραφα-ιστοσελίδες, σχόλια χρηστών, σημασιολογικές επισημειώσεις κειμένων και βιολογικές οντότητες. Επέφεραν δε αποτελέσματα που βελτίωναν τις προϋπάρχουσες βασικές μεθόδους στο κάθε πρόβλημα και οδήγησαν σε περισσότερες από δέκα δημοσιεύσεις σε διεθνή συνέδρια, workshops και περιοδικά. Επιπλέον, μέσω των παραπάνω εργασιών, προέκυψαν περαιτέρω ερευνητικά προβλήματα, τα οποία έχουν περιγραφεί στις δημοσιευμένες εργασίες και θα μπορούσαν να αποτελέσουν αντικείμενο μελλοντικής δουλειάς.

Abstract

The thesis handles re-ranking problems, including personalization, diversification, and hybrid search of entities on the web. Specifically, we studied and proposed novel methods for re-ranking web search results by capturing information needs of users or groups of users. We base our methods on ranking function training models, utilizing information extractedfrom user's search history (clickstream data - queries, results and clicked results). Further, we propose methods for semi-automatic semantic annotation of documents using ontology classes, for hybrid document search (using keywords and ontology classes) and for personalization of keyword search on semantic (RDF) data. Moreover, we evaluate/propose heuristics and introduce criteria for diversification of user comments on social networks, as well as for diversification of keyword search on semantic, structured data. Finally, we propose a first cut approach on re-ranking search results on name changing biological entities. Next, we discuss each of the above methods in more detail.

Through the presented research, we implemented methods for more effective utilization of users' search histories, through ranking function training. Specifically, first, we proposed a method for enriching the extracted information from user's clickstream data (search history), for faster ranking function training. Next, we proposed and implemented methods for training multiple ranking functions, based either on search content or on user behavior. The novelty of the methods lies on gathering collaborative information from all users and grouping this information into clusters that represent diverse content or diverse search behavior. The final ranking of the results is achieved by combining rankings produced by models trained on different clusters. Moreover, we studied the adaptation of the problem of search result diversification into the scenario of diversifying user comments on news articles. We defined problem specific diversification criteria and applied several heuristic diversification algorithms. In order to assess the effectiveness of the proposed methods, we defined problem specific evaluation measures. Beyond that, we proposed a first cut approach for diversifying keyword search results on semantic (RDF) data, utilizing the schema and structure characterizing the data and the properties interconnecting the data. Finally, we examined indexing schemes and ranking algorithms for entities whose naming changes through time, as it stands for certain categories of biological entities.

The aforementioned works were evaluated in several search scenarios, as well as on diverse datasets, such as documents-web pages, user comments, semantic annotations on documents and biological entities. The evaluation results showed that the above methods

improved the effectiveness of baseline methods in the specific research problems, leading to the publication of more than ten articles in international conferences, workshops and journals. Further, through the work done on the specific areas, new, interesting problems arised, that are described in the individual publications and can be handled in future works.

Κεφάλαιο 1

Εισαγωγή

Οι μηχανές αναζήτησης αποτελούν σήμερα ένα ευρύ και αρκετά διαδεδομένο πεδίο έρευνας και ανάπτυξης. Η σημασία τους φαίνεται τόσο από την χρήση του διαδικτύου, όπου οι εμπορικές μηχανές αναζήτησης Google, Bing, Yahoo! ανήκουν στους πιο δημοφιλείς προορισμούς των χρηστών, όσο και από τη βαρύτητα που δίνεται από την ερευνητική κοινότητα, με πληθώρα συνεδρίων και δημοσιεύσεων που καταπιάνονται με διάφορα προβλήματα στο συγκεκριμένο ερευνητικό πεδίο.

Ο ερευνητικός τομέας που ασχολείται κατά κύριο λόγο με προβλήματα αναζήτησης είναι η Ανάκτηση Πληροφορίας (Information Retrieval). Τα τελευταία χρόνια έχει δοθεί ιδιαίτερο βάρος σε προβλήματα *εξατομίκευσης αναζήτησης* (search personalization), η οποία συνίσταται στην μεταβολή/προσαρμογή των αποτελεσμάτων αναζήτησης ανάλογα με τις ανάγκες του εκάστοτε χρήστη. Ένα από τα πιο διαδεδομένα παραδείγματα που καταδεικνύουν τη σημασία της εξατομίκευσης είναι αυτό της αναζήτησης με λέξη κλειδί 'java'. Τα αποτελέσματα μίας τέτοιας αναζήτησης μπορεί να αναφέρονται στο νησί java, σε ποικιλία καφέ java ή, φυσικά, στη γλώσσα προγραμματισμού java. Στόχος της εξατομίκευσης είναι η κατανόηση της έννοιας με την οποία ο κάθε χρήστης χρησιμοποιεί τη λέξη κλειδί 'java', έτσι ώστε, αντί να του παρουσιαστούν αποτελέσματα από όλες τις παραπάνω ετερογενείς κατηγορίες, να επιστραφούν αποτελέσματα μόνο από συγκεκριμένη κατηγορία για την οποία ενδιαφέρεται.

Η εξατομίκευση των αποτελεσμάτων επιτυγχάνεται με διάφορες μεθοδολογίες που διαφέρουν ως προς την προσέγγιση, αλλά και ως προς το τελικό αποτέλεσμα που εμφανίζεται στο χρήστη. Συγκεκριμένα, χρησιμοποιούνται μεθοδολογίες εξόρυξης δεδομένων, μηχανικής μάθησης, επεξεργασίας φυσικής γλώσσας κ.α. Επίσης, τα εξατομικευμένα αποτελέσματα μπορεί να εμφανίζονται στο χρήστη αναταξινομημένα, φιλτραρισμένα ή επεκτεταμένα με νέα αποτελέσματα, προτεινόμενα από το σύστημα εξατομίκευσης.

Κοινός τόπος όλων των προσεγγίσεων είναι η εξασφάλιση κάποιας μορφής ανάδρασης από το χρήστη (user feedback), πάνω στην οποία θα στηρίξουν τη διαδικασία εξατομίκευσης. Η ανάδραση αυτή μπορεί να είναι (α) άμεση, ζητώντας από το χρήστη να αξιολογήσει τα αποτελέσματα των αναζητήσεών του ή (β) έμμεση, εξετάζοντας το ιστορικό αναζήτησης του χρήστη και εντοπίζοντας ποιά αποτελέσματα προτίμησε ('πάτησε') για συγκεκριμένα ερωτήματα. Επίσης, μπορεί να είναι ευέλικτη και να εξελίσσεται συνεχώς ανάλογα με τυχόν

αλλαγές στις προτιμήσεις αναζήτησης του χρήστη (όπως τα παραπάνω παραδείγματα), ή να είναι πιο στατική, στηριζόμενη στην κατασκευή ενός αρχικού προφίλ χρήστη με βάση το οποίο γίνεται η εξατομίκευση.

Από την άλλη πλευρά, όσο χρήσιμη είναι η εξατομίκευση αποτελεσμάτων σε ορισμένα σενάρια, άλλο τόσο σημαντική είναι και η διαφοροποίηση αποτελεσμάτων, δηλαδή η συγκέντρωση ενός μικρού συνόλου αποτελεσμάτων (και γενικότερα οντοτήτων), τα οποία να είναι όσο το δυνατόν πιο ετερογενή μεταξύ τους. Η παραπάνω λειτουργικότητα βρίσκει εφαρμογή, μεταξύ άλλων, σε εφαρμογές κοινωνικών δικτύων, όπου η συνεισφορά υλικού από τους χρήστες αποτελεί τη βάση των προσφερόμενων υπηρεσιών.

Η παρούσα εργασία επικεντρώνεται σε προβλήματα αναταξινόμησης αποτελεσμάτων με χρήση μεθοδολογιών μηχανικής μάθησης (machine learning), καθώς και ευριστικών μεθόδων, με σκοπό την εξατομίκευση (personalization) ή διαφοροποίηση (diversification) των αποτελεσμάτων αναζήτησης. Οι υλοποιημένες μέθοδοι αφορούν (α) αναταξινόμηση απλών κειμενικών αποτελεσμάτων που προκύπτουν από αναζήτηση με λέξεις κλειδιά, (β) αναταξινόμηση σημασιολογικών οντοτήτων προερχόμενων από δεδομένα που ακολουθούν συγκεκριμένη δομή και σχήμα (RDF δεδομένα) και (γ) αναταξινόμηση σχολίων χρηστών σε κοινωνικά δίκτυα. Συγκεκριμένα, υιοθετούνται οι Μηχανές Διανυσμάτων Στήριξης (Support Vector Machines) [1, 2] ως η χρησιμοποιούμενη μεθοδολογία μηχανικής μάθησης για την παραγωγή συναρτήσεων αναταξινόμησης/εξατομίκευσης αποτελεσμάτων αναζήτησης (ranking functions) και άπληστοι ευριστικοί αλγόριθμοι (greedy heuristics) [91] για την υλοποίηση αλγορίθμων διαφοροποίησης αποτελεσμάτων. Οι συναρτήσεις μηχανικής μάθησης εκπαιδεύονται με δεδομένα εκπαίδευσης που προκύπτουν από την ανάδραση του χρήστη, ενώ οι άπληστες ευριστικές χρησιμοποιούν κριτήρια διαφοροποίησης που ορίζουμε με βάση τα συγκεκριμένα προβλήματα με τα οποία καταπιανόμαστε. Στα πλαίσια της διατριβής προτείνονται και υλοποιούνται μεθοδολογίες για τα εξής προβλήματα:

1. Αυτόματος εμπλουτισμός των δεδομένων εκπαίδευσης, έτσι ώστε να επιτυγχάνεται η εκπαίδευση των συναρτήσεων αναταξινόμησης σε συντομότερο χρονικό διάστημα. Η συγκεκριμένη μέθοδος συνεισφέρει κυρίως στο πρώτο στάδιο της διαδικασίας εξατομίκευσης αποτελεσμάτων αναζήτησης, επιταχύνοντας την εφαρμογή της.

2. Συνεργατική εκπαίδευση πολλαπλών συναρτήσεων ταξινόμησης και την επιλεκτική χρησιμοποίησή τους, με σκοπό τη βελτίωση της ποιότητας εκπαίδευσης. Η συγκεκριμένη μέθοδος συνεισφέρει άμεσα στη διαδικασία εξατομίκευσης αποτελεσμάτων αναζήτησης, εκπαιδεύοντας επιμέρους συναρτήσεις αναταξινόμησης, κάθε μία από τις οποίες αντιπροσωπεύει διαφορετικούς σκοπούς και συμπεριφορές αναζήτησης (search intent - search behavior). Κάθε μία από αυτές τις συναρτήσεις μπορεί να χρησιμοποιηθεί στη συνέχεια για να εξατομικεύσει αποτελέσματα, ανάλογα με τον αναγνωριζόμενο σκοπό/συμπεριφορά αναζήτησης.

3. Εξατομίκευση και διαφοροποίηση αναζήτησης σημασιολογικών δεδομένων και συνδυασμός αναζήτησης με λέξεις κλειδιά με σημασιολογική περιήγηση στο σχήμα των δεδο-

μένων. Με τις μεθοδολογίες που προτείνουμε σε αυτό το κομμάτι της δουλειάς προσπαθούμε: (α) Να προσαρμόσουμε τις μεθοδολογίες μηχανικής μάθησης και να ορίσουμε κατάλληλα κριτήρια διαφοροποίησης, ώστε να λαμβάνουν υπόψη τις ιδιαιτερότητες των σημασιολογικών δεδομένων (συγκεκριμένη δομή και σχήμα, συσχετίσεις μεταξύ των αναζητήτουμενων οντοτήτων) και (β) να ορίσουμε τρόπους σημασιολογικής επισημείωσης κειμένων και υβριδικής αναζήτησης, τόσο με λέξεις κλειδιά, όσο και με περιήγηση στην οντολογία που χρησιμοποιείται για την επισημείωση, ορίζοντας ταυτόχρονα μετρικές ταξινόμησης, που συνδυάζουν κειμενική ομοιότητα και ποσοστό επισημείωσης των κειμένων με κλάσεις της οντολογίας.

4. Διαφοροποίηση σχολίων χρηστών σε κοινωνικά δίκτυα. Μέσω της συγκεκριμένης μεθοδολογίας προτείνονται κριτήρια διαφοροποίησης και εξετάζονται ευριστικοί αλγόριθμοι που έχουν ως στόχο την απομόνωση ετερογενών υποσυνόλων σχολίων χρηστών σε κοινωνικά δίκτυα, τα οποία εμπεριέχουν διάφορες εκφάνσεις της θεματολογίας και των απόψεων που εκφράζονται σε αυτά. Επιπλέον, επεκτείνονται υπάρχουσες μετρικές αξιολόγησης, προκειμένου να έχουν εφαρμογή στο σενάριο διαφοροποίησης σχολίων. Η προτεινόμενη μεθοδολογία, παρόλο που εξετάζεται στο σενάριο των ειδησεογραφικών άρθρων και σχολίων χρηστών, είναι αρκετά γενική ώστε να μπορεί να εφαρμοστεί και σε άλλα σενάρια (σχόλια σε facebook, twitter, forums).

1.1 Προβλήματα και προκλήσεις

1.1.1 Δεδομένα εκπαίδευσης

Προκειμένου να είναι δυνατή η εκπαίδευση συναρτήσεων αναταξινόμησης αποτελεσμάτων για εξατομίκευση, είναι απαραίτητη η συλλογή δεδομένων εκπαίδευσης που αντικατοπτρίζουν τις εκαστότε ανάγκες αναζήτησης των χρηστών. Η συλλογή των δεδομένων αυτών, όπως αναλύθηκε παραπάνω, μπορεί να γίνει είτε με άμεσο, είτε με έμμεσο τρόπο. Διάφορες μελέτες [3, 4], αλλά και η εμπειρία από τη χρήση μηχανών αναζήτησης δείχνουν ότι οι χρήστες συνήθως δυσανασχετούν με και αγνοούν συστήματα που τους ζητάνε να αξιολογήσουν τα αποτελέσματα των αναζητήσεών τους. Οπότε, η άμεση ανάδραση από τους χρήστες, αν και πιο αξιόπιστη, αφού οι χρήστες δηλώνουν ρητά τις προτιμήσεις τους για τα αποτελέσματα, δεν είναι πρακτικά εφικτή σε πραγματικά συστήματα. Αντιθέτως, η λύση που χρησιμοποιείται είναι αυτή της έμμεσης ανάδρασης, δηλαδή της ανάλυσης του ιστορικού αναζήτησης των χρηστών και της εξαγωγής συμπερασμάτων σχετικά με τις προτιμήσεις τους.

Ακόμα και στην περίπτωση της έμμεσης ανάδρασης, όμως, υπάρχει το πρόβλημα της ποσότητας των δεδομένων που συγκεντρώνεται από το ιστορικό του χρήστη. Και αυτό γιατί οι χρήστες κατά τις αναζητήσεις τους κοιτάζουν συνήθως μόνο την πρώτη σελίδα αποτελεσμάτων και, μάλιστα, επιλέγουν ('πατάνε') πολύ λιγότερα αποτελέσματα [3, 4]. Έτσι, τα δεδομένα σχετικά με τις προτιμήσεις των χρηστών που συγκεντρώνονται ανά αναζήτηση είναι ελάχιστα. Οι τεχνικές μηχανικής μάθησης που χρησιμοποιούνται, όμως, απαιτούν μία σεβαστή ποσότητα δεδομένων εκπαίδευσης έτσι ώστε να εκπαιδεύσουν ακριβείς συναρτήσεις ταξινόμησης. Άρα,

προκειμένου να συγκεντρωθεί η απαιτούμενη ποσότητα δεδομένων, το σύστημα θα πρέπει να βρίσκεται σε φάση εκπαίδευσης για μεγάλο χρονικό διάστημα ή/και να συμμετέχουν πολλοί διαφορετικοί χρήστες στην εκπαίδευση του ίδιου συστήματος. Τα παραπάνω δημιουργούν τα εξής δύο προβλήματα:

1. Το σύστημα για μεγάλο χρονικό διάστημα δεν επιτελεί τη λειτουργία στην οποία στοχεύει: την αναταξινόμηση αποτελεσμάτων με βάση τις προτιμήσεις των χρηστών. Και αυτό γιατί βρίσκεται σε φάση εκπαίδευσης η οποία προηγείται της φάσης αναταξινόμησης. Έτσι, οι χρήστες, για μεγάλο χρονικό διάστημα δεν λαμβάνουν εξατομικευμένα αποτελέσματα.

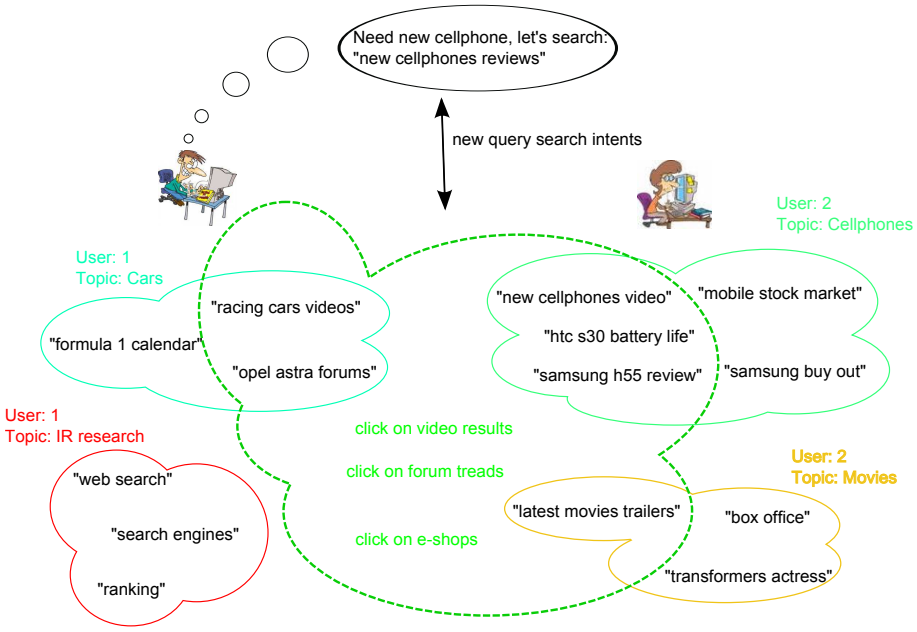
2. Στην περίπτωση που συμμετέχουν πολλοί χρήστες, συλλέγονται μεν περισσότερα δεδομένα σε λιγότερο χρονικό διάστημα, χάνεται δε η ομοιογένεια της εκπαίδευσης του συστήματος. Αυτό συμβαίνει γιατί αυξάνονται οι θεματικές περιοχές αναζήτησης, καθώς και οι διαφορετικές συμπεριφορές αναζήτησης (search behaviors). Ως αποτελέσμα, οι εκπαιδευόμενες συναρτήσεις ταξινόμησης προσπαθούν να ικανοποιήσουν όλες τις διαφορετικές συμεριφορές αναζήτησης, καταλήγοντας τελικά, να είναι μεν γενικευμένες, αλλά λιγότερο ακριβείς.

Στην παρούσα διατριβή προτείνεται μία μέθοδος εμπλουτισμού των δεδομένων εκπαίδευσης με την εκμετάλλευση του *πλήρους* ιστορικού των χρηστών και όχι μόνο του κομματιού εκείνου για το οποίο υπάρχει διαθέσιμη ανάδραση. Συγκεκριμένα, ξεκινώντας από ένα αρχικό σύνολο από δεδομένα για τα οποία υπάρχει ανάδραση, δηλαδή αξιολογήσεις των αποτελεσμάτων ή κρίσεις σχετικότητας (relevance judgments) από τους χρήστες, επεκτείνουμε αυτές τις αξιολογήσεις στο σύνολο των δεδομένων, ακόμα και αν αυτά τα δεδομένα δεν έχουν εξεταστεί από τους χρήστες. Η επέκταση των αξιολογήσεων γίνεται με βάση την θεματική ομοιότητα αξιολογημένων αποτελεσμάτων με αντίστοιχα μη αξιολογημένα. Με αυτόν τον τρόπο, πρακτικά, επιτυγχάνεται ο πολλαπλασιασμός των δεδομένων εκπαίδευσης, χωρίς καμία επιπλέον επιβάρυνση στους χρήστες.

1.1.2 Ποιότητα εκπαίδευσης

Ανεξάρτητα από το μέγεθος των δεδομένων εκπαίδευσης και τον τρόπο με τον οποίο έχουν συγκεντρωθεί, ένα πολύ σημαντικό πρόβλημα είναι η ποιότητα/ομοιογένεια αυτών των δεδομένων. Όπως αναφέρθηκε και παραπάνω, όταν το σύστημα εκπαιδεύεται με βάση το ιστορικό περισσοτέρων του ενός χρηστών, τότε πρέπει να λάβει υπόψη του διαφορετικές συμπεριφορές αναζήτησης, πάνω σε διαφορετικές θεματικές κατηγορίες, οι οποίες ενδέχεται να μειώσουν σημαντικά την ποιότητα της εκπαίδευσης. Ακόμα και στην περίπτωση που το σύστημα εξειδικεύεται σε ένα μόνο χρήστη, είναι πολύ πιθανό ο χρήστης να ψάχνει για ετερογενή θέματα και με εντελώς διαφορετικές συμπεριφορές αναζήτησης.

Γενικά, οι προσεγγίσεις που ακολουθούνται σήμερα είναι η εκπαίδευση μίας ξεχωριστής συνάρτησης ταξινόμησης για κάθε χρήστη ή η εκπαίδευση κοινής συνάρτησης ταξινόμησης για χρήστες που θεωρούνται 'όμοιοι' με βάση κάποια συγκεκριμένα κριτήρια. Επιπλέον, αυτά



Σχήμα 1.1: Κλασσικές μέθοδοι εξατομίκευσης ομαδοποιούν τα ιστορικά αναζήτησης με βάση τη θεματική περιοχή (cellphones, cars, κλπ.). Η δική μας προσέγγιση ομαδοποιεί τα ερωτήματα βασιζόμενη στη συμπεριφορά/σκοπό αναζήτησης, περιλαμβάνοντας ερωτήματα από διαφορετικές θεματικές περιοχές (διακεκομμένο σύννεφο).

τα κριτήρια συνίστανται κυρίως στην κειμενική (θεματική) ομοιότητα των αναζητήσεων των χρηστών ή/και στην ομοιότητα των προφίλ τους [5, 6].

Στην παρούσα διατριβή, μελετούμε μία ορθογώνια προσέγγιση αναταξινόμησης αποτελεσμάτων αναζήτησης, η οποία επιτρέπει σε όλους τους χρήστες να επωφελούνται το ίδιο από τα εκπαιδευόμενα μοντέλα αναταξινόμησης. Η μέθοδός μας βασίζεται στην παρατήρηση ότι οι τρέχουσες μέθοδοι επικεντρώνονται κυρίως στις θεματικές περιοχές των ερωτημάτων και στα επιμέρους ιστορικά των χρηστών, αγνοώντας έναν σημαντικό παράγοντα: την ανάλυση της συμπεριφοράς/σκοπού αναζήτησης του χρήστη, η οποία υπονοείται από τα ερωτήματα. Η συμπεριφορά αναζήτησης μπορεί να παρατηρηθεί και να κωδικοποιηθεί μέσω των αποτελεσμάτων που επιλέγουν οι χρήστες ανά ερώτημα και να χρησιμοποιηθεί για την εξαγωγή του σκοπού αναζήτησης των χρηστών. Έτσι, ο σκοπός αναζήτησης δρα ως μία λανθάνουσα μεταβλητή, η οποία εντοπίζεται μέσω της συμπεριφοράς αναζήτησης.

Στην Εικόνα 1.1 φαίνεται ένα παράδειγμα που διαφοροποιεί την προσέγγισή μας από προηγούμενες. Έστω δύο χρήστες για τους οποίους έχουν καταγραφεί από δύο ιστορικά αναζήτησης. Ο χρήστης 1 θέτει ένα νέο ερώτημα που αφορά «κριτικές για νέα κινητά τηλέφωνα». Μέχρι τώρα το ιστορικό του περιλαμβάνει αναζητήσεις στις θεματικές περιοχές της «έρευνας στην Ανάκτηση Πληροφορίας» και των «αυτοκινήτων». Ένα εξατομικευμένο μοντέλο αναζήτησης που περιορίζεται μόνο στο ιστορικό αναζήτησης του χρήστη θα χρησιμοποιούσε μόνο τις παραπάνω δύο περιοχές αναζητήσεων, παράγοντας ένα 'μέσο' μοντέλο για να αναταξινομήσει τα αποτελέσματα του νέου ερωτήματος. Έτσι, για παράδειγμα, πιθανόν

να ευνοούσε αποτελέσματα-έγγραφα της μορφής .pdf, βασιζόμενο στο γεγονός ότι ο χρήστης, σε προηγούμενες αναζητήσεις του στην Ανάκτηση Πληροφορίας θα είχε επιλέξει πολλά αποτελέσματα-δημοσιεύσεις σε μορφή .pdf. Από την άλλη πλευρά, ακόμα κι αν οι θεματικές περιοχές αναζητήσεων των δύο χρηστών δεν σχετίζονται (Ανάκτηση Πληροφορίας και αυτοκίνητα για το χρήστη 1 και κινητά και ταινίες για το χρήστη 2), ο χρήστης 1 θα μπορούσε να εκμεταλλευτεί το ιστορικό αναζήτησης του χρήστη 2 στα κινητά τηλέφωνα. Τέλος, η συμπεριφορά αναζήτησης, που εξάγεται από τα ερωτήματα και τα αντίστοιχα αποτελέσματα που επιλέχθηκαν από το χρήστη, δεν εξαρτάται πάντα από τη θεματική περιοχή του ερωτήματος. Για παράδειγμα, παρόλο που οι αναζητήσεις του χρήστη 2 φαινομενικά σχηματίζουν ένα συμπαγές ιστορικό αναζήτησης για κινητά, στην πραγματικότητα αφορούν δύο επιμέρους θεματικές: (α) αναζητήσεις για κριτικές/πληροφορίες πάνω σε κινητά και (β) αναζητήσεις για χρηματοοικονομικές πληροφορίες σχετικά με εταιρείες κινητής τηλεφωνίας. Στην πρώτη περίπτωση, σελίδες από φόρουμ ή σελίδες που περιέχουν βίντεο αναμένεται να ταιριάζουν περισσότερο στην ανάγκη αναζήτησης, ενώ στη δεύτερη περίπτωση αποτελέσματα από ειδησεογραφικές σελίδες αναμένεται να είναι πιο κατάλληλα. Έτσι, το νέο ερώτημα του χρήστη 1 θα επωφελούταν μόνο από το ιστορικό αναζήτησης του χρήστη 2 που αφορά τη θεματική (α). Το παραπάνω παράδειγμα δείχνει ότι η αναγνώριση συμπεριφορών αναζήτησης χρηστών και η εκμετάλλευσή τους για εξατομίκευση των αποτελεσμάτων αναζήτησης επεκτείνεται πέρα από το περιεχόμενο ή τους μεμονωμένους χρήστες και εξαρτάται επίσης από τους λανθάνοντες σκοπούς αναζήτησης που περικλείονται σε κάθε ερώτημα.

Δεδομένων των παραπάνω, η προτεινόμενη μέθοδός της διατριβής δεν βασίζεται σε επικεντρωμένα στο χρήστη μοντέλα, αλλά στοχεύει στην αναγνώριση του σκοπού αναζήτησης, ομαδοποιώντας ερωτήματα που επιφέρουν παρόμοιες συμπεριφορές αναζήτησης. Ο χρήστης 2, για παράδειγμα, δείχνει ιδιαίτερο ενδιαφέρον στην παρακολούθηση βίντεο σχετικά με ταινίες και με κινητά τηλέφωνα, δηλαδή επιλέγει (πατάει) συχνά τέτοιου τύπου αποτελέσματα. Η μέθοδός μας μοντελοποιεί αυτή τη συμπεριφορά, ομαδοποιώντας ερωτήματα που καταλήγουν στην επιλογή αποτελεσμάτων-βίντεο. Αυτή η συσταδοποίηση είναι ανεξάρτητη από τους επιμέρους χρήστες και αντιμετωπίζει όλα τα ερωτήματα και τους χρήστες με τον ίδιο τρόπο. Τα τελικά αποτελέσματα που θα εμφανιστούν στο χρήστη 1 για το νέο ερώτημα θα αποτελούνται από βίντεο αποτελέσματα που σχετίζονται με κινητά τηλέφωνα.

Η προτεινόμενη μέθοδος αφορά στη βελτίωση της ποιότητας εκπαίδευσης συναρτήσεων ταξινόμησης και βασίζεται στην εκπαίδευση πολλαπλών συναρτήσεων ταξινόμησης, κάθε μία από τις οποίες αντιστοιχεί (α) είτε σε συγκεκριμένη θεματική κατηγορία (αφελής μέθοδος) (β) είτε σε συγκεκριμένη συμπεριφορά αναζήτησης. Δηλαδή, με βάση το προηγούμενο παράδειγμα, και ανάλογα με την προσέγγιση που ακολουθείται, για τον χρήστη θα εκπαιδευτούν (α) δύο διαφορετικές συναρτήσεις ταξινόμησης που θα αντιστοιχούν στις διαφορετικές περιοχές αναζήτησης (δημοσιεύσεις, κινητά τηλέφωνα) ή (β) δύο διαφορετικές συναρτήσεις ταξινόμησης που θα αντιστοιχούν στις διαφορετικές συμπεριφορές αναζήτησης (επιλογή .pdf αποτελεσμάτων, επιλογή αποτελεσμάτων από forum). Με αυτόν τον τρόπο, ουσιαστικά ομαδοποιούμε το ιστορικό του χρήστη (δηλαδή τα δεδομένα εκπαίδευσης) με βάση είτε το περιεχόμενο, είτε τις λανθάνουσες συμπεριφορές αναζήτησής του. Στη συνέχεια, τα πολλαπλά μοντέλα (συ-

ναρτήσεις ταξινόμησης) συνδυάζονται σε κάθε *νέα* αναζήτηση του χρήστη, ανάλογα με την ομοιότητα της νέας αναζήτησης με κάθε μία από τις ομάδες που έχουν δημιουργηθεί, ώστε να εκτελεστεί η τελική αναταξινόμηση/εξατομίκευση των αποτελεσμάτων.

Η παραπάνω διαδικασία γίνεται με τρόπο συνεργατικό (collaborative), δηλαδή τα δεδομένα εκπαίδευσης συγκεντρώνονται από το *σύνολο* των χρηστών, ακόμα και αν αυτό το σύνολο είναι αρκετά ετερογενές. Αυτό δεν δημιουργεί πρόβλημα, αφού σκοπός της διαδικασίας είναι η εξαγωγή πολλαπλών μοτίβων (περιεχομένου ή συμπεριφοράς) και η εκμετάλλευσή τους για την εκπαίδευση αντιστοίχων πολλαπλών εξειδικευμένων συναρτήσεων ταξινόμησης. Επιπλέον, με αυτόν τον τρόπο, διαχωρίζονται οι ετερογενείς (από άποψη περιεχομένου ή συμπεριφορά-ς) αναζητήσεις του ίδιου χρήστη, ενώ παράλληλα συνδυάζονται οι παρόμοιες συμπεριφορές διαφορετικών χρηστών σε κοινά μοντέλα ταξινόμησης.

1.1.3 Μέθοδοι συνδυαστικής σημασιολογικής αναζήτησης, εξατομίκευσης και διαφοροποίησης αποτελεσμάτων σε σημασιολογικά δεδομένα

Η επισημείωση εγγράφων έχει προσελκύσει, τα τελευταία χρόνια, την προσοχή των κοινοτήτων του Σημασιολογικού Ιστού [51] και των Ψηφιακών Βιβλιοθηκών [52]. Η σημασιολογική επισημείωση συνίσταται στον χαρακτηρισμό κειμένων με έννοιες (concepts), για παράδειγμα κλάσεις (classes) μίας οντολογίας (ontology), έτσι ώστε το περιεχόμενο να αποκτά σαφή και οργανωμένη σημασιολογία. Οι επισημειώσεις βοηθούν τους χρήστες να οργανώνουν καλύτερα τα έγγραφά τους, αλλά και βελτιώνουν τις δυνατότητες αναζήτησής τους: μέσω των επισημειώσεων, οι χρήστες μπορούν να αναζητούν πληροφορίες στα έγγραφά τους, όχι μόνο μέσω της αναζήτησης με λέξεις κλειδιά, αλλά και επιλέγοντας σαφώς ορισμένες έννοιες που περιγράφουν τα πεδία αναζήτησης τών χρηστών.

Αν και οι παραδοσιακές τεχνικές ανάκτησης πληροφορίας έχουν καθιερωθεί και χρησιμοποιούνται από πληθώρα εφαρμογών, αποδεικνύονται λιγότερο αποτελεσματικές σε σενάρια ασάφειας ή συνωνυμίας εννοιών. Από την άλλη πλευρά, αναζήτηση βασισμένη μόνο στα σημασιολογικά μεταδεδομένα των εγγράφων αναμένεται επίσης να μην είναι αποτελεσμάτική, αφού: (α) δε λαμβάνει υπόψη το ίδιο το κειμενικό περιεχόμενο, (β) σε πολλές περιπτώσεις τα σημασιολογικά μεταδεδομένα δεν είναι διαθέσιμα και (γ) σημασιολογικές επισημειώσεις μπορεί να καλύπτουν ένα μικρό μέρος του κειμένου των εγγράφων. Υβριδικές μέθοδοι, οι οποίες συνδυάζουν αναζήτηση βασισμένη σε λέξεις κλειδιά και σημασιολογική αναζήτηση/περιήγηση σε έννοιες, μπορούν να ξεπεράσουν τα παραπάνω προβλήματα. Η ανάπτυξη μεθόδων και εργαλείων που ολοκληρώνουν σημασιολογική επισημείωση και αναζήτηση είναι ιδιαίτερα σημαντική. Για παράδειγμα, ερευνητές έχουν την ανάγκη να οργανώνουν, να κατηγοριοποιούν και να αναζητούν επιστημονικό υλικό (π.χ. δημοσιεύσεις) με αποτελεσματικό και αποδοτικό τρόπο. Παρόμοια, μία υπηρεσία αποδελτίωσης έχει την ανάγκη εντοπισμού ειδησεογραφικών άρθρων, επισημείωσης σημαντικών θεμάτων και αναζήτησης πληροφορίας σε αυτά.

Στη διατριβή προτείνεται, μέσω του εργαλείου GoNTogle, ένα πλαίσιο σημασιολογικής επισημείωσης και ανάκτησης κειμένων, το οποίο συνδυάζει τεχνολογίες Σημασιολογικού Ιστού

και κλασσικής Ανάκτησης Πληροφορίας. Το GoNTogle δίνει τη δυνατότητα χειροκίνητης και ημι-αυτόματης επισημείωσης βασιζόμενης σε έννοιες οντολογίας. Η επισημείωση βασίζεται σε καθιερωμένες τεχνολογίες Σημασιολογικού Ιστού, όπως η γλώσσα οντολογιών OWL -Web Ontology Language. Ταυτόχρονα, μία μέθοδος μηχανικής μάθησης (k-NN) η οποία εκμεταλλεύεται κειμενική πληροφορία και το ιστορικό επισημείωσης του χρήστη προτείνεται για την υποστήριξη του μηχανισμού αυτόματης επισημείωσης.

Ένα άλλο σημαντικό πρόβλημα που άπτεται της αναζήτησης πληροφορίας στο Σημασιολογικό Ιστό είναι η εξατομίκευση αναζήτησης σημασιολογικών δεδομένων. Η εξατομίκευση αναζήτησης είναι ένα ευρέως γνωστό πρόβλημα ανάκτησης πληροφορίας με το οποίο έχει καταπιαστεί ένας μεγάλος αριθμός εργασιών ανά τα χρόνια. Έγκειται στην αλλαγή της λίστας αποτελεσμάτων που επιστρέφονται σε ένα χρήστη (αναταξινόμηση (reranking), φιλτράρισμα (filtering) ή αναζήτηση/προσθήκη νέων αποτελεσμάτων μέσω πρότασης ερωτημάτων (query suggestion», ώστε να ταιριάζουν στις εξειδικευμένες ανάγκες αναζήτησής του. Αυτές οι ανάγκες καθορίζονται, έμμεσα, από το προφίλ (profile) του χρήστη ή το ιστορικό αναζήτησής (search history) του.

Στη βιβλιογραφία υπάρχει διαθέσιμη πληθώρα εργασιών που αντιμετωπίζουν διάφορες εκφάνσεις του προβλήματος. Υπάρχουν μελέτες πάνω στη συμπεριφορά αναζήτησης και στην ανάδραση χρήστη [9], [2], [13], [33], μελέτες πάνω στο ποιες κατηγορίες ερωτημάτων θα έπρεπε να εξατομικευθούν [38], [34], καθώς επίσης δουλειές που προτείνουν ή βελτιώνουν τεχνικές μηχανικής μάθησης για εκπαίδευση συναρτήσεων ταξινόμησης αποτελεσμάτων [35], [20], [36], [9], [2], [13]. Διάφορες εργασίες καταπιάνονται με την εκμετάλλευση του βραχυπρόθεσμου/μακροπρόθεσμου ιστορικού αναζήτησης [5], [6], των συμφραζόμενων (context) [37], [6] ή επικεντρώνονται στο να βρίσκουν κοινά ενδιαφέροντα αναζήτησης και, πάνω σε αυτά, να εφαρμόζουν συνεργατικές τεχνικές εξατομίκευσης αποτελεσμάτων [5].

Στα πλαίσια της διατριβής επεκτείνουμε το κλασσικό σενάριο της εξατομίκευσης αναζήτησης σε αδόμητα δεδομένα (unstructured data), όπως έγγραφα ή ιστοσελίδες, στο σενάριο της αναζήτησης οντοτήτων οι οποίες είναι δομημένες, συνδέονται με σχέσεις μεταξύ τους και οργανώνονται κάτω από κάποιο σχήμα (schema), όπως ισχύει για τα σημασιολογικά δεδομένα, τα οποία οργανώνονται σε RDF[1] μορφή και ακολουθούν (προαιρετικά) κάποιο σχήμα (για παράδειγμα RDFS[2] ή OWL[3]). Έστω, για παράδειγμα, το ακόλουθο σενάριο αναζήτησης, το οποίο καταδεικνύει τη χρησιμότητα της εφαρμογής εξατομικευμένης αναζήτησης σε RDF δεδομένα: ένας χρήστης ενδιαφέρεται να ψάξει για ταινίες που σχετίζονται με το Woody Allen στην DBpedia[4], ένα μεγάλο σύνολο σημασιολογικών δεδομένων. Σε αυτήν την περίπτωση θα έθετε ένα ερώτημα της μορφής: $Q = \{film, woody\ allen\}$. Όμως, ο χρήστης είναι ταυτόχρονα και θαυμαστής της Scarlett Johansson και αυτό αντανακλάται στο ιστορικό αναζήτησής/βαθμολόγησής του, όπου έχει κάνει πολλές αναζητήσεις και έχει επιλέξει αποτελέσματα σχετικά με τη Scarlett Johansson ή συνεχώς βαθμολογούσε με υψηλό σκορ ταινίες

[1] http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
[2] http://www.w3.org/TR/rdf-schema/
[3] http://www.w3.org/TR/owl2-primer/
[4] http://dbpedia.org/About

της. Έτσι, μία εξατομικευμένη λίστα αποτελεσμάτων θα έπρεπε να έχει στις υψηλότερες θέσεις αποτελέσματα από ταινίες του Woody Allen οι οποίες, με κάποιον τρόπο, σχετίζονται με τη Scarlett Johansson.

Το παραπάνω πρόβλημα δεν είναι τετριμμένο για τους εξής λόγους:

(α) Οι σχέσεις που συνδέουν τις προς αναζήτηση οντότητες και το σχήμα που τις περιγράφει επιβάλλουν τον ορισμό νέων, βασιζόμενων στη δομή και στο σχήμα χαρακτηριστικών εκπαίδευσης, πέρα από τα κλασσικά, βασιζόμενα σε Ανάκτηση Πληροφορίας, χαρακτηριστικά. Ο ορισμός χαρακτηριστικών εκπαίδευσης είναι ένα κρίσιμο κομμάτι της διαδικασίας εξατομίκευσης αποτελεσμάτων (όπως περιγράφεται και στην Ενότητα 2), αφού, μέσω αυτών των χαρακτηριστικών, *ποσοτικοποιείται* η ποιότητα ενός αποτελέσματος σε σχέση με το ερώτημα και τις ανάγκες αναζήτησης του χρήστη. Έτσι, αν χρησιμοποιηθούν κατάλληλα χαρακτηριστικά εκπαίδευσης, το ιστορικό αναζήτησης του χρήστη κωδικοποιείται και χρησιμοποιείται καλύτερα για την εξατομίκευση μελλοντικών ερωτημάτων.

(β) Ενώ μία λίστα αποτελεσμάτων στο κλασσικό σενάριο αναζήτησης αδόμητων δεδομένων αποτελείται από αυτόνομα έγγραφα, αυτό δεν ισχύει για σημασιολογικά (δομημένα) δεδομένα, όπου ένα αποτέλεσμα μπορεί να αποτελείται από διάφορες οντότητες, **μαζί με** τις σχέσεις που τις συνδέουν. Έτσι, θέματα συνδυασμού επιμέρους οντοτήτων, ώστε να σχηματίσουν ένα σύνθετο αποτέλεσμα, σε μορφή γράφου, ή θέματα υπολογισμού ενός συνολικού σκορ εξατομίκευσης από επιμέρους σκορ πρέπει να αντιμετωπιστούν.

1.1.4 Μέθοδοι διαφοροποίησης σχολίων χρηστών

Τα τελευταία χρόνια, το μέγεθος των κοινωνικών δικτύων στο διαδίκτυο μεγαλώνει εκθετικά. Όλο και περισσότεροι χρήστες κοινωνικοποιούνται μέσω Facebook, συζητούν θέματα επικαιρότητας σε φόρουμ forums ή εκφράζουν τις απόψεις τους μέσα από μπλογκς (blogs) ή το Twitter[5]. Ο κοινωνικός ιστός έχει διεισδύσει ακόμα και σε πιο παραδοσιακές εκφάνσεις του παγκόσμιου ιστού, όπως οι ειδησεογραφικές ιστοσελίδες. Μεγάλες διαδικτυακές εταιρείες, όπως η Yahoo! News[6], επιτρέπουν στους χρήστες τους να σχολιάζουν ειδησεογραφικά άρθρα, διευκολύνοντας, έτσι, την ολοκλήρωση και το διαμοιρασμό της συνεισφερόμενης πληροφορίας και απόψεων των χρηστών. Αν και αυτή η κατάσταση συνεισφέρει σημαντικά στην εξάπλωση της πληροφορίας και προωθεί την ελευθερία της έκφρασης, ταυτόχρονα εισάγει θέματα διαχείρισης και επεξεργασίας δεδομένων, λόγω του μεγάλου και ετερογενούς όγκου πληροφορίας προς επεξεργασία.

Στο σενάριο που εξετάζεται στα πλαίσια της διατριβής, δηλαδή τα σχόλια χρηστών σε ειδησεογραφικά άρθρα, πολλές φορές, άρθρα ενδέχεται να συγκεντρώνουν εκατοντάδες ή χιλιάδες σχόλια χρηστών, κάτι που καθιστά αδύνατη την επισκόπηση του συνόλου των σχολίων από τους χρήστες. Αρκετές φορές, όμως, το περιεχόμενο του άρθρου από μόνο του δεν είναι αρκετό για να σχηματίσει ο χρήστης μία πλήρη εικόνα πάνω στα θέματα με τα οποία καταπιάνεται. Η κοινή γνώμη είναι ένας σημαντικός παράγοντας που συμπληρώνει το άρθρο και αντιπροσωπεύει τη «σοφία του πλήθους». Σε αυτήν την περίπτωση, ο χρήστης χρειάζεται να

[5]https://twitter.com/

[6]http://news.yahoo.com/

μπορεί να δει ένα μικρό και όσο το δυνατόν πιο ετερογενές υποσύνολο σχολίων, το οποίο αντιπροσωπεύει διάφορες εκφάνσεις των θεμάτων του άρθρου και γνώμες/συναισθήματα των χρηστών. Επιπλέον, ο χρήστης θα πρέπει να έχει τη δυνατότητα, επιλέγοντας ένα συγκεκριμένο σχόλιο, να δει παρόμοια σχόλια. Ένα άλλο σενάριο χρήσης αφορά έναν αρχειοθέτη, ο οποίος επιθυμεί να οργανώσει υλικό σχετικό με ένα συγκεκριμένο θέμα. Και σε αυτήν την περίπτωση, ο αρχειοθέτης θα έπρεπε να μπορεί να 'επισυνάψει' στο βασικό πόρο (άρθρο) συμπληρωματική πληροφορία (ενδεικτικά σχόλια χρηστών). Αυτή η διαδικασία θα ήταν ιδιαίτερα χρήσιμη, για παράδειγμα, σε ένα μελλοντικό δημοσιογράφο που επισκοπεί παρελθοντικά γεγονότα, βοηθώντας τον να συγκεντρώσει όσο το δυνατόν πιο ετερογενές (κα άρα πιο σφαιρικό και αντικειμενικό) υλικό.

Στα πλαίσια της διατριβής, προτείνουμε ένα σύνολο από κριτήρια διαφοροποίησης, εξειδικευμένα για σχόλια χρηστών, τα οποία μπορούν να χρησιμοποιηθούν για τη συγκέντρωση ετερογενών σχολίων χρηστών σε ειδησεογραφικά άρθρα. Ισχυριζόμαστε ότι, αν και η διαφοροποίηση βασιζόμενη μόνο στο κειμενικό περιεχόμενο ενός πόρου λειτουργεί αποτελεσματικά σε διάφορα άλλα σενάρια διαφοροποίησης (για παράδειγμα στη διαφοροποίηση αποτελεσμάτων αναζήτησης), δεν επαρκεί στο σενάριο διαφοροποίησης σχολίων χρηστών. Έτσι, πέρα από το κειμενικό περιεχόμενο, μέσω των κριτηρίων που προτείνουμε, αιχμαλωτίζουμε το συναίσθημα που εκφράζεται μέσω των σχολίων, σημαντικές Ονοματικές Οντότητες (Named Entities) και την ποιότητα γραφής των σχολίων. Δηλαδή, ουσιαστικά, ορίζουμε κριτήρια που αντιπροσωπεύουν σημασιολογικά μεταδεδομένα των σχολίων, υποστηρίζοντας ότι ετερογένεια σε αυτά τα κριτήρια συνεπάγεται και ετερογένεια στα θέματα/έννοιες/απόψεις που περιέχονται στα σχόλια. Εφαρμόζουμε τα παραπάνω κριτήρια σε τρεις καθιερωμένους ευριστικούς αλγόριθμους διαφοροποίησης που παρουσιάζονται στο [96], καθώς επίσης προτείνουμε μία δική μας παραλλαγή ευριστικού αλγορίθμου. Επίσης, επεκτείνουμε υπάρχουσες έννοιες και μετρικές για την αξιολόγηση της διαφοροποίησης στο σενάριο των σχολίων χρηστών. Τέλος, επιβεβαιώνουμε την αποτελεσματικότητα των μεθόδων μας και υλοποιούμε ένα πρότυπο σύστημα διαφοροποίησης σχολίων χρηστών σε ειδησεογραφικά άρθρα.

1.2 Συνεισφορά

Η συνεισφορά της διατριβής συνοψίζεται στα παρακάτω σημεία:

1. Μελετάμε το πρόβλημα της αύξησης της ποσότητας των αρχικών δεδομένων εκπαίδευσης για ταχύτερη και ομοιογενέστερη εκπαίδευση συναρτήσεων ταξινόμησης αποτελεσμάτων.

2. Μελετάμε το πρόβλημα της εκπαίδευσης πολλαπλών συναρτήσεων ταξινόμησης εξαρτώμενων από το περιεχόμενο ή τη συμπεριφορά αναζήτησης για την βελτίωση της ποιότητας εκπαίδευσης.

3. Μελετάμε το πρόβλημα της υβριδικής αναζήτησης εγγράφων με τη βοήθεια λέξεων κλειδιών και εννοιών οντολογιών, καθώς και θέματα αναταξινόμησης αποτελεσμάτω-

ν αναζήτησης σημασιολογικών δεδομένων, με στόχο είτε την εξατομίκευση, είτε τη διαφοροποίηση των αποτελεσμάτων.

4. Μελετάμε το πρόβλημα της διαφοροποίησης σχολίων χρηστών σε κοινωνικά δίκτυα και, συγκεκριμένα, όσον αφορά τα σχόλια ειδησεογραφικών άρθρων.

5. Υλοποιούμε τις παραπάνω μεθόδους βασιζόμενοι και επεκτείνοντας/βελτιώνοντας καθιερωμένους αλγόριθμους και ευριστικές και εκτελούμε πειράματα σε διαφορετικά σετ δεδομένων τα οποία δείχνουν την αποδοτικότητα των προτεινόμενων μεθόδων, συγκρινόμενων με τις υπάρχουσες βασικές μεθόδους.

6. Εντοπίζουμε θέματα για περαιτέρω διευρεύνηση πάνω στις μεθόδους που προτείνουμε, τα οποία προσφέρουν τη δυνατότητα για μελλοντική δουλειά.

7. Παράλληλα με το προηγούμενο, παρουσιάζουμε μία προεργασία που έχουμε πραγματοποιήσει όσον αφορά: (α) στην μοντελοποίηση αναζήτησης βιολογικών οντοτήτων με αλλαγές στην ονοματολογία τους και (β) στην μοντελοποίηση του προβλήματος της διαφοροποίησης αποτελεσμάτων (σε μορφή γράφων) αναζήτησης σημασιολογικών (RDF) δεδομένων.

1.3 Δομής της έκθεσης

Η υπόλοιπη έκθεση αναπτύσσεται ως εξής: Στο Κεφάλαιο 2 δίνεται συνοπτικά το υπόβαθρο στο οποίο στηρίζεται η παρούσα δουλειά, καθώς και μία επισκόπηση σχετικών εργασιών. Στο Κεφάλαιο 3 παρουσιάζεται το πρώτο σκέλος της δουλειάς, που αφορά τη μέθοδο που προτείνουμε για επέκταση του αρχικού σετ δεδομένων εκπαίδευσης για εκπαίδευση συναρτήσεων ταξινόμησης. Στο Κεφάλαιο 4 παρουσιάζεται το δεύτερο σκέλος της δουλειάς, που αφορά τη μέθοδο που προτείνουμε για τη βελτίωση της ποιότητας εκπαίδευσης συναρτήσεων ταξινόμησης, με σκοπό τη βελτίωση της αναταξινόμησης των αποτελεσμάτων για εξατομίκευση. Στο κεφάλαιο 5 παρουσιάζονται οι προτεινόμενες προσεγγίσεις για προσαρμοστική αναζήτηση με τη βοήθεια ή πάνω σε σημασιολογικά δεδομένα και, συγκεκριμένα, η υβριδική αναζήτηση με λέξεις κλειδιά και με έννοιες οντολογίας πάνω σε σημασιολογικά επισημειωμένα δεδομένα, καθώς και η εξατομικευμένη αναζήτηση σημασιολογικών δεδομένων. Στο κεφάλαιο 6 παρουσιάζεται η εργασία πάνω στη διαφοροποίηση σχολίων χρηστών σε ειδησεογραφικά άρθρα (και γενικότερα σε κοινωνικά δίκτυα), όπου ορίζονται εξειδικευμένα κριτήρια και εφαρμόζονται ευριστικοί αλγόριθμοι διαφοροποίησης. Στο κεφάλαιο 7 παρουσιάζονται δύο προκαταρκτικές εργασίες που αφορούν τα εξής: διαφοροποίηση αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα και μοντελοποίηση και έξυπνη αναζήτηση βιολογικών οντοτήτων των οποίων η ονοματολογία μεταβάλλεται με την πάροδο του χρόνου. Οι συγκεκριμένες εργασίες μοντελοποιούν τα επιμέρους προβλήματα και προτείνουν κατευθύνσεις για την επίλυσή τους. Τέλος, στο Κεφάλαιο 8 συνοψίζουμε την παρουσίαση δουλειάς που επιτελέστηκε στα πλαίσια της διατριβής.

Κεφάλαιο 2

Σχετικές Εργασίες

2.1 Υπόβαθρο

Στις επόμενες ενότητες παρουσιάζονται οι βασικές έννοιες για την κατανόηση της διατριβής: Εκπαίδευση συναρτήσεων ταξινόμησης αποτελεσμάτων, Μηχανές Διανυσμάτων Στήριξης, τεχνολογίες Σημασιολογικού Ιστού και αλγόριθμοι και αντικειμενικές συναρτήσεις - στόχοι διαφοροποίησης.

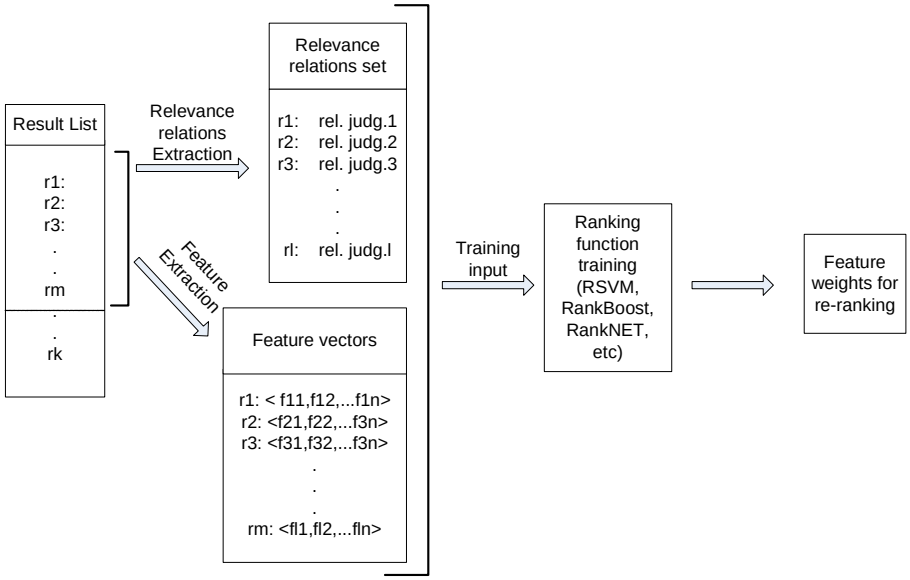
2.1.1 Εκπαίδευση συναρτήσεων ταξινόμησης αποτελεσμάτων

Στην πλειοψηφία των περιπτώσεων η εκπαίδευση μίας συνάρτησης ταξινόμησης αποτελεσμάτων αποτελείται από τρεις φάσεις: (α) Εξαγωγή προτιμήσεων χρηστών για ζεύγη ερωτημάτων-αποτελεσμάτων, (β) εξαγωγή χαρακτηριστικών (features) που περιγράφουν τη σχέση ερωτημάτων-αποτελεσμάτων και (γ) τροφοδότηση ενός αλγορίθμου μηχανικής μάθησης για την εκπαίδευση της συνάρτησης (ή αλλιώς μοντέλου) ταξινόμησης.

Εξαγωγή προτιμήσεων χρηστών. Όπως αναλύθηκε στην Εισαγωγή, οι προτιμήσεις των χρηστών μπορούν να εξαχθούν είτε άμεσα, είτε έμμεσα, με τον δεύτερο τρόπο να είναι πιο ρεαλιστικός σε πραγματικά συστήματα. Επιπλέον, οι προτιμήσεις αυτές μπορεί να εκφράζονται είτε με απόλυτο τρόπο (αριθμητικές τιμές που αντιπροσωπεύουν τη σημασία κάθε αποτελέσματος για συγκεκριμένο ερώτημα - absolute relevance) ή με σχετικό τρόπο (δυαδικές σχέσεις που δηλώνουν την προτίμηση ενός αποτελέσματος σε σχέση με κάποιο άλλο - relative relevance). Στην παρούσα εργασία θεωρούμε, άνευ βλάβης της γενικότητας, προτιμήσεις της πρώτης κατηγορίας. Οι τιμές με τις οποίες εκφράζονται οι προτιμήσεις για ένα αποτέλεσμα ονομάζονται *κρίσεις σχετικότητας*.

Εξαγωγή χαρακτηριστικών. Κάθε ζεύγος ερωτήματος - αποτελέσματος αναπαρίσταται από ένα διάνυσμα χαρακτηριστικών (feature vector) το οποίο ποσοτικοποιεί την ποιότητα 'ταιριάσματος ' μεταξύ του ερωτήματος και του αποτελέσματος. Υπάρχει μία μεγάλη ποικιλία από κατηγορίες χαρακτηριστικών που μπορούν να χρησιμοποιηθούν. Χαρακτηριστικά βασισμένα στο περιεχόμενο, τα οποία μπορούν να εξαχθούν από τον τίτλο, το σώμα κειμένου και τη διεύθυνση του αποτελέσματος, χρησιμοποιούνται για να υπολογίσουν κειμενική ομοιότητα (textual similarity) ανάμεσα σε ερώτημα και αποτέλεσμα [7]. Κάποια άλλα χαρακτηριστικά

βασίζονται σε πληροφορία από υπερσυνδέσμους (για παράδειγμα τιμές pagerank) [8] ή σε συγκεκριμένες ιδιότητες των αποτελεσμάτων, όπως το domain της διεύθυνσης ή την κατάταξη του αποτελέσματος σε διάφορες μηχανές αναζήτησης [2]. Επιπλέον, κάποια χαρακτηριστικά μπορεί να ενσωματώνουν στατιστική πληροφορία για τη συμπεριφορά των χρηστών, όπως για παράδειγμα, την απόκλιση από τη μέση τιμή που ξοδεύεται στην επισκόπηση ιστοσελίδων [9].



Σχήμα 2.1: Εκπαίδευση συνάρτησης ταξινόμησης για εξατομίκευση αποτελεσμάτων.

Εκπαίδευση. Η εκπαίδευση στοχεύει στη δημιουργία ενός διανύσματος βαρών weight vector) το οποίο αντιστοιχίζει ένα βάρος σε κάθε ένα από τα χαρακτηριστικά του διανύσματος χαρακτηριστικών. Η επιλογή των βαρών γίνεται λαμβάνοντας υπόψη τις τιμές των προτιμήσεων των χρηστών (*κρίσεις σχετικότητας*), καθώς και τις τιμές των χαρακτηριστικών για κάθε ένα από τα ζεύγη ερωτημάτων-αποτελεσμάτων του σετ δεδομένων εκπαίδευσης. Διάφορες μέθοδοι εκπαίδευσης έχουν προταθεί κατά καιρούς, όπως Ranking SVM [2], RankNET citeburges-icml05, RankBoost [11], όπως επίσης μέθοδοι βελτιστοποίησης για τις παραπάνω τεχνικές [12]. Η δική μας δουλειά υιοθετεί την πρώτη μέθοδο, δηλαδή τις Μηχανές Διανυσμάτων Στήριξης για ταξινόμηση, η οποία περιγράφεται στην επόμενη ενότητα.

2.1.2 Μηχανές Διανυσμάτων Στήριξης

Σε αυτήν την ενότητα δίνονται πληροφορίες σχετικά με το μοντέλο των Μηχανών Διανυσμάτων Στήριξης, αφενός μεν για να γίνει κατανοητή η λειτουργία τους στην εκπαίδευση συναρτήσεων ταξινόμησης, αφετέρου δε για να γίνει ευκολότερα αντιληπτή η προσαρμογή της μεθόδου μας στο συγκεκριμένο μοντέλο, όπως θα περιγραφεί σε επόμενη ενότητα.

Έστω S ένα σετ δεδομένων εκπαίδευσης, αποτελούμενο από ερωτήματα, τις λίστες των αποτελεσμάτων τους και κρίσεις σχετικότητας για (κάποια από) τα αποτελέσματα. Άνευ βλάβης της γενικότητας, έστω ότι έχουμε 3 κλάσεις σχετικότητας: $r = 0$ (άσχετο), $r = 1$ (μερικώς σχετικό) και $r = 2$ (πολύ σχετικό) για την αξιολόγηση της σχετικότητας κάθε

αποτελέσματος με το αντίστοιχο ερώτημα. Έστω, επίσης, ένας χώρος χαρακτηριστικών διάστασης d. Όπως προείπαμε, τα χαρακτηριστικά αυτά περιγράφουν την ποιότητα 'ταιριάσματος ' μεταξύ του ερωτήματος και του αποτελέσματος.

Έτσι, το σετ δεδομένων εκπαίδευσης έχει τη μορφή: $S = \{(\mathbf{x}_1, r_1), (\mathbf{x}_2, r_2), ..., (\mathbf{x}_n, r_n)\}$, όπου $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{di})$ είναι το διάνυσμα χαρακτηριστικών ενός ζεύγους ερωτήματος - αποτελέσματος, r_i είναι η κρίση (κλάση) σχετικότητας του συγκεκριμένου αποτελέσματος i και n ο αριθμός των ζευγών ερωτήματος -αποτελέσματος στο σετ δεδομένων. Οι Μηχανές Διανυσμάτων Στήριξης για ταξινόμηση επιδιώκουν να βρουν μία συνάρτηση ταξινόμησης που ικανοποιεί την παρακάτω σχέση:

$$\mathbf{x}_i \succ \mathbf{x}_j \Leftrightarrow g(\mathbf{x}_i) > g(\mathbf{x}_j) \tag{2.1}$$

όπου $\mathbf{x}_i \succ \mathbf{x}_j$ σημαίνει ότι το αποτέλεσμα i έχει υψηλότερη κατάταξη (άρα και κρίση σχετικότητας) από το αποτέλεσμα j, δηλαδή $r_i > r_j$. Θεωρώντας τη συνάρτηση g ως γραμμική συνάρτηση του $\mathbf{x}_i$, τότε $g(\mathbf{x}_i) = \langle \mathbf{w} \cdot \mathbf{x}_i \rangle$. Έτσι, η εξίσωση 2.1 ξαναγράφεται ως εξής:

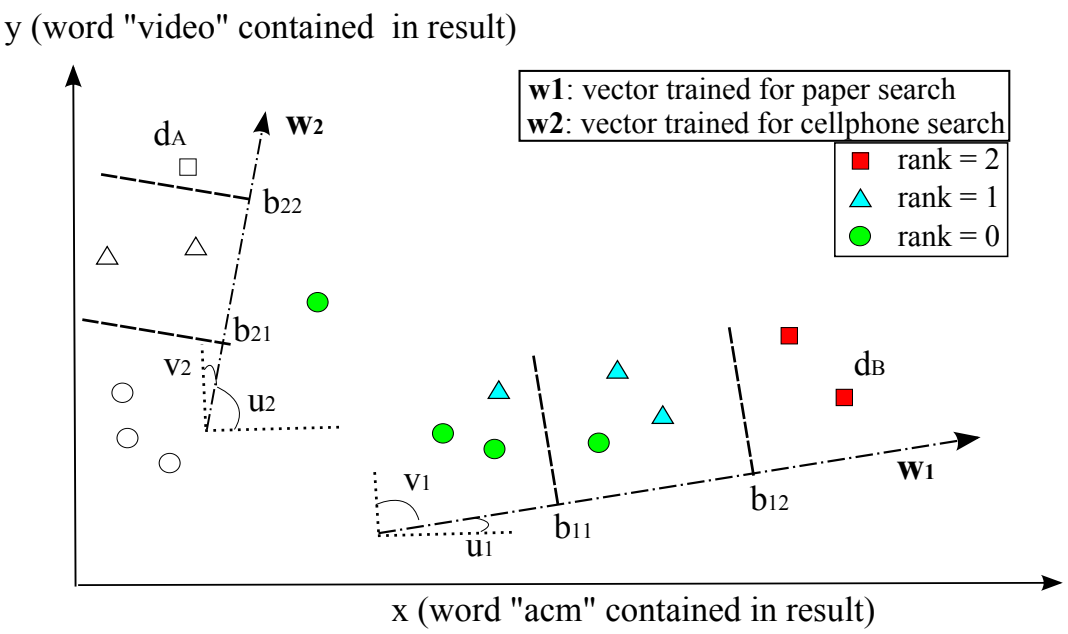
$$\mathbf{x}_i \succ \mathbf{x}_j \Leftrightarrow \langle \mathbf{w} \cdot \mathbf{x}_i - \mathbf{x}_j \rangle > 0 \tag{2.2}$$

Θεωρώντας όλα τα ζεύγη αποτελεσμάτων του σετ δεδομένων εκπαίδευσης για τα οποία $\mathbf{x}_i \succ \mathbf{x}_j$, και εισάγοντας μεταβλητές χαλάρωσης (slack variables) (για αποτελέσματα που ταξινομούνται λανθασμένα), το πρόβλημα μετατρέπεται στο εξής πρόβλημα βελτιστοποίησης:

$$\begin{aligned} &\text{ελαχιστοποίησε το: } \frac{1}{2}\|\mathbf{w}\|^2 + c\sum \xi_{ij} \\ &\quad \text{υποκείμενο σε: } \langle \mathbf{w} \cdot \mathbf{x}_i - \mathbf{x}_j \rangle > 1 - \xi_{ij}, \forall \mathbf{x}_i \succ \mathbf{x}_j, \xi_{ij} > 0 \end{aligned} \tag{2.3}$$

όπου $\|\mathbf{w}\|$ η νόρμα του $\mathbf{w}$. Σημειώνουμε ότι το $\mathbf{w}$ είναι το διάνυσμα βαρών που ποσοτικοποιεί τη σημασία κάθε χαρακτηριστικού για μία συγκεκριμένη εκπαίδευση, σε ένα συγκεκριμένο σετ δεδομένων. Το διάνυσμα αυτό είναι κάθετο σε υπερεπιφάνειες (οριζόμενες από τα σημεία b_{pq}, στα οποία 'κόβουν' το διάνυσμα) οι οποίες διαχωρίζουν αποτελέσματα διαφορετικών κλάσεων σχετικότητας, όπως φαίνεται στο Σχήμα 2.2. Στην Εξίσωση 2.3, ο πρώτος όρος $\frac{1}{2}\|\mathbf{w}\|^2$ σχετίζεται με την απόσταση των κοντινότερων στην υπερεπιφάνεια, *ορθά* ταξινομημένων αποτελεσμάτων. Ο δεύτερος όρος $\sum \xi_{ij}$ σχετίζεται με το σφάλμα που εισάγεται από τα *λανθασμένα* ταξινομημένα αποτελέσματα. Για παράδειγμα, θεωρώντας ότι τα τριγωνικά σημεία θα έπρεπε να τοποθετηθούν (ταξινομηθούν) στα δεξιά της υπερεπιφάνειας που ορίζεται από το σημείο b_{11} και τα κυκλικά σημεία στα δεξιά, τότε: (α) το $\frac{1}{2}\|\mathbf{w}_1\|^2$ σχετίζεται με την απόσταση του b_{11} από το κοντινότερό του δεξί τριγωνικό σημείο και από το κοντινότερό του αριστερό κυκλικό σημείο και (β) το $\sum \xi_{ij}$ σχετίζεται με την απόσταση του b_{11} από το κοντινότερό του αριστερό τριγωνικό σημείο από το κοντινότερό του δεξί κυκλικό σημείο. Τέλος, το c καθορίζει τη σχετική σημασία των δύο όρων στη διαδικασία εκπαίδευσης.

Το παράδειγμα της Εικόνας 2.2 δίνει μία γεωμετρική ερμηνεία του υιοθετούμενου μοντέλου. Χάριν απλότητας, θεωρούμε ότι τα διανύσματα χαρακτηριστικών αποτελούνται από δύο μόνο χαρακτηριστικά. Το χαρακτηριστικό x αντιπροσωπεύει της συχνότητα της λέξης «acm»



Σχήμα 2.2: Εκπαιδευμένα διανύσματα βάρους και υπερεπιφάνειες στον χώρο χαρακτηριστικών

στο κείμενο του αποτελέσματος, ενώ το χαρακτηριστικό y τη συχνότητα της λέξης «video» στο αποτέλεσμα. Τα γραμμοσκιασμένα γεωμετρικά σχήματα αντιπροσωπεύουν αποτελέσματα σχετικά με αναζητήσεις για δημοσιεύσεις, ενώ τα κενά σχήματα αποτελέσματα από αναζητήσεις για κινητά τηλέφωνα. Τα τετράγωνα αντιστοιχούν σε *πολύ σχετικά* αποτελέσματα, τα τρίγωνα σε *μερικώς σχετικά* και οι κύκλοι σε *άσχετα* αποτελέσματα. Σημειώνουμε ότι τα δύο παραπάνω χαρακτηριστικά εξαρτώνται από τη συχνότητα των συγκεκριμένων λέξεων στα αποτελέσματα, δηλαδή οι τιμές τους δεν επηρεάζονται από τα ερωτήματα. Στη γενική περίπτωση, βέβαια, αυτό δεν ισχύει πάντα, απλά διαλέξαμε τα συγκεκριμένα χαρακτηριστικά για ευκολία στην παρουσίαση.

Στο χώρο χαρακτηριστικών της Εικόνας 2.2 εκπαιδεύουμε δύο μοντέλα αναταξινόμησης αποτελεσμάτων, εκφραζόμενα από αντίστοιχα διανύσματα βαρών, τα $\boldsymbol{\omega}_1$ και $\boldsymbol{\omega}_2$. Αυτά τα διανύσματα αντιστοιχούν σε αναζητήσεις για δημοσιεύσεις και αναζητήσεις για κινητά αντίστοιχα. Η κατεύθυνση του κάθε διανύσματος, δηλαδή η γωνία ανάμεσα στο διάνυσμα και σε κάποιον από τους άξονες δείχνει τη σημασία του κάθε χαρακτηριστικού για τη διαδικασία της ταξινόμησης αποτελεσμάτων. Για παράδειγμα, η γωνία u_1 ανάμεσα στο διάνυσμα $\boldsymbol{\omega}_1$ και στον άξονα X είναι μικρότερη από τη γωνία v_1 ανάμεσα στο διάνυσμα και στον άξονα Y. Αυτό σημαίνει ότι μία αλλαγή στην τιμή του χαρακτηριστικού x (συχνότητα της λέξης «acm») είναι πιο πιθανό να προκαλέσει μία αλλαγή στην κατάταξη του του αποτελέσματος από ότι μία αλλαγή στην τιμή του χαρακτηριστικού y (συχνότητα της λέξης «video»). Το αντίθετο ισχύει όσον αφορά στο διάνυσμα βαρών $\boldsymbol{\omega}_2$, που σχετίζεται με αναζητήσεις για κινητά: το χαρακτηριστικό y είναι πιο σημαντικό από το x, όπως φαίνεται από την κατεύθυνση του διανύσματος $\boldsymbol{\omega}_2$.

Το παραπάνω παράδειγμα περιγράφει δύο διαφορετικές συμπεριφορές αναζήτησης, δηλαδή διαφορετικά μοτίβα αναζήτησης που ακολουθούνται από χρήστες για συγκεκριμένες κατηγο-

ρίες αναζήτησης. Το συμπέρασμα που προκύπτει είναι ότι οι συμπεριφορές αναζήτησης δεν εκφράζονται μέσω του περιεχομένου, αλλά μέσω του χώρου χαρακτηριστικών $\Xi \epsilon R^d$ που έχει επιλεγεί για να αντιπροσωπεύει τα δεδομένα αναζήτησης (clickthrough data), δηλαδή τα αποτελέσματα του ερωτήματος με την κατάταξή τους και την πληροφορία για το ποια από αυτά επέλεξε ο χρήστης. Από τα παραπάνω προκύπτει ότι οι συμπεριφορές αναζήτησης μπορούν να αναγνωριστούν μέσω της εκμετάλλευσης της κατανομής των δεδομένων αναζήτησης στο χώρο χαρακτηριστικών.

2.1.3 Τεχνολογίες σημασιολογικού ιστού

Ο Σημασιολογικός Ιστός (Semantic Web) είναι ένα σύνολο πληροφοριών, διασυνδεδεμένων με κατάλληλο τρόπο, ώστε να είναι εύκολα και με αποδοτικό τρόπο προσβάσιμες και επεξεργάσιμες από προγράμματα σε παγκόσμια κλίμακα. Μπορεί να θεωρηθεί ως μία «παγκόσμια βάση δεδομένων», της οποίας η δομή και οργάνωση επιτρέπει όχι μόνο στους ανθρώπους, αλλά και στις μηχανές να χρησιμοποιήσουν την αποθηκευμένη πληροφορία. Ο Σημασιολογικός Ιστός αποτελεί ουσιαστικά μία επέκταση του Παγκόσμιου Ιστού (World Wide Web), η οποία επιτρέπει την αποτελεσματικότερη συνεργασία ανθρώπων και υπολογιστών, σύμφωνα με τον εμπνευστή του (και εμπνευστή των WWW, URIs, HTTP και HTML) Tim Berners-Lee.

Το πρόβλημα με τον Παγκόσμιο Ιστό είναι ότι το μεγαλύτερο μέρος της διαθέσιμης πληροφορίας οργανώνεται με τρόπο που καθιστά δύσκολη την επεξεργασία της από μία μηχανή. Η πληροφορία αποθηκεύεται συνήθως στη μορφή HTML αρχείων, τα οποία προσφέρουν μόνο οπτική απεικόνιση και όχι σημασιολογική ταξινόμηση της πληροφορίας. Αυτό έχει ως αποτέλεσμα, ένας άνθρωπος, διαβάζοντας μία σελίδα HTML, να μπορεί να διαχωρίσει σημασιολογικά την πληροφορία που περιέχει, αλλά αυτό να είναι αδύνατο για μία εφαρμογή.

Αυτό που επιδιώκεται με το Σημασιολογικό Ιστό είναι η ρητή και με σαφήνεια σημασιολογική επισημείωση της πληροφορίας, έτσι ώστε να διευκολύνεται η αυτοματοποιημένη επεξεργασία και ολοκλήρωσή της από μία μηχανή. Στηρίζεται στην δυνατότητα ορισμού από την XML (Extensible Markup Language) σχημάτων προσαρμοσμένων ετικετών, στη δυνατότητα ευέλικτης παρουσίασης δεδομένων της RDF (Resource Description Framework) και στη δυνατότητα της OWL για τυπική περιγραφή της σημασιολογίας και ορολογίας ενός εγγράφου.

Το παραπάνω πρόβλημα που επιδιώκεται να λυθεί με το Σημασιολογικό Ιστό δεν περιορίζεται μόνο στο διαδίκτυο, αλλά αφορά και οποιονδήποτε άλλο τομέα στον οποίο προκύπτει η ανάγκη για αποθήκευση και ανάκτηση πληροφορίας. Εν προκειμένω, όσον αφορά στην αναζήτηση εγγράφων στο σκληρό δίσκο ενός υπολογιστή, η μέχρι στιγμής υπάρχουσα κλασσική αναζήτηση με λέξεις κλειδιά εμφανίζει διαφόρων ειδών ατέλειες. Για παράδειγμα, ένα έγγραφο μπορεί να αναφέρεται σε μία έννοια, αλλά να μην περιέχει (αρκετές) λέξεις κλειδιά που να περιγράφουν τη συγκεκριμένη έννοια ή και το αντίστροφο, να περιέχει πολλές λέξεις κλειδιά μίας έννοιας στην οποίο όμως δεν αναφέρεται. Επιπλέον, μπορεί διαφορετικά σημεία του εγγράφου να αναφέρονται σε διαφορετικές έννοιες.

Με τη χρήση οντολογιών και οντολογικών γλωσσών (OWL, RDFS), δίνεται η δυνατότητα

σημασιολογικής επισημείωσης της έννοιας (ή και των εννοιών) κάθε εγγράφου (ή μέρους ενός εγγράφου), έτσι ώστε η αναζήτηση με τη βοήθεια της οντολογίας να εξειδικεύεται και να γίνεται πιο έγκυρη και αποτελεσματική. Παρακάτω περιγράφεται η έννοια της οντολογίας και αναλύεται η γλώσσα OWL (Web Ontology Language).

Οντολογία

Οντολογία είναι η περιγραφή, με τη χρησιμοποίηση ενός συγκεκριμένου λεξιλογίου, ενός συνόλου από έννοιες, αντικείμενα και σχέσεις μεταξύ τους που αφορούν μία συγκεκριμένη περιοχή γνώσης. Ουσιαστικά, μία οντολογία είναι μία ιεραρχία από κλάσεις, ιδιότητες και στιγμιότυπα των κλάσεων, που περιγράφουν ένα γνωστικό αντικείμενο.

RDF

Η RDF[1] είναι μία γλώσσα αναπαράστασης πληροφορίας για την περιγραφή διαδικτυακών οντοτήτων/πόρων. Χρησιμεύει στην οργάνωση της πληροφορίας σε συγκεκριμένο μορφότυπο, επεξεργάσιμο από εφαρμογές, παρέχοντας έτσι ένα κοινό πλαίσιο για την ανταλλαγή/διαμοιρασμό πληροφορίας μεταξύ εφαρμογών, χωρίς απώλεια της σημασιολογίας των δεδομένων. Η RDF βασίζεται στην ιδέα της αντιπροσώπευσης οντοτήτων μέσω καθολικών αναγνωριστικών πόρων (Internationalized Resource Identifiers - IRIs) και της περιγραφής οντοτήτων μέσω απλών ιδιοτήτων και τιμών που παίρνουν οι ιδιότητες. Αυτό οδηγεί στη δημιουργία απλών δηλώσεων (statements) για οντότητες σε μορφή ενός γράφου από κόμβους και ακμές που αντιπροσωπεύουν οντότητες ή τιμές και ιδιότητες αντίστοιχα. Συγκεκριμένα, οι δηλώσεις είναι σε μορφή τριπλετών (υποκείμενο, κατηγόρημα, αντικείμενο - subject, predicate, object) αποτελούμενες από την οντότητα (υποκείμενο) που περιγράφεται, την ιδιότητα (κατηγόρημα) που το περιγράφει και την τιμή που παίρνει (αντικείμενο). Το υποκείμενο και το κατηγόρημα επιτρέπεται να αντιπροσωπεύονται μόνο από καθολικά αναγνωριστικά, ενώ το αντικείμενο μπορεί να είναι τόσο αναγνωριστικό, όσο και μία σταθερή τιμή RDF (RDF Literal), όπως μία συμβολοσειρά, ένας ακέραιος αριθμός, μία ημερομηνία, κ.α. Μία συλλογή από RDF δηλώσεις (ή τριπλέτες) μπορεί να θεωρηθεί ως ένας κατευθυνόμενος, ονοματισμένος γράφος (directed labelled graph).

RDFS

Η RDF παρέχει έναν τρόπο έκφρασης απλών δηλώσεων για οντότητες, χρησιμοποιώντας ονοματισμένες ιδιότητες και τιμές. Παρόλα αυτά, είναι αναγκαία η ύπαρξη ενός λεξιλογίου (vocabulary) το οποίο θα χρησιμοποιείται στις δηλώσεις, προκειμένου να περιγράφει τις ιδιότητες και τις έννοιες/κλάσεις που χαρακτηρίζουν τις περιγραφόμενες οντότητες. Αυτό επιτυγχάνεται με το RDF σχήμα (RDF Schema - RDFS[2]), το οποίο επεκτείνει το λεξιλόγιο της RDF με σκοπό να περιγράψει, με τη βοήθεια δεσμευμένων λέξεων, οντότητες και σχέσεις μεταξύ ιδιοτήτων. Παρέχει δομές για την περιγραφή του τύπου των αντικειμένων (κλάσεις

[1] http://www.w3.org/TR/rdf-primer/
[2] http://www.w3.org/TR/rdf-schema/

-classes), ιεραρχίες τύπων (υποκλάσεις subclasses), ιδιότητες (properties) που περιγράφουν χαρακτηριστικά των οντοτήτων και ιεραρχίες ιδιοτήτων. Το λεξιλόγιο που χηρσιμοποιείται από το RDFS είναι ένα εξειδικευμένο σύνολο από προκαθορισμένους πόρους RDF, με ειδική σημασιολογία. Μία κλάση στο RDFS αντιστοιχεί στην γενικευμένη έννοια του τύπου ή της κατηγορίας και ορίζεται με τη δομή rdfs:Class. Οντότητες που ανήκουν σε μία κλάση ονομάζονται στιγμιότυπα της κλάσης. Ένα στιγμιότυπο μίας κλάσης συνδέεται με την κλάση μέσω της ιδιότητας rdf:type. Μία οντότητα μπορεί να είναι στιγμιότυπο πολλών κλάσεων. Μία ιδιότητα χαρακτηρίζει στιγμιότυπα μίας κλάσης ή ενός συνόλου κλάσεων και ορίζεται μέσω της δομής rdf:Property. Κάθε ιδιότητα έχει ένα πεδίο ορισμού, δηλαδή το σύνολο των κλάσεων τις οποίες χαρακτηρίζει και πεδίο τιμών, δηλαδή τις τιμές που μπορεί να πάρει.

OWL

Η γλώσσα οντολογιών ιστού (Web Ontology Language - OWL[3]) είναι η καθιερωμένη γλώσσα ορισμού και υλοποίησης οντολογιών στον ιστό. Η OWL και το RDFS έχουν αρκετές ομοιότητες. Η OWL ορίζεται ως ένα εξειδικευμένο λεξιλόγιο, όπως και το RDFS, έχει, όμως, πλουσιότερη σημασιολογία. Μία κλάση ορίζεται χρησιμοποιώντας τη δομή owl:Class και αντιπροσωπεύει ένα σύνολο από οντότητες με παρόμοιες ιδιότητες/χαρακτηριστικά. Η OWL προσφέρει επιπλέον δομές για ορισμό κλάσεων, συμπεριλαμβανομένων βασικών τελεστών συνόλων, όπως ένωση, τομή, συμπλήρωμα, καθώς και άλλες δομές, όπως ισοδυναμία κλάσης (owl:equivalentClass). Όσον αφορά τις οντότητες, η OWL επιτρέπει τον ορισμό ισοδυναμίας ή διαφορετικότητας δύο οντοτήτων μέσω των δομών owl:sameAs και owl:differentFrom. Επίσης διαχωρίζει ιδιότητες των οποίων η τιμή είναι μία οντότητα (owl:ObjectProperty) με αυτές οι οποίες παίρνουν σταθερές τιμές (owl:DatatypeProperty). Επίσης, ορίζεται ισοδυναμία μεταξύ ιδιοτήτων.

SPARQL 1.0

Το πρωτόκολλο και γλώσσα ερωτήσεων SPARQL[4] αποτελεί σύσταση της ομάδας W3C[5] και έχει καθιερωθεί ως η στάνταρ γλώσσα ερωτήσεων για RDF δεδομένα. Η αποτίμηση ερωτήσεων σε SPARQL βασίζεται στο ταίριασμα μοτίβων γράφων (graph pattern matching). Ένα μοτίβο γράφου ορίζεται αναδρομικά και περιέχει εκφράσεις RDF τριπλετών και τελεστές που τις συνδέουν. Οι τελεστές αυτοί είναι: σύζευξης (AND), διάζευξης (UNION), προαιρετικότητας (OPTIONAL) και φιλτραρίσματος (FILTER). Τα μοτίβα τριπλετών - εκφράσεις έχουν ακριβώς την ίδια μορφή με τις RDF τριπλέτες, με τη διαφορά ότι ένας τουλάχιστον από τους τρεις όρους είναι μεταβλητή. Η δομή ενός SPARQL ερωτήματος προσιδιάζει αρκετά τη δομή της SQL. Επιπλέον, η SPARQL παρέχει διάφορους τελεστές αλλαγής της σειράς των αποτελεσμάτων, όπως: Distinct, Reduced, Limit, Offset, Order By.

[3]http://www.w3.org/TR/owl-features/
[4]http://www.w3.org/TR/rdf-sparql-query/
[5]http://www.w3.org/

SPARQL 1.1

Η SPARQL 1.1[6] αποτελεί επέκταση της SPARQL, προτεινόμενη επίσης από την ομάδα W3C. Περιλαμβάνει τα ακόλουθα δομικά στοιχεία: Query, Update, Protocol, Service Description, Property Paths, Entailment Regimes, Uniform HTTP Protocol for Managing RDF Graphs και Federation Extensions. Η SPARQL 1.1 εξαλείφει αρκετούς από τους περιορισμούς της SPARQL 1.0, εισάγοντας λειτουργικότητα που αφορά συναθροιστικές συναρτήσεις, εμφωλευμένα ερωτήματα, λειτουργίες ανανέωσης δεδομένων, χειρισμό μονοπατιών, άρνησης κ.α. Γενικά, η SPARQL 1.1 επεκτείνεται, από γλώσσα ανάκτησης δεδομένων, σε γλώσσσα χειρισμού δεδομένων. Για παράδειγμα, εκτός από λειτουργίες ανανέωσης δεδομένων, υποστηρίζει συναρτήσεις εισαγωγής και διαγραφής, τόσο ατομικών όσο και μαζικών δεδομένων.

2.1.4 Τεχνικές διαφοροποίησης

Το πρόβλημα της διαφοροποίησης είναι στενά συνδεδεμένο με το πρόβλημα p-διασποράς (p-dispersion) [90]:

Ορισμός 2.1 (Πρόβλημα p-διασποράς). *Επίλεξε, από ένα σύνολο n δεδομένων σημείων, p σημεία έτσι ώστε η ελάχιστη απόσταση, ανά ζεύγη, των επιλεγμένων σημείων να μεγιστοποιείται.*

Το πρόβλημα έχει διάφορες παραλλαγές και ονόματα [92], όπως *facility dispersion*, *p-defense*, *maxsum/min dispersion*. Επίσης, είναι NP-complete [90], οπότε διάφοροι ευριστικοί αλγόριθμοι έχουν προταθεί, προκειμένου να λύνουν το πρόβλημα όσο το δυνατόν πιο αποδοτικά [91]. Στη συνέχεια, ανατρέχουμε εν συντομία σε μία κατηγοριοποίηση αυτών των αλγορίθμων, όπως παρουσιάζεται διεξοδικά στο [91].

2.1.5 Κατηγοριοποίηση αλγορίθμων διασποράς

Οι αλγόριθμοι p-διασποράς μπορούν να χωριστούν σε τέσσερις κατηγορίες: *κατασκευαστικοί* (*construction*), *γειτονιάς* (*neighborhood*), *προβολής* (*projection*) και *εναλλαγής* (*interchange*). Για τα ακόλουθα, θεωρούμε N το σύνολο των υποψήφιων σημείων, με $|N| = n$ και S το σύνολο των διαφοροποιημένων σημείων, με $|S| = p$, όταν ο αλγόριθμος διασποράς/διαφοροποίησης έχει ολοκληρωθεί.

Κατασκευαστικοί αλγόριθμοι

Χωρίζονται σε τρεις υποκατηγορίες:

- **Άπληστες κατασκευαστικές ευριστικές.** Οι συγκεκριμένοι αλγόριθμοι επιλέγουν, αρχικά, το ζεύγος σημείων που έχουν τη μεγαλύτερη μεταξύ τους απόσταση, για να αρχικοποιήσει το σύνολο των διαφοροποιημένων σημείων S. Στη συνέχεια, και μέχρι να ισχύσει ότι $|S| = p$, σε κάθε επανάληψη, το στοιχείο που εισάγεται είναι το σημείο που μεγιστοποιεί την ελάχιστη απόσταση από τα ήδη εισηγμένα στο S σημεία.

[6] http://www.w3.org/TR/sparql11-query/

- **Άπληστες ευριστικές διαγραφής.** Οι συγκεκριμένοι αλγόριθμοι αρχικά θεωρούν όλα τα υποψήφια σημεία ως στοιχεία του τελικού συνόλου, δηλαδή $S = N$ και μετά, σε κάθε βήμα, αφαιρούν ένα από τα δύο σημεία τα οποία έχουν τη μικρότερη μεταξύ τους απόσταση και, συγκεκριμένα, το σημείο που απέχει, συγκεντρωτικά, λιγότερο από τα υπόλοιπα σημεία του S.

- **Ημι-άπληστες ευριστικές διαγραφής** Οι συγκεκριμένοι αλγόριθμοι είναι ίδιοι με τις *Άπληστες ευριστικές διαγραφής*, απλά η διαγραφή του ενός από τα δύο επιλεγμένα προς διαγραφή στοιχεία (από το ζεύγος με τη μικρότερη απόσταση) γίνεται με τυχαίο τρόπο.

Αλγόριθμοι γειτονιάς

Οι συγκεκριμένοι αλγόριθμοι διαλέγουν και εισάγουν ένα αρχικό σημείο στο S και ορίζουν μία γειτονιά γύρω από αυτό, για παράδειγμα έναν κύκλο με κέντρο το επιλεγμένο σημείο x_i. Στη συνέχεια, επιλέγουν το επόμενο στοιχείο προς εισαγωγή, αποκλείοντας όλα τα σημεία στη γειτονιά του x_i. Σε κάθε βήμα, η γειτονιά ορίζεται με βάση το τελευταίο εισηγμένο στοιχείο. Οι αλγόριθμοι διαφοροποιούνται ανάλογα με την ευριστική που χρησιμοποιείται για να επιλεγεί το επόμενο σημείο x_i' (που βρίσκεται έξω από τη γειτονιά του x_i): κάποιος μπορεί να επιλέξει το *πρώτο* σημείο, το *πιο κοντινό* ή το *πιο μακρυνό* σημείο από τα υποψήφια.

Αλγόριθμοι εναλλαγής

Οι συγκεκριμένοι αλγόριθμοι θεωρούν μία αρχική τυχαία λύση S και μετά, σε κάθε βήμα, εναλλάσσουν ένα σημείο $x \in S$ με ένα σημείο $y \notin S$, στοχεύοντας να βελτιώσουν τη διασπορά του S. Όπως και στην περίπτωση των αλγορίθμων *άπληστης διαγραφής*, το σημείο προς διαγραφή x προέρχεται από το ζεύγος σημείων του S με την μικρότερη μεταξύ τους απόσταση. Το σημείο y προς εισαγωγή, μπορεί να είναι το πρώτο που αυξάνει τη διασπορά ή αυτό που την αυξάνει περισσότερο. Μία σημαντική παραλλαγή αυτών των αλγορίθμων είναι οι αλγόριθμοι προσομοιωμένης ανόπτησης (simulated annealing), οι οποίοι στοχεύουν στην αποφυγή τοπικών μεγίστων, εναλλάσσοντας, περιοδικά, στοιχεία με τρόπο που να μειώνει τη διασπορά του S.

Αλγόριθμοι προβολής

Οι συγκεκριμένοι αλγόριθμοι προβάλλουν το πρόβλημα σε μία γραμμή και το λύνουν εκεί.

Η αξιολόγηση που περιγράφεται στο [91] δείχνει ότι, γενικά, οι αλγόριθμοι εναλλαγής και γειτονιάς είναι λίγο πιο ακριβείς από τους κατασκευαστικούς αλγορίθμους, ενώ οι αλγόριθμοι προβολής έχουν τη χειρότερη απόδοση. Παρόλα αυτά, οι αλγόριθμοι γειτονιάς χρειάζονται παραμετροποίηση της ακτίνας της γειτονιάς και, ανάλογα με αυτήν την παραμετροποίηση, μπορεί να μην καταλήξουν σε λύση. Επίσης, χρειάζεται να τρέξουν αρκετές φορές με διαφορετικά σημεία αρχικοποίησης, ώστε να μην επηρεάζεται η τελική λύση από την αρχικοποίηση. Επιπλέον, τόσο οι αλγόριθμοι γειτονιάς, όσο και οι εναλλαγής, γίνονται πιο απαιτητικοί σε χρόνο για μεγάλο αριθμό υποψήφιων σημείων N. Από την άλλη πλευρά, οι κατασκευαστικοί

αλγόριθμοι είναι εύκολοι στην υλοποίηση, δε χρειάζονται παραμετροποίηση και έχουν σχετικά καλή αποτελεσματικότητα, όντας αποδοτικοί για μεγάλα μεγέθη του N και μικρά κλάσματα $|S|/|N|$. Για αυτό το λόγο, μεγάλος αριθμός πρόσφατων εργασιών σε διαφοροποίηση αποτελεσμάτων υιοθετεί παραλλαγές των άπληστων κατασκευαστικών αλγορίθμων.

2.1.6 Συναρτήσεις -στόχοι και αποστάσεις

Στη συνέχεια παρουσιάζουμε υπάρχουσες εργασίες στην περιοχή της διαφοροποίησης αποτελεσμάτων αναζήτησης, από τις οποίες εμπνεύστηκε μέρος της εργασίας μας πάνω στη διαφοροποίηση σχολίων χρηστών.

Οι συγγραφείς του [95] βασίζουν την ανάλυση τους στην έννοια της πληροφοριακής ανάγκης (information need) u, η οποία σχετίζεται με ένα ερώτημα q και τις αντίστοιχες *μονάδες πληροφορίας* information nuggets. Οι μονάδες πληροφορίας είναι διαφορετικές εκφάνσεις της πληροφοριακής ανάγκης που σχετίζεται με ένα ερώτημα, δηλαδή διαφορετικές όψεις ή ερωτήσεις που απαντώνται από τα αποτελέσματα ενός ερωτήματος. Για παράδειγμα, για το ερώτημα «jaguar», δύο προφανείς μονάδες πληροφορίας είναι το αυτοκίνητο jaguar και το ζώο jaguar. Για το ερώτημα «εθνικές εκλογές», πιθανές μονάδες πληροφορίας θα απαντούσαν στις παρακάτω ερωτήσεις: «πότε γίνονται οι επόμενες εθνικές εκλογές;», «ποιοι είναι οι υποψήφιοι αρχηγοί κομμάτων για τις επόμενες εθνικές εκλογές;» ή «ποια είναι τα εκτιμώμενα αποτελέσματα για τις επόμενες εκλογές;».

Βασιζόμενοι στα παραπάνω, ορίζουν την πιθανότητα ενός εγγράφου d να ικανοποιεί την ανάγκη αναζήτησης u ενός ερωτήματος ως:

$$P(R = 1|u, d) = 1 - \prod_{i=1}^{m} (1 - P(n_i \in u) \cdot P(n_i \in d)) \tag{2.4}$$

όπου η ποσότητα $P(R = 1|u, d)$ δηλώνει ότι υπάρχει τουλάχιστον μία μονάδα πληροφορίας n_i που ικανοποιεί την πληροφοριακή ανάγκη u και καλύπτεται από το έγγραφο d, και το m δηλώνει το συνολικό αριθμό μονάδων πληροφορίας που σχετίζονται με το u. Το γινόμενο στον παραπάνω τύπο δηλώνει την πιθανότητα να μην υπάρχει καμία μονάδα πληροφορίας που να συνδέει την πληροφοριακή ανάγκη u με το έγγραφο d.

Ο παραπάνω τύπος εξετάζει μόνο ένα έγγραφο, όσον αφορά τη σχέση του με την πληροφοριακή ανάγκη. Οπότε θα μπορούσε να χρησιμοποιηθεί ως συνάρτηση στόχος προς βελτιστοποίηση, όταν θέλουμε να επιλέξουμε το πρώτο στοιχείο από ένα σύνολο υποψήφιων στοιχείων προς εισαγωγή στο σύνολο S. Στη γενική περίπτωση, όπου έχουν εισαχθεί ήδη k στοιχεία στο σύνολο S, η Εξίσωση 2.4 μετατρέπεται σε:

$$P(R_k = 1|u, d_1, d_2, ..., d_k) = 1 - \prod_{i=1}^{m} (1 - P(n_i \in u) \cdot (\prod_{j=1}^{k-1} P(n_i \notin d_j)) \cdot P(n_i \in d)) \tag{2.5}$$

όπου η ποσότητα $\prod_{j=1}^{k-1} P(n_i \notin d_j)$ δηλώνει την πιθανότητα ότι κανένα από τα ήδη εισηγμένα στοιχεία δεν ικανοποιεί τη μονάδα πληροφορίας n_i. Το $P(R_k = 1|u, d_1, d_2, ..., d_k)$, τελικά,

δίνει την πιθανότητα ένα στοιχείο k να περιέχει μονάδες πληροφορίας που δεν περιέχονται σε προηγουμένως εισηγμένα στο S στοιχεία. Δεδομένων των παραπάνω, ο στόχος ενός ευριστικού αλγορίθμου διαφοροποίησης είναι, σε κάθε βήμα, να μεγιστοποιεί την πιθανότητα που περιγράφεται στην Εξίσωση 2.5, έως ότου το σύνολο S να περιέχει τον επιθυμητό αριθμό στοιχείων.

Οι συγγραφείς του [97] επεκτείνουν την παραπάνω ανάλυση θεωρώντας κατηγορίες στις οποίες ερωτήματα και έγγραφα (αποτελέσματα) μπορούν να ανήκουν. Όμοια με το [95], ορίζουν το $V(d|q,c)$ ως την ποιοτική τιμή του εγγράφου d, για το ερώτημα q, όσον αφορά την κατηγορία c. Ο στόχος είναι η εύρεση ενός συνόλου από k διαφοροποιημένα αποτελέσματα, που να μεγιστοποιούν την παρακάτω ποσότητα:

$$P(S|q) = \sum_{c} P(c|q)(1 - \prod_{d \in S}(1 - V(d|q,c))) \tag{2.6}$$

όπου $P(c|q)$ είναι η κατανομή πιθανότητας των κατηγοριών για το ερώτημα q.

Η κύρια διαφορά μεταξύ αυτής της μεθόδου και της αντίστοιχης του [95] είναι ότι εδώ λαμβάνουν υπόψη και τη σχετική σημασία ανάμεσα σε διαφορετικές μονάδες πληροφορίας, καθώς επίσης και ότι έγγραφα που περιέχουν την ίδια μονάδα πληροφορίας μπορεί να καλύπτουν, σε διαφορετικό βαθμό, μία πληροφοριακή ανάγκη. Για παράδειγμα, αν ένα έγγραφο d αναφέρεται επιφανειακά στη μονάδα πληροφορίας n_i, τότε το $V(d|q,c_i)$ αναμένεται να είναι χαμηλό, οπότε επιπλέον έγγραφα σχετικά με τη μονάδα n_i πρέπει να εισαχθούν σε επόμενα βήματα του αλγορίθμου, ώστε η πληροφοριακή ανάγκη να καλυφθεί πλήρως.

Ο υιοθετούμενος αλγόριθμος είναι ένας άπληστος κατασκευαστικός αλγόριθμος κατά τον οποίο, σε κάθε βήμα, το έγγραφο με τη μεγαλύτερη *οριακή χρησιμότητα* (marginal utility) εισάγεται στο S. Η οριακή χρησιμότητα ορίζεται ως:

$$g(d|q,c,S) = \sum_{c \in C(d)} U(c|q,S)V(d|q,c) \tag{2.7}$$

όπου $C(d)$ είναι το σύνολο των κατηγοριών που σχετίζονται με το έγγραφο d και $U(c|q,S)$ είναι η δεσμευμένη πιθανότητα το ερώτημα q να ανήκει στην κατηγορία c, δεδομένου ότι όλα τα ήδη εισηγμένα στο S έγγραφα αποτυγχάνουν να καλύψουν την πληροφοριακή ανάγκη. Ουσιαστικά, ο αλγόριθμος, σε κάθε βήμα, ψάχνει να βρει το έγγραφο με τη μεγαλύτερη συνεισφορά στο να ικανοποιήσει οποιεσδήποτε από τις κατηγορίες του, δεδομένης της υπάρχουσας συνεισφοράς των ήδη εισηγμένων στο S εγγράφων.

Οι συγγραφείς του [96] εισάγουν την έννοια της κειμενικής ομοιότητας στις συναρτήσεις στόχους τους. Δηλαδή, στόχος είναι πλέον η βελτιστοποίηση μίας συνάρτησης που συναθροίζει την ομοιότητα του εγγράφου με το ερώτημα και τη διασπορά μεταξύ των εγγράφων. Στη συγκεκριμένη δουλειά, τρεις συναρτήσεις στόχοι ορίζονται: *Max-Sum*, *Max-Min* και *Mono-objective*. Ακολουθούν οι τύποι που τις περιγράφουν:

- **Max-Sum**

$$f(S) = (k-1)\sum_{u \epsilon S} w(u) + 2\lambda \sum_{u,v \epsilon S} d(u,v) \tag{2.8}$$

όπου $|S| = k$ το πλήθος των επιθυμητών διαφοροποιημένων εγγράφων, $w(u)$ είναι το σκορ ομοιότητας του εγγράφου u με το ερώτημα, $d(u, v)$ είναι το σκορ διαφοροποίησης (απόσταση) μεταξύ των εγγράφων u και v και $\lambda > 0$ είναι η παράμετρος που ρυθμίζει τη σημασία του σκορ διαφοροποίησης σε σχέση με το σκορ ομοιότητας.

- **Max-Min**

$$f(S) = \min_{u \epsilon S} w(u) + \lambda \min_{u,v \epsilon S} d(u, v) \tag{2.9}$$

- **Mono-objective**

$$f(S) = \sum_{u \epsilon S} w(u)' \tag{2.10}$$

όπου $w(u)' = (w(u) + \frac{\lambda}{|N|-1} \sum_{v \epsilon N} d(u, v))$.

Τρεις προσεγγιστικοί αλγόριθμοι για τις παραπάνω συναρτήσεις στόχους ορίζονται επίσης στη συγκεκριμένη δουλειά. Οι συγκεκριμένοι αλγόριθμοι υιοθετούνται και στη δική μας εργασία, οπότε και παρουσιάζονται αναλυτικά στο αντίστοιχο κεφάλαιο που περιγράφει τη διαφοροποίηση σχολίων χρηστών σε ειδησεογραφικά άρθρα.

2.2 Σχετικές εργασίες

2.2.1 Δεδομένα εκπαίδευσης

Προηγούμενες εργασίες σχετικά με το πρόβλημα της αύξησης της ποσότητας/ποιότητας των δεδομένων εκπαίδευσης βασίζονται στην προαγωγή σε υψηλότερες θέσεις κατάταξης αποτελεσμάτων από χαμηλή θέση. Στο [13] η αρχική κατάταξη (ταξινόμηση) που παρουσιάζεται στους χρήστες παραλλάσσεται όσον αφορά τα 2 πρώτα αποτελέσματα. Συγκεκριμένα, τα αποτελέσματα των θέσεων 1 και 2, τα οποία αντικαθιστούν τα αρχικά αποτελέσματα, επιλέγονται με βάση διάφορες στρατηγικές (τυχαία επιλογή, ζεύγος αποτελεσμάτων που μεγιστοποιεί μία συγκεκριμένη μετρική απώλειας, κ.α.). Στο [14], στόχος είναι η προαγωγή νέων σελίδων, οι οποίες αδικούνται από αλγόριθμους που υπολογίζουν τη δημοφιλία σελίδων στο σκορ ταξινόμησης. Έτσι, αποτελέσματα από χαμηλές θέσεις κατάταξης επιλέγονται τυχαία και τοποθετούνται σε υψηλότερες θέσεις. Αυτή η διαδικασία επιτρέπει σε διαφορετικά (νέα) έγγραφα να κριθούν για τη σχετικότητά τους από τους χρήστες.

Κατ΄ αντιστοιχία με τη μέθοδό μας, και οι δύο παραπάνω προσεγγίσεις στοχεύουν στην εκμετάλλευση επιπλέον ανάδρασης από το χρήστη για την εκπαίδευση μίας συνάρτησης ταξινόμησης. Παρόλα αυτά, οι συγκεκριμένες μέθοδοι αλλοιώνουν την αρχική κατάταξη των αποτελεσμάτων που εμφανίζεται στο χρήστη κατά τη διάρκεια της φάσης εκπαίδευσης, μεταφέροντας αποτελέσματα χαμηλής κατάταξης σε υψηλότερες θέσεις. Σε αντίθεση με αυτές, η μέθοδός μας παράγει νέες κρίσεις σχετικότητας για αποτελέσματα που δεν έχουν επισκοπηθεί από το χρήστη (αποτελέσματα χαμηλής κατάταξης), χωρίς να μεταβάλλει την αρχική κατάταξη των αποτελεσμάτων.

2.2.2 Ποιότητα εκπαίδευσης

Στις ακόλουθες παραγράφους παρουσιάζονται διάφορες προσεγγίσεις που σχετίζονται με (α) εξατομικευμένη αναζήτηση στον ιστό και (β) εκπαίδευση συναρτήσεων ταξινόμησης/αναταξινόμησης αποτελεσμάτων.

Στο [15] ο συγγραφέας προτείνει μία, βασιζόμενη στη θεματική περιοχή, βελτίωση του αλγορίθμου pagerank, που επιτρέπει τον προεπεξεργασμένο υπολογισμό ενός *σταθερού* αριθμού από διανύσματα pagerank, τα οποία θα αντιστοιχούν σε συγκεκριμένες θεματικές κατηγορίες. Αυτά τα διανύσματα, στη συνέχεια, χρησιμοποιούνται για να μεταβάλλουν τον υπολογισμό της λίστας αποτελεσμάτων κάθε ερωτήματος, βασιζόμενα στην ομοιότητα του ερωτήματος με κάθε θεματική κατηγορία. Στο [16] αντιμετωπίζεται πάλι το πρόβλημα της εξατομίκευσης του αλγόριθμου pagerank, με έμφαση στην κλιμακωσιμότητα. Στο [17] οι συγγραφείς εκμεταλλεύονται ιεραρχίες εννοιών, όπως το ODP [28] , για να κατηγοριοποιήσουν ερωτήματα και να δημιουργήσουν προφίλ χρηστών. Έπειτα, χρησιμοποιούν τεχνικές συνεργατικού φιλτραρίσματος (collaborative filtering) για να ανακατατάξουν τα αποτελέσματα ερωτημάτων, βασιζόμενοι στα παραπάνω προφίλ. Συγκρινόμενη με τα παραπάνω, η μέθοδός μας διαφέρει στο ότι δεν κατασκευάζει διαφορετικά προφίλ χρηστών, ούτε θεωρεί προκαθορισμένο αριθμό θεματικών κατηγοριών. Αντιθέτως, εκμεταλλευόμαστε συνεργατική πληροφορία από διάφορους χρήστες για να δημιουργήσουμε ομάδες δεδομένων (με κοινό θεματικό περιεχόμενο ή με κοινή συμπεριφορά αναζήτησης), στις οποίες οι χρήστες ενδέχεται να συμμετέχουν με διαφορετικές βαρύτητες. Επιπλέον, κάποιες εκδοχές της μεθόδου μας δεν στηρίζονται μόνο στο περιεχόμενο, αλλά υπολογίζουν και ομοιότητες στη συμπεριφορά αναζήτησης.

Στα [5, 6, 18, 19], οι συγγραφείς δομούν τα μοντέλα τους εκμεταλλευόμενοι το βραχυπρόθεσμο ή μακροπρόθεσμο ιστορικό αναζήτησης ή το περιβάλλον (context) (για παράδειγμα τα έγγραφα στον υπολογιστή) του χρήστη. Αυτά τα μοντέλα είναι, ουσιαστικά, προφίλ χρηστών που χρησιμοποιούνται για να επεκτείνουν μελλοντικά ερωτήματα ή να βελτιώσουν τα αποτελέσματά τους. Συγκρινόμενες με τη μέθοδό μας, οι παραπάνω προσεγγίσεις θεωρούν μόνο ομοιότητα περιεχομένου και, επιπλέον, δεν εκμεταλλεύονται συνεργατική πληροφορία από όλους τους χρήστες.

Υπάρχουν επίσης δουλειές που τροποποιούν διαδεδομένες τεχνικές μηχανικής μάθησης για να βελτιώσουν την κατάταξη των αποτελεσμάτων. Στο [12], ο κλασικός αλγόριθμος Μηχανών Διανυσμάτων Στήριξης για ταξινόμηση τροποποιείται έτσι ώστε (α) να μειώνονται τα σφάλματα ταξινόμησης στις υψηλότερες κλάσεις σχετικότητας (β) να αυξάνεται η βαρύτητα ερωτημάτων με λιγότερες διαθέσιμες κρίσεις σχετικότητας, όσον αφορά τη διαδικασία της εκπαίδευσης. Στο [20] οι συγγραφείς προτείνουν την εκπαίδευση πολλαπλών συναρτήσεων ταξινόμησης, που θα αντιστοιχούν, όμως, στις διαφορετικές κλάσεις σχετικότητας των δεδομένων εκπαίδευσης και όχι σε διαφορετικές θεματικές περιοχές ή διαφορετικές συμπεριφορές αναζήτησης. Συγκρινόμενες με τη δουλειά μας, οι παραπάνω εργασίες δεν λαμβάνουν υπόψη τους τις ενδογενείς σχέσεις μεταξύ ερωτημάτων και αποτελεσμάτων, που καθορίζουν διαφορετικές συμπεριφορές αναζήτησης.

Τέλος, μέθοδοι συσταδοποίσης συνδυάζονται με τεχνικές Μηχανών Διανυσμάτων Στήρι-

ξης, αλλά για να επιτύχουν διαφορετικούς σκοπούς από τους δικούς μας. Στο [21] εφαρμόζεται συσταδοποίηση με σκοπό την εκλέπτυνση μεγάλων σετ δεδομένων εκπαίδευσης, απορρίπτοντας εκείνα τα δεδομένα που δεν χρειάζονται στη φάση εκπαίδευσης του μοντέλου. Επίσης, στο [22] εκτελείται συσταδοποίηση σε συναρτήσεις ταξινόμησης, εκπαιδευμένες ανά χρήστη, για σκοπούς παραγωγής προτάσεων (recommendation).

2.2.3 Σημασιολογική επισημείωση

Ένας μεγάλος αριθμός προσεγγίσεων όσον αφορά τη σημασιολογική επισημείωση έχει προταθεί στη βιβλιογραφία [53, 54]. Οι περισσότερες επικεντρώνονται στην επισημείωση διαδικτυακών πόρων (web resources), όπως HTML ιστοσελίδες [55, 56, 57, 58, 59, 60, 61]. Στην επισημείωση απλού κειμένου, υπάρχουν προσεγγίσεις που διαφέρουν όσον αφορά τόσο στην επισημείωση, όσο και στην αναζήτηση κειμένου. Το GATE [62] είναι μία πλατφόρμα που ενσωματώνει συγκεκριμένη αρχιτεκτονική, πλαίσιο και γραφικό εργαλείο για γλωσσική επεξεργασία. Ταυτόχρονα προσφέρει εργαλεία και πόρους για κειμενική επισημείωση, τόσο χειροκίνητη όσο και αυτόματη, χρησιμοποιώντας τεχνολογίες εξαγωγής πληροφορίας (Information Extraction - IE). Η δουλειά που περιγράφεται στο [63] προσφέρει μία υποδομή για σημασιολογική επισημείωση κειμένων, περιοριζόμενη, όμως, από τη δική της οντολογία, ονομαζόμενη KIMO. Τα υποσυστήματα για εξαγωγή πληροφορίας, διαχείριση εγγράφων και επισημείωση βασίζονται στο GATE. Σκοπός του υποσυστήματος εξαγωγής πληροφορίας είναι η αναγνώριση ονοματικών οντοτήτων που να σχετίζονται με τις έννοιες της οντολογίας KIMO. Συγκρινόμενη με τις παραπάνω προσεγγίσεις, η δουλειά που περιγράφεται στην παρούσα διατριβή υλοποιεί προχωρημένες δυνατότητες αναζήτησης, συνδυάζοντας κλασσική αναζήτηση με λέξεις κλειδιά και σημασιολογική αναζήτηση με περιήγηση στις κλάσεις της οντολογίας. Επιπλέον, εισάγει δυνατότητες αυτόματης επισημείωσης κειμένου, στηριζόμενες σε μοντέλα μηχανικής μάθησης που εκπαιδεύονται με βάση το ιστορικό επισημείωσης των χρηστών.

Η δουλειά του [64] (AKTiveMedia) υποστηρίζει επισημείωση κειμένου, εικόνων και ιστοσελίδων, χρησιμοποιώντας τόσο οντολογίες, όσο και επισημειώσεις ελεύθερου κειμένου (tags). Για την υποστήριξη αυτόματης επισημείωσης χρησιμοποιείται ένα υποκείμενο σύστημα εξαγωγής πληροφορίας, το οποίο εκπαιδεύεται από προηγούμενες επισημειώσεις για να προτείνει επισημειώσεις στο χρήστη. Παρόλα αυτά, η συγκεκριμένη δουλειά δεν υποστηρίζει δυνατότητες αναζήτησης, ενώ ο μηχανισμός αυτόματης επισημείωσης έχει διάφορους περιορισμούς: έχει χαμηλή απόδοση, όταν η επισημείωση επεκτείνεται πέραν του ενός όρου, ενώ δεν υποστηρίζεται επισημείωση με περισσότερες της μίας έννοιες.

Τα παραπάνω εργαλεία περιορίζονται σε επισημείωση απλού κειμένου ή HTML κειμένου. Όσον αφορά άλλους μορφότυπους κειμένου, το PDFTab [65] είναι ένα plug-in του Protege[7] που υποστηρίζει επισημείωση pdf εγγράφων με κλάσεις από οντολογίες OWL. Οι επισημειώσεις αποθηκεύονται εσωτερικά στο κάθε έγγραφο. Ομοίως, το SemanticWord [66] είναι ένα plug-in του MS Word που υποστηρίζει επισημείωση word εγγράφων με κλάσεις από οντο-

[7] http://protege.stanford.edu/

λογίες DAML+OIL. Συγκριτικά με τη δική μας προσέγγιση, τα δύο παραπάνω εργαλεία δεν υποστηρίζουν αναζήτηση ή αυτόματη επισημείωση.

Όσον αφορά στη σημασιολογική αναζήτηση, τα τελευταία χρόνια έχουν, επίσης, προταθεί αρκετές μέθοδοι στη βιβλιογραφία [67]. Μία προσέγγιση αρκετά κοντινή στη δική μας παρουσιάζεται στο [68], όπου ένας συνδυασμός αναζήτησης με λέξεις κλειδιά και σημασιολογικής αναζήτησης σε διαδικτυακούς πόρους υλοποιείται πάνω στο AKTiveMedia ([64]). Ένα σημαντικό μειονέκτημα αυτής της μεθόδου είναι ότι η ταξινόμηση αποτελεσμάτων στηρίζεται μόνο στην αναζήτηση με λέξεις κλειδιά, ενώ η σημασιολογική αναζήτηση χρησιμοποιείται μόνο για να φιλτραριστούν αποτελέσματα. Επιπλέον, δεν υποστηρίζονται προχωρημένες επιλογές σημασιολογικής αναζήτησης βασισμένες στη σημασιολογία των οντολογιών. Τέλος, μία ενδιαφέρουσα αλλά λιγότερο σχετική προσέγγιση παρουσιάζεται στο [69], όπου γίνεται ανάλυση της σημασίας των λέξεων και φράσεων, ώστε να οριστούν σημασιολογικές σχέσεις μεταξύ εννοιών. Έτσι, η αναζήτηση επεκτείνεται με σημασιολογία, μετατρέποντας τις απλές λέξεις σε έννοιες, ώστε να γίνει δυνατή η εκμετάλλευση των μεταξύ τους σημασιολογιών.

2.2.4 Εξατομίκευση σημασιολογικών δεδομένων

Οι περισσότερες υπάρχουσες, μέχρι στιγμής, προσεγγίσεις είτε αντιμετωπίζουν έμμεσα το πρόβλημα της εκπαίδευσης μοντέλων αναταξινόμησης για εξατομίκευση, είτε ακολουθούν προσεγγίσεις βασισμένες στη μνήμη (memory based approaches) για εξατομίκευση. Η δουλειά που περιγράφεται στη διατριβή είναι η πρώτη που πραγματεύεται την υιοθέτηση προσέγγισης βασισμένης σε μοντέλο (model based approach) για εξατομίκευση σημασιολογικής αναζήτησης.

Η πιο κοντινή δουλειά στη δική μας περιγράφεται στο [39], όπου οι συγγραφείς αναπτύσσουν τεχνικές εξατομίκευσης αναζήτησης σε σημασιολογικά δεδομένα, στο πλαίσιο της μηχανής αναζήτησης NAGA[8], ενσωματώνοντας τα ενδιαφέροντα του χρήστη στο μηχανισμό βαθμολόγησης αποτελεσμάτων της μηχανής. Το προφίλ του χρήστη κατασκευάζεται ως ένα απόσπασμα (snippet) του συνολικού γράφου γνώσης που περιέχει τα σημασιολογικά δεδομένα. Οι σχέσεις μεταξύ των οντοτήτων της βάσης (γράφου) γνώσης χρησιμοποιούνται για να διαδοθούν σκορ προτίμησης του χρήστη, από οντότητες που τον αφορούν, σε περαιτέρω οντότητες, υλοποιώντας, έτσι, ένα πιθανοτικό μοντέλο εξατομίκευσης αποτελεσμάτων αναζήτησης.

Στο [40], οι συγγραφείς χρησιμοποιούν τεχνικές διάδοσης ενεργοποίησης (spreading activation), συνδυαζόμενες με κλασσική αναζήτηση με σκοπό τη βελτίωση της αναζήτησης σε κάποια συγκεκριμένη θεματική περιοχή. Εφαρμόζουν την τεχνική τους χρησιμοποιώντας ένα υβριδικό δίκτυο οντοτήτων, όπου κάθε σχέση μεταξύ οντοτήτων χαρακτηρίζεται από μία επιγραφή (label) και ένα αριθμητικό βάρος (weight). Η τεχνική τους καθιστά δυνατή την ανάκτηση οντοτήτων οι οποίες δεν περιέχουν τις λέξεις κλειδιά του ερωτήματος, αλλά συνδέονται με άλλες οντότητες που τις περιέχουν. Στη δουλειά του [41], υιοθετείται μία στατιστική προσέγγιση για την εκμάθηση ενός οντολογικού μοντέλου χρήστη με βάση κάποια θεματική

[8] http://www.mpi-inf.mpg.de/yago-naga/naga/

οντολογία με τη βοήθεια διάδοσης ενεργοποίησης.

Στο [42], οι συγγραφείς πραγματεύονται το πρόβλημα πρότασης ερωτημάτων, διασυνδέοντας με αυτόματο τρόπο ερωτήματα με έννοιες από την οντολογία της DBpedia. Στα πλαίσια της μεθόδου, σχετικές έννοιες εντοπίζονται για το πλήρες ερώτημα, αλλά και για κάθε ν-γράμμα (n-gram) (δηλαδή μία ακολουθία από ν λέξεις) του ερωτήματος και, κατόπιν, τεχνικές μηχανικής μάθησης εφαρμόζονται για να αποφασιστεί ποιες από τις έννοιες θα κρατηθούν και ποιες θα απορριφθούν. Η συγκεκριμένη μέθοδος εκμεταλλεύεται την κειμενική περιγραφή κάθε έννοιας, καθώς επίσης χαρακτηριστικά βασισμένα στο ερώτημα και βασισμένα στις έννοιες για την εκπαίδευση των μοντέλων μηχανικής μάθησης. Η προσέγγιση που προτείνεται στο [43] χρησιμοποιεί τόσο εξαρτημένα, όσο και ανεξάρτητα από το σύνολο δεδομένων χαρακτηριστικά εκπαίδευσης. Τα πρώτα εξάγονται από τον γράφο RDF των δεδομένων. Τα δεύτερα εξάγονται από εξωτερικές πηγές, όπως διαδικτυακές μηχανές αναζήτησης ή βασεις δεδομένων ν-γραμμάτων. Τα χαρακτηριστικά, επίσης, κατηγοριοποιούνται σε (α) βασισμένα στην συχνότητα, τα οποία προκύπτουν μετρώντας διάφορα μοτίβα στον γράφο δεδομένων ή στις εξωτερικές πηγές και (β) βασισμένα στην κεντρικότητα, τα οποία προκύπτουν εφαρμόζοντας γραφοθεωρητικούς αλγορίθμους, όπως ο PageRank στο γράφο RDF δεδομένων.

Στο [44], ορίζεται μία οντολογία από θεματικές περιοχές που χρησιμοποιείται για την κατηγοριοποίηση ιστοσελίδων. Το προφίλ του χρήστη κωδικοποιείται ως εκδοχή μίας προϋπάρχουσας θεματικής οντολογίας αναφοράς, στην οποία εκδοχή οι έννοιες -κλάσεις χαρακτηρίζονται από σκορ ενδιαφέροντος, τα οποία εξάγονται και ανανεώνονται με βάση τη συμπεριφορά του χρήστη (δηλαδή ποιες έννοιες χρησιμοποίησε/είδε). Η ιεραρχική σχέση μεταξύ των εννοιών στην οντολογία χρησιμοποιείται σε συνδυασμό με τεχνικές διάδοσης ενεργοποίησης για να χαρακτηριστεί το σύνολο των εννοιών της οντολογίας -προφίλ του χρήστη. Τέλος, στο [45], ένα εκτεταμένο σύνολο από προτιμήσεις εννοιών εξάγεται για κάθε χρήστη, βασισμένο σε έννοιες που προκύπτουν από τα δεδομένα αναζήτησης του χρήστη. Έτσι, το προφίλ του χρήστη αναπαρίσταται, αρχικά, ως ένα οντολογικό δέντρο και, στη συνέχεια, δίνεται ως είσοδος σε μία μηχανή διανυσμάτων στήριξης για να εκπαιδεύσει ένα εξατομικευμένο μοντέλο προτιμήσεων χρήστη.

2.2.5 Ανάλυση σχολίων χρηστών και τεχνικές διαφοροποίησης

Από μελέτη της βιβλιογραφίας δεν προκύπτουν μέχρι και σήμερα εργασίες που να ασχολούνται με το πρόβλημα της διαφοροποίησης σχολίων χρηστών σε ειδησεογραφικά άρθρα, το οποίο πραγματεύεται η συγκεκριμένη διατριβή. Στη συνέχεια παραθέτουμε εργασίες που καταπιάνονται με (α) το πρόβλημα της ανάλυσης σχολίων χρηστών και, γενικά, δεδομένων κοινωνικών δικτύων και (β) με το πρόβλημα της διαφοροποίησης αποτελεσμάτων αναζήτησης.

Η δουλειά που περιγράφεται στο [88] είναι η κοντινότερη στη δική μας. Οι συγγραφείς παρουσιάζουν το (υπό ανάπτυξη) σύστημά τους που χειρίζεται διαδικτυακές ομάδες συζήτησης. Το σύστημα απαιτεί οι χρήστες να δηλώνουν ρητά τις απόψεις τους σε συγκεκριμένες θεματικές περιοχές. Στη συνέχεια, εκμεταλλεύεται αυτήν την ανάδραση χρηστών για να προτείνει (συστήσει) διάφορες απόψεις, επιτρέποντας στο χρήστη να ρυθμίσει το βαθμό της ομοιότητα-

ς/ετερογένειας των συστάσεων σε σχέση με τις δικές του απόψεις. Πέρα από τις διαφορές στα κριτήρια διαφοροποίησης που χρησιμοποιούνται, η συγκεκριμένη δουλειά διαφέρει από τη δική μας στο ότι απαιτεί άμεση (ρητή) και συγκεκριμένη ανάδραση από τους χρήστες και, επιπλέον, διαφοροποιεί τις προτεινόμενες απόψεις με βάση τις προσωπικές απόψεις του χρήστη και όχι με έναν καθολικό τρόπο.

Οι συγγραφείς του [84] προτείνουν ένα σύστημα συστάσεων ειδησεογραφικών άρθρων σε κοινωνικά δίκτυα (φόρουμ), το οποίο εκμεταλλεύεται τα σχόλια χρηστών για να προτείνει άρθρα. Η προσέγγισή τους στοχεύει στην κατασκευή θεματικών προφίλ, χρησιμοποιώντας τόσο το άρθρο, όσο και τα σχόλιά του. Τα προφίλ αυτά χρησιμοποιούνται για να ανακτηθούν σχετικά άρθρα. Ομοίως, στο [85], παρουσιάζεται μία μέθοδος για συστάσεις άρθρων σε χρήστες, τα οποία είναι πιθανόν να σχολιαστούν από αυτούς. Οι συγγραφείς προτείνουν μία υβριδική μέθοδο συστάσεων με την οποία εκμεταλλεύονται, πέρα από το κειμενικό περιεχόμενο των άρθρων, και μοτίβα συν-σχολιασμού άρθρων από χρήστες.

Οι συγγραφείς του [79] πρώτα προβλέπουν αν ένα άρθρο είναι πιθανό να σχολιαστεί και, στη συνέχεια, αν θα λάβει μεγάλο αριθμό σχολίων ή όχι. Για να το επιτύχουν, εφαρμόζουν δύο διαδοχικές φάσεις κατηγοριοποίησης (classification). Στο [78], οι ίδιοι συγγραφείς μοντελοποιούν και συγκρίνουν κατανομές σχολιασμού άρθρων από διάφορες ειδησεογραφικές πηγές και προβλέπουν το συνολικό όγκο σχολίων για ένα άρθρο, παρατηρώντας μία μικρή αρχική περίοδο σχολιασμού του.

Η δουλειά του [81] προσπαθεί να αναγνωρίσει τα μοτίβα συναισθημάτων των σχολιαστών, απέναντι σε πολιτικά άρθρα, καθώς και να προβλέψει τον πολιτικό προσανατολισμό τους από τα συναισθήματα που εκφράζονται στα σχόλια. Οι συγγραφείς εφαρμόζουν διάφορες τεχνικές, ανάλογα με το αν οι προβλέψεις αφορούν ένα χρήστη ή μία ομάδα χρηστών. Επίσης, λαμβάνουν υπόψη συμφραζόμενη πληροφορία, όπως τις ψήφους που έλαβε ένα σχόλιο. Στο [89], οι συγγραφείς μελετούν σχόλια χρηστών σε πολιτικά ειδησεογραφικά άρθρα και αξιολογούν την ικανοποίηση χρηστών όσον αφορά πολιτικές γνώμες. Με αυτόν τον τρόπο προσπαθούν να διακρίνουν μεταξύ χρηστών που ψάχνουν παρόμοιες απόψεις με τις δικές τους και χρηστών που ψάχνουν ετερογενείς απόψεις. Στο [77] πραγματοποιείται ανάλυση συναισθημάτων σε αναρτήσεις χρήστων στην ιστοσελίδα Yahoo! Answers. Οι συγγραφείς αναλύουν την επιρροή διαφόρων παραγόντων, όπως δημογραφικά στοιχεία, θεματικές κατηγορίες και χρονική στιγμή της ανάρτησης.

Οι συγγραφείς του [83] μελετούν την περιγραφικότητα των σχολίων, δηλαδή το βαθμό στον οποίο τα σχόλια είναι παρόμοια με το θέμα στο οποίο αναφέρονται. Τα αποτελέσματα που προκύπτουν είναι θετικά, από την άποψη ότι είναι δυνατή η εύρεση ενός συγκεκριμένου πλήθους σχολίων που αντιπροσωπεύουν επαρκώς το αντίστοιχο άρθρο. Στο [80], οι συγγραφείς μελετούν τις ανάγκες των χρηστών όσον αφορά το σχολιασμό άρθρων και διαξάγουν μία ποιοτική ανάλυση πάνω στα σχόλια που αναρτώνται στην ιστοσελίδα μίας διαδικτυακής εφημερίδας. Στο [82], πραγματοποιείται μία ανάλυση σχολίων, υπερσυνδέσμων και διασυνδέσεων μεταξύ μπλογκς. Οι συγγραφείς του [86], στοχεύουν στην παραγωγή περιλήψεων κειμένων, εκμεταλλευόμενοι τα αντίστοιχα σχόλια. Προκειμένου να συνθέσουν τις περιλήψεις, εξάγουν προτάσεις από το αρχικό κείμενο (π.χ. ανάρτηση σε μπλογκ), οι οποίες είναι

πολωμένες (biased) με όρους από τα σχόλια. Στο [87], οι συγγραφείς αναλύουν σχόλια αναρτήσεων σε μπλογκ και τη μεταξύ τους συσχέτιση. Συγκεκριμένα, εκτιμούν το συνολικό όγκο σχολίων στη μπλογκόσφαιρα, αναλύουν τη σχέση μεταξύ της δημοφιλίας ενός μπλογκ και των μοτίβων σχολιασμού του και μετρούν τη συνεισφορά του περιεχομένου των σχολίων στην προσβασιμότητα των μπλογκ.

Μία εκτενής ανασκόπηση θεμελειωδών εργασιών στη διαφοροποίηση δίνεται στο [99]. Η δουλειά του [93] περιγράφει τη μέθοδο/αρχή της μέγιστης οριακής σχετικότητας (maximal marginal relevance), η οποία επιδιώκει να μεγιστοποιήσει τη σχετικότητα ενός αποτελέσματος με το ερώτημα, ελαχιστοποιώντας ταυτόχρονα την ομοιότητα του αποτελέσματος με τα αποτελέσματα που έχουν ανακτηθεί πριν από αυτό. Για το σκοπό αυτό, η σχετικότητα των αποτελεσμάτων υπολογίζεται χρησιμοποιώντας δύο επιμέρους συναρτήσεις: μία που υπολογίζει την ομοιότητά τους με το ερώτημα και μία που υπολογίζει την ομοιότητα μεταξύ των αποτελεσμάτων. Το [94] θεωρεί μία μετρική αξιολόγησης η οποία δίνει ποινή σε ένα μοντέλο ανάκτησης αποτελεσμάτων, μόνο αν δεν επιστρέψει καθόλου σχετικά αποτελέσματα. Δεδομένου αυτού, οι συγγραφείς προτείνουν μία μέθοδο στην οποία κάθε αποτέλεσμα επιλέγεται με βάση την πιθανότητα να (μην) είναι όμοιο με τα προηγουμένως επιλεγμένα αποτελέσματα.

Στο [96], οι συγγραφείς εισάγουν ένα σύνολο από αξιώματα διαφοροποίησης και δείχνουν ότι δεν είναι δυνατόν κάποιος αλγόριθμος διαφοροποίησης να τα ικανοποιεί όλα ταυτόχρονα. Επίσης, παρουσιάζουν τρεις συναρτήσεις -στόχους για διαφοροποίηση, οι οποίες διαφέρουν όσον αφορά τον τρόπο σύγκρισης των αποτελεσμάτων. Οι συγγραφείς του [95] παρουσιάζουν ένα πλαίσιο αξιολόγησης των εννοιών της καινοτομίας (novelty) και της διαφοροποίησης σε ένα σύνολο αποτελεσμάτων. Ομοίως, στο [97], προτείνεται ένας άπληστος ευριστικός αλγόριθμος, καθώς επίσης και επεκτείνονται κάποιες καθιερωμένες μετρικές αξιολόγησης, ώστε να έχουν εφαρμογή στο σενάριο της διαφοροποιημένης αναζήτησης αποτελεσμάτων. Τέλος, το [98] παρουσιάζει μία μέθοδο για αποδοτική διαφοροποίηση δομημένων δεδομένων, όπου τα στοιχεία προς διαφοροποίηση δεν είναι έγγραφα, αλλά στοιχεία με διακριτές ιδιότητες, δηλαδή εγγραφές σε ένα πίνακα μίας βάσης.

Κεφάλαιο 3

Αύξηση Δεδομένων Εκπαίδευσης

Σε αυτό το κεφάλαιο παρουσιάζουμε τη μέθοδο που προτείνουμε για επέκταση του αρχικού σετ δεδομένων εκπαίδευσης για εκπαίδευση συναρτήσεων ταξινόμησης [113]. Η επέκταση αφορά πρακτικά τις κρίσεις σχετικότητας, οι οποίες επεκτείνονται σε δεδομένα που δεν έχει επισκοπήσει ο χρήστης. Με αυτό τον τρόπο, από ένα αρχικά μικρό σετ δεδομένων που έχει αξιολογηθεί από το χρήστη (για παράδειγμα τα πρώτα 10 αποτελέσματα που είδε), μπορεί να προκύψει ένα πολύ μεγαλύτερο σετ δεδομένων εκπαίδευσης. Αυτή η διαδικασία οδηγεί τόσο σε γρηγορότερη, όσο και πιο ομοιογενή εκπαίδευση, αφού επιτρέπει την εκπαίδευση συναρτήσεων ταξινόμησης σε μικρότερες (και άρα πιο ομοιογενείς θεματικά) ομάδες χρηστών.

Η μέθοδός μας συνοψίζεται ως εξής: Για κάθε ερώτημα, χρησιμοποιούμε τα πρώτα 10 αποτελέσματα για να εξάγουμε ένα αρχικό σετ από κρίσεις σχετικότητας R. Η επιλογή του αριθμού των αρχικών αποτελεσμάτων γίνεται αυθαίρετα και με βάση διάφορες μελέτες [4] που δείχνουν ότι, στην πλειοψηφία των περιπτώσεων, οι χρήστες κοιτάνε την πρώτη σελίδα των αποτελεσμάτων (10 πρώτα αποτελέσματα). Η διαδικασία που ακολουθείται, πάντως, είναι ανεξάρτητη από τον αρχικό αριθμό αποτελεσμάτων. Στη συνέχεια, εκτελείται συσταδοποίηση στο *σύνολο* των αποτελεσμάτων, ανεξάρτητα από το αν υπάρχουν ή όχι διαθέσιμες κρίσεις σχετικότητας για το κάθε αποτέλεσμα. Οι συστάδες που προκύπτουν περνούν από ένα βήμα προ-επεξεργασίας για να διαπιστωθεί αν είναι αρκετά ομοιογενείς ως προς τις κρίσεις σχετικότητας που χαρακτηρίζουν τα αποτελέσματα που περιέχουν. Οι συστάδες που δεν ικανοποιούν την παραπάνω απαίτηση απορρίπτονται από τη συνέχεια της διαδικασίας. Οι συστάδες που παραμένουν χρησιμοποιούνται για να εξαχθούν κρίσεις σχετικότητας για αποτελέσματα που δεν έχουν επισκοπηθεί από τους χρήστες, σχηματίζοντας ένα νέο σύνολο R'. Τέλος, το επεκτεταμένο σύνολο κρίσεων σχετικότητας $R \cup R'$ χρησιμοποιείται ως είσοδος για την εκπαίδευση συνάρτησης ταξινόμησης, με το μοντέλο των Μηχανών Διανυσμάτων Στήριξης για ταξινόμηση.

3.1 Εξαγωγή Κρίσεων Σχετικότητας

Έστω ότι έχουμε μία αρχική ταξινόμηση αποτελεσμάτων που εμφανίζεται στους χρήστες για κάθε ερώτημα που θέτουν. Η ταξινόμηση αυτή συνήθως προέρχεται από κάποια μηχανή

αναζήτησης πάνω στην οποία θέλουμε να εκτελέσουμε εξατομίκευση αναταξινομώντας τα αποτελέσματα. Όπως προείπαμε, άνευ βλάβης της γενικότητας, θεωρούμε ως ανάδραση από το χρήστη τα πρώτα 10 αποτελέσματα από κάθε ερώτημα, μαζί με τις κρίσεις σχετικότητας που τους αντιστοιχούν. Θεωρούμε, επίσης, ότι για τα υπόλοιπα αποτελέσματα κάθε ερωτήματος (αποτελέσματα κάτω από τη δέκατη θέση κατάταξης) δεν υπάρχουν κρίσεις σχετικότητας. Η προσέγγισή μας αποτελείται από δύο βήματα: συσταδοποίηση αποτελεσμάτων και επέκταση κρίσεων σχετικότητας, τα οποία αναλύονται στη συνέχεια.

3.1.1 Συσταδοποίηση Αποτελεσμάτων Αναζήτησης

Έστω ένας χώρος F διάστασης n, με $F = \{t_1, t_2, \ldots, t_n\}$, όπου t_n κάθε ένας από τους διακριτούς όρους (λέξεις) που αποτελούν τις διαστάσεις του χώρου. Οι διακριτοί αυτοί όροι προέρχονται από το σύνολο των εγγράφων που αποτελούν το διαθέσιμο σετ δεδομένων. Τότε, μπορούμε να αναπαραστήσουμε κάθε αποτέλεσμα ως ένα διάνυσμα $v_i = \{wtd_{i1}, wtd_{i2}, \ldots, wtd_{in}\}$, όπου $wtd_{ik} = tf_{ik} * log(N/df_k)$, tf_{ik}: η συχνότητα εμφάνισης του όρου t_k στο κείμενο του αποτελέσματος i, N ο συνολικός αριθμός των αποτελεσμάτων και df_k ο συνολικός αριθμός των αποτελεσμάτων που περιέχουν τον όρο t_k. Σημειώνουμε ότι στην προσέγγισή μας θεωρούμε ως κείμενο ενός αποτελέσματος τον τίτλο και την περίληψή του.

Αφού δημιουργήσουμε το αντίστοιχο διάνυσμα όρων για κάθε αποτέλεσμα, εφαρμόζουμε συσταδοποίηση με σκοπό να δημιουργήσουμε ομάδες αποτελεσμάτων με παρόμοιο κειμενικό περιεχόμενο. Η τεχνική συσταδοποίησης που χρησιμοποιούμε εφαρμόζει επαναλαμβανόμενες διχοτομήσεις (partitions) στην εκάστοτε αρχική συστάδα, μέχρι να καταλήξει στο επιθυμητό αποτέλεσμα [23, 24]. Έχει αποδειχθεί ότι αυτή η προσέγγιση δίνει εξαιρετικά αποτελέσματα σε συλλογές κειμένων αντίστοιχου μεγέθους (σχετικά μικρό μέγεθος κειμένων) [23]. Πιο αναλυτικά, η διαδικασία συσταδοποίησης είναι η εξής:

Όλα τα αποτελέσματα βρίσκονται αρχικά σε μία συστάδα. Στη συνέχεια, αρχίζει μία διαδικασία διχοτόμησης. Σε κάθε βήμα διχοτομείται μόνο μία από τις υπάρχουσες συστάδες. Η επιλογή της κατάλληλης συστάδας για περαιτέρω διχοτόμηση γίνεται έτσι ώστε να βελτιστοποιείται μία συγκεκριμένη συνάρτηση-κριτήριο (criterion function). Η συνάρτηση-κριτήριο στη δική μας περίπτωση στοχεύει στη μεγιστοποίηση της παρακάτω ποσότητας:

$$\sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} sim(v,u)}$$

όπου k ο αριθμός των συστάδων, S_i το σύνολο των αποτελεσμάτων που περιέχονται στη συστάδα i, και $sim(v,u)$ το σκορ ομοιότητας μεταξύ των αποτελεσμάτων v και u της συστάδας S_i.

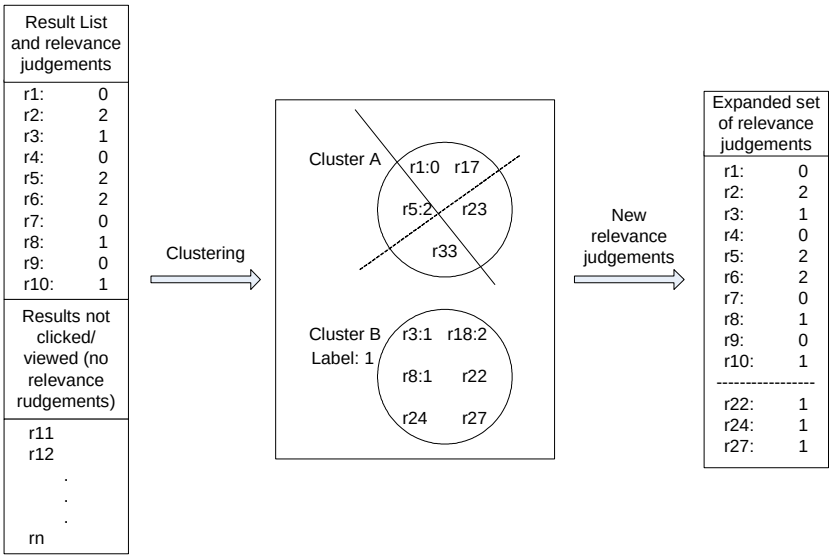
Για μετρική ομοιότητας μεταξύ δύο αποτελεσμάτων εφαρμόζουμε την cosine similarity στα διανύσματα όρων, σύμφωνα με την παρακάτω Εξίσωση:

$$sim(v,u) = \frac{\sum_{i=1}^{n}(wtd_{vi} \times wtd_{ui})}{\sqrt{\sum_{i=1}^{n} wtd_{vi}^2} \times \sqrt{\sum_{i=1}^{n} wtd_{ui}^2}}$$

Εκτελώντας την παραπάνω διαδικασία για κάθε ερώτημα ξεχωριστά, παράγουμε συστάδες των αποτελεσμάτων του. Στη συνέχεια δείχνουμε πώς τις εκμεταλλευόμαστε για να επεκτείνουμε τις κρίσεις σχετικότητας.

3.1.2 Επέκταση Κρίσεων Σχετικότητας

Η διαδικασία που ακολουθείται αναπαρίσταται στο παράδειγμα του Σχήματος 3.1. Έστω ότι τα αποτελέσματα r_1, r_5, r_{17}, r_{23}, r_{33} ανήκουν όλα στη ίδια συστάδα. Κάθε συστάδα μπορεί να περιέχει αποτελέσματα με ή χωρίς κρίσεις σχετικότητας. Για παράδειγμα, και με δεδομένο ότι ο χρήστης έχει κρίνει (δει ή πατήσει) αποτελέσματα μόνο από τις 10 πρώτες θέσεις της κατάταξης, κρίσεις σχετικότητας μπορούν να εξαχθούν για τα r_5, r_1, αλλά όχι για τα r_{17}, r_{23}, r_{33}. Σημειώνουμε επίσης ότι, στην περίπτωσή μας, θεωρούμε 3 κλάσεις σχετικότητας: 0 για *άσχετα*, 1 για *μερικώς σχετικά* και 2 για *πολύ σχετικά* αποτελέσματα.



Σχήμα 3.1: Παράδειγμα επέκτασης κρίσεων σχετικότητας με τη χρήση συσταδοποίησης

Από τη στιγμή που χρησιμοποιούμε τις συστάδες για να υπολογίσουμε κρίσεις σχετικότητας για μη επισκοπημένα αποτελέσματα, θα πρέπει να βεβαιωθούμε ότι οι συγκεκριμένες συστάδες είναι αρκετά ομοιογενείς. Ξέρουμε ήδη από τη διαδικασία συσταδοποίησης ότι η κάθε συστάδα περιέχει αποτελέσματα παρόμοιου περιεχομένου. Επιπλέον, όμως, απαιτούμε να περιέχουν και αποτελέσματα με παρόμοιες κρίσεις σχετικότητας. Πιο συγκεκριμένα, απορρίπτουμε για το υπόλοιπο της διαδικασίας συστάδες οι οποίες περιέχουν αποτελέσματα τα οποία διαφέρουν στις κρίσεις σχετικότητάς τους περισσότερο από μία μονάδες. Εν προκειμένω, αν μία συστάδα περιέχει κάποια αποτελέσματα χαρακτηρισμένα με 0 (άσχετα) και κάποια άλλα με 2 (πολύ σχετικά), τότε απορρίπτεται. Για τις συστάδες που παραμένουν, κάθε μία χαρακτηρίζεται από την πιο συχνή κρίση σχετικότητας μεταξύ των αποτελεσμάτων που περιέχει. Πιο τυπικά, τα παραπάνω μεταφράζονται ως εξής:

1. Έστω η συστάδα $c = \{r_1, r_2, \ldots r_n\}$ αποτελεσμάτων αναζήτησης

2. Έστω nj_c^0, nj_c^1, nj_c^2 ο αριθμός των αποτελεσμάτων στην c χαρακτηρισμένων ως 0 (άσχετα), 1 (μερικώς σχετικά), και 2 (πολύ σχετικά) αντίστοιχα, με $nj_c^0+nj_c^1+nj_c^2 \leq n$.

3. Εάν $nj_c^0 = 0$, τότε η κρίση του c προκύπτει από το $argmax(nj_c^1, nj_c^2)$.

4. Αλλιώς, εάν $nj_c^2 = 0$, τότε η κρίση του c προκύπτει από το $argmax(nj_c^0, nj_c^1)$.

5. Αλλιώς, απόρριψε τη c.

Στη συνέχεια, για κάθε συστάδα με κρίση j, αναθέτουμε σε όλα τα αποτελέσματα χωρίς κρίση σχετικότητας την κρίση j.

Επιστρέφοντας στο παράδειγμα του Σχήματος 3.1, και βασιζόμενοι στα προηγούμενα, έχουμε: Η συστάδα Α απορρίπτεται, αφού $nj_A^0 = 1$ και $nj_A^2 = 1$. Η συστάδα Β χαρακτηρίζεται με την κρίση 1, αφού $nj_B^1 = 2$ και $nj_B^2 = 1$. Όλα τα αποτελέσματα της συστάδας που δεν είχαν αρχικά κρίση σχετικότητας (r_{22}, r_{24}, r_{27}) χαρακτηρίζονται πλέον με την κρίση 1 (μερικώς σχετικά).

3.2 Πειραματική Μελέτη

Η μέθοδός μας στοχεύει στην επέκταση ενός αρχικού συνόλου κρίσεων σχετικότητας με σκοπό την παραγωγή περισσότερων δεδομένων εκπαίδευσης από τα ίδια δεδομένα ανάδρασης του χρήστη. Αρχικά παρουσιάζουμε το πειραματικό σετ δεδομένων και τις μετρικές πειραματικής αποτίμησης και στη συνέχεια αξιολογούμε τη μέθοδό μας με δύο τρόπους. Πρώτα εξετάζουμε την ποιότητα των νέων κρίσεων σχετικότητας που προκύπτουν. Κατόπιν, εξετάζουμε αν το επεκτεταμένο σύνολο κρίσεων σσχετικότητας οδηγεί σε καλής ποιότητας εκπαίδευση, η οποία μετράται με όρους ακρίβειας στα αποτελέσματα αναζήτησης. Συγκεκριμένα, εκπαιδεύουμε μία συνάρτηση ταξινόμησης με το επεκτεταμένο σύνολο κρίσεων σχετικότητας (υποσύνολο του αρχικού συνόλου, επεκτεταμένο με τη μέθοδό μας) και συγκρίνουμε την ακρίβεια ταξινόμησης αποτελεσμάτων του με την ακρίβεια μίας συνάρτησης που έχει εκπαιδευτεί στο αρχικό σύνολο των κρίσεων σχετικότητας. Για να εκτελέσουμε μία τέτοια αποτίμηση θα πρέπει να 'προσομοιώσουμε' μία τυπική συμπεριφορά αναζήτησης και ανάδρασης χρήστη πάνω στο συγκεκριμένο σετ δεδομένων που έχουμε. Αυτό γίνεται ως εξής:

- Οι χρήστες θέτουν ερωτήματα. Τα αποτελέσματα ταξινομούνται αρχικά σύμφωνα με το BM25 σκορ τους [25].

- Μόνο τα πρώτα 10 αποτελέσματα παρουσιάζονται στους χρήστες για να τα δουν/πατήσουν (αξιολογήσουν). Για αυτά τα αποτελέσματα κρατάμε τις κρίσεις σχετικότητας.

- Τα υπόλοιπα αποτελέσματα, για τα οποία επίσης διαθέτουμε κρίσεις σχετικότητας, αλλά δεν τις χρησιμοποιούμε στη μέθοδό μας, χρησιμοποιούνται ως μέσο επαλήθευσης της μεθόδου.

3.2.1 Πειραματικό Σετ Δεδομένων

Η μέθοδός μας ελέγχθηκε πειραματικά πάνω στη συλλογή κειμένων OHSUMED, ένα υποσύνολο της MEDLINE (βάση δεδομένων ιατρικών δημοσιεύσεων), που περιέχει 348566 εγγραφές (έγγραφα προς αναζήτηση) από 270 ιατρικά περιοδικά από το 1987 έως το 1991. Κάθε έγγραφη έχει την ακόλουθη δομή:

.I(id)

.U(MEDLINE id)

.M(Human-assigned MeSH terms)

.T(title)

.P(publication type)

.W(abstract)

.A (author)

.S(source).

από τα οποία εμείς χρησιμοποιήσαμε τα πεδία id (αναγνωριστικό), title (τίτλος) και abstract (περίληψη).

Επίσης, υπάρχουν διαθέσιμα 106 ερωτήματα με τα αντίστοιχα αποτελέσματά τους. Για κάθε ζεύγος ερωτήματος - αποτελέσματος υπάρχει και μία κρίση σχετικότητας : 0 (άσχετο), 1 (μερικώς σχετικό) και 2 (πολύ σχετικό). Συνολικά, υπάρχουν 16140 κρίσεις σχετικότητας για τα 106 ερωτήματα.

Τέλος, το σετ δεδομένων έχει οργανωθεί σε 5 υποσύνολα, συμβολιζόμενα ως $S1$, $S2$, $S3$, $S4$ και $S5$. Συνδυάζοντας κάθε φορά διαφορετικά υποσύνολα για να παράγουν τα: (α) σετ εκπαίδευσης (training set), (β) σετ επαλήθευσης (validation set) και (γ) σετ αξιολόγησης (test set), οι δημιουργοί του πειραματικού σετ έχουν παράξει 5 ανακατατάξεις (Folds). Το σετ επαλήθευσης χρησιμοποιείται για τη βελτιστοποίηση των παραμέτρων του μοντέλου εκπαίδευσης. Στη συγκεκριμένη δουλειά, όμως, δεν ενδιαφερόμαστε να συγκρίνουμε διαφορετικά μοντέλα εκπαίδευσης, παρά να εξετάσουμε πώς η μέθοδός μας βελτιώνει την αποδοτικότητα του ίδιου μοντέλου. Για αυτό, συγχωνεύουμε τα σετ εκπαίδευσης με τα σετ επαλήθευσης, καταλήγοντας τελικά στις 5 ανακατατάξεις του σετ δεδομένων που φαίνονται στον Πίνακα 3.1.

Ανακατατάξεις	Σετ Εκπαίδευσης	Σετ Αξιολόγησης
Fold1	{$S1$, $S2$, $S3$, $S4$}	$S5$
Fold2	{$S2$, $S3$, $S4$, $S5$}	$S1$
Fold3	{$S3$, $S4$, $S5$, $S1$}	$S2$
Fold4	{$S4$, $S5$, $S1$, $S2$}	$S3$
Fold5	{$S5$, $S1$, $S2$, $S3$}	$S4$

Πίνακας 3.1: Ανακατατάξεις του Σετ Εκπαίδευσης

3.2.2 Ποιότητα επεκτεταμένων κρίσεων σχετικότητας (Ποιότητα Συσταδοποίησης)

Σκοπός του πρώτου μας πειράματος είναι να εξετάσουμε την ακρίβεια της επέκτασης των κρίσεων σχετικότητας που επιτυγχάνουμε μέσω συσταδοποίησης. Ο πίνακας 3.2 παρουσιάζει τα επί τοις εκατό αποτελέσματα για διάφορες λύσεις συσταδοποίησης που δοκιμάσαμε. Η στήλη *numOfClusters* δείχνει τον αριθμό των συστάδων που παράγονται σε κάθε λύση συσταδοποίησης. Η στήλη *Correct* δείχνει το ποσοστό των κρίσεων σχετικότητας που προβλέφθηκαν σωστά. Η στήλη *Part. Correct* δείχνει το ποσοστό των κρίσεων σχετικότητας που ήταν μερικώς σωστές. Αναφέρεται στις περιπτώσεις όπου μία κρίση προβλέφθηκε 0 αντί για 1, 1 αντί για 0, 2 αντί για 1 ή 1 αντί για 2. Η στήλη Wrong δείχνει το ποσοστό των κρίσεων σχετικότητας που ήταν λάθος.

numOfClusters	Correct	Part. Correct	Wrong
5	55.8%	27.2%	17%
10	55.5%	26.3%	18.2%
15	54.8%	26.3%	18.9%
20	55.4%	25.7%	18.9%
25	55.5%	26.3%	18.2%
30	55.2%	27%	17.8%

Πίνακας 3.2: Ποσοστά επιτυχίας στο σύνολο των επεκτεταμένων αποτελεσμάτων

Αν και τα αποτελέσματα είναι σχεδόν ίδια για όλες τις συσταδοποιήσεις που δοκιμάζονται, η λύση με τις 5 συστάδες είναι λίγο καλύτερη από τις υπόλοιπες. Η λύση αυτή μας δίνει 55.8% σωστές προβλέψεις και 83% σωστές και μερικώς σωστες προβλέψεις, ποσοστά αρκετά ικανοποιητικά.

3.2.3 Ποιότητα αποτελεσμάτων αναταξινόμησης

Σε αυτήν την υποενότητα η προτεινόμενη μέθοδος αξιολογείται σε σχέση με την ποιότητα της συνάρτησης ταξινόμησης που εκπαιδεύει. Χρησιμοποιούμε το εργαλείο $\Sigma^{v}\mathrm{M}^{light}$ και τρέχουμε τα πειράματά μας στο σετ δεδομένων του Πίνακα 3.1. Χρησιμοποιούμε τα ακόλουθα μέτρα αξιολόγησης: *Precision at position n (P@n)*, *Mean average precision (MAP)* and *Normalized discount cumulative gain (NDCG)*.

Ακολουθούν οι ορισμοί των παραπάνω μέτρων:

$$P@n = \frac{\#relevant\ results\ in\ top\ n\ results}{n}$$

όπου ως 'relevant' θεωρούνται μόνο τα αποτελέσματα με κρίση 2.

$$MAP = \frac{\sum_{i=1}^{n} P@n * rel(n)}{\#total\ relevant\ results\ for\ the\ query}$$

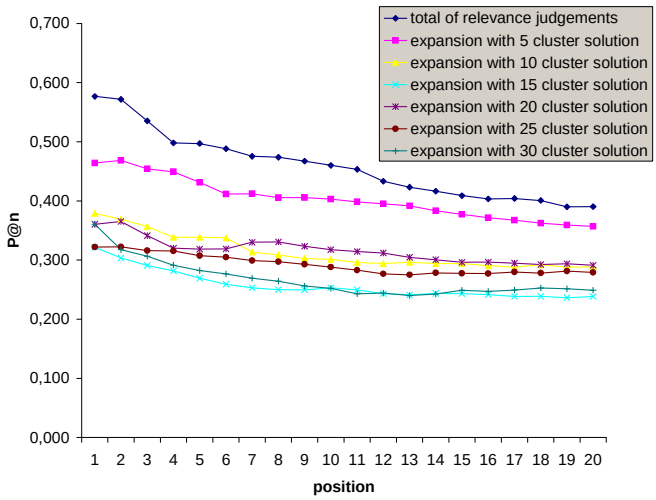
όπου N ο συνολικός αριθμός των αποτελεσμάτων και το $rel(n)$ ορίζεται ως εξής:

$$rel(n) = \begin{cases} 1 & \text{αν το } n\text{-ιοστό αποτέλεσμα είναι σχετικό} \\ 0 & \text{διαφορετικά} \end{cases}$$

$$NDCG(n) = Z_n \sum_{j=1}^{n} \begin{cases} 2^{r(j)} - 1 & j = 1 \\ \frac{2^{r(j)}-1}{\log j} & j > 1 \end{cases}$$

όπου $r(j)$ η θέση του j-οστού αποτελέσματος στην κατάταξη των αποτελεσμάτων και Z_n μία σταθερά κανονικοποίησης που εξασφαλίζει ότι, για την ιδανική ταξινόμηση, ισχύει $NDCG = 1$.

Τα πειραματικά αποτελέσματα φαίνονται στα Σχήματα 3.2 και 3.3 καθώς και στον Πίνακα 3.3. Το Σχήμα 3.2 δείχνει την ακρίβεια στις 20 πρώτες θέσεις κατάταξης των αποτελεσμάτων. Η γραφική παράσταση με την ένδειξη 'total of relevance judgement' δείχνει την ακρίβεια που επιτυγχάνεται με εκπαίδευση του συστήματος χρησιμοποιώντας το σύνολο των κρίσεων σχετικότητας που είναι διαθέσιμες στο σετ εκπαίδευσης. Αυτό σημαίνει ότι οι συγκεκριμένες τιμές ακρίβειας για το μέτρο $P@n$ αντιστοιχούν στην 'ιδανική' εκπαίδευση του συστήματος, όπου υπάρχουν διαθέσιμες κρίσεις σχετικότητας για όλα τα αποτελέσματα κάθε ερωτήματος. Οι υπόλοιπες γραφικές παραστάσεις παρουσιάζουν την ακρίβεια που επιτυγχάνεται με επέκταση ενός αρχικού σετ κρίσεων σχετικότητας (που αντιστοιχούν στα 10 αποτελέσματα με τη μεγαλύτερη τιμή του BM25 σκορ τους) με βάση τη μέθοδό μας, για διαφορετικές λύσεις συσταδοποίησης (όσον αφορά τον αριθμό των συστάδων που προκύπτουν).



Σχήμα 3.2: Σύγκριση των τιμών $P@n$ του ιδανικά εκπαιδευμένου συστήματος και των μεθόδων μας.

Αν και είναι δεδομένο ότι οι τιμές ακρίβειας του ιδανικά εκπαιδευμένου συστήματος θα είναι υψηλότερες από τις δικές μας, αυτό που θέλουμε να εξετάσουμε είναι κατά πόσο οι δικές μας τιμές προσεγγίζουν τις ιδανικές. Στο Σχήμα 3.2 βλέπουμε ότι η καλύτερη εκδοχή από τις μεθόδους που προτείνουμε είναι αυτή με τις λιγότερες συστάδες 'expansion with 5 clusters solution'. Παρατηρούμε, επίσης, ότι η συγκεκριμένη λύση προσσεγίζει αρκετά την ιδανική,

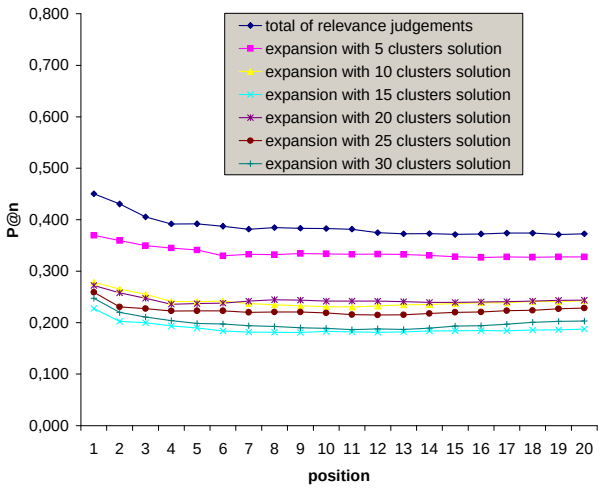
ειδικά όταν μιλάμε για χαμηλότερες θέσεις κατάταξης. Το πλεονέκτημα της μεθόδου μας έγκειται στο γεγονός ότι οι συγκεκριμένες τιμές ακρίβειας έχουν επιτευχθεί χρησιμοποιώντας μόνο 10 κρίσεις σχετικότητας ανά ερώτημα, ενώ το ιδανικό σύστημα χρησιμοποιεί κατά μέσο όρο 160 κρίσεις σχετικότητας ανά ερώτημα. Δηλαδή, χρησιμοποιώντας ένα πολύ μικρό ποσοστό κρίσεων σχετικότητας, επιτυγχάνουμε ακρίβεια πολύ κοντά στην ιδανική.

Ακόμα ενδεικτικότερες για την αποτελεσματικότητα της μεθόδού μας είναι οι τιμές μέσης ακρίβειας MAP που παρουσιάζονται στον Πίνακα 3.3. Ξανά, όπως είναι αναμενόμενο, η καλύτερη από τις λύσεις μας είναι αυτή με τις 5 συστάδες, η οποία επιτυγχάνει μέση ακρίβεια ίση με το 91% της ακρίβειας του ιδανικού συστήματος.

	trj	5 cl	10 cl	15 cl	20 cl	25 cl
MAP	0.399	0.364	0.325	0.286	0.320	0.312

Πίνακας 3.3: Σύγκριση MAP τιμών

Τέλος, στο Σχήμα 3.3 συγκρίνουμε τις $NDCG$ τιμές της μεθόδού μας με την ιδανική. Τα συμπεράσματα είναι όμοια με αυτά που προκύπτουν από τη σύγκριση των $P@n$ τιμών, δηλαδή, η καλύτερη λύση μας προσεγγίζει ικανοποιητικά την ιδανική.



Σχήμα 3.3: Σύγκριση των τιμών NDCG του ιδανικά εκπαιδευμένου συστήματος και των μεθόδων μας.

Κεφάλαιο 4

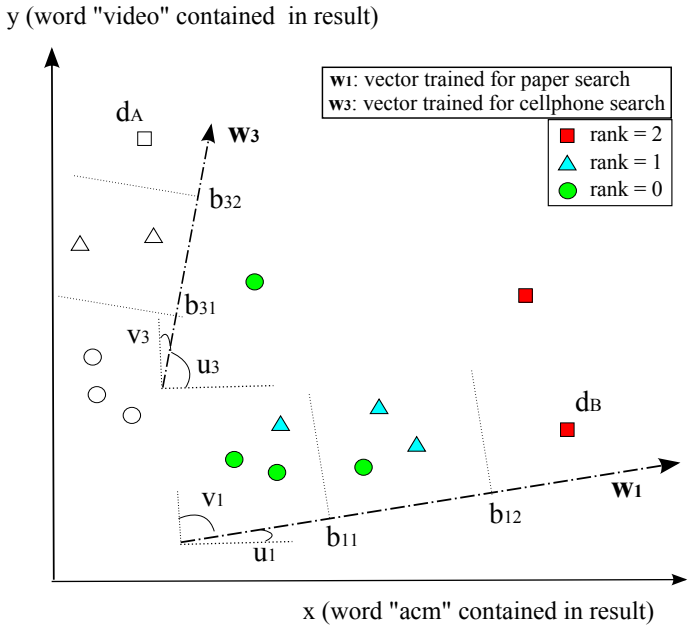
Βελτίωση Ποιότητας Εκπαίδευσης

Στο κεφάλαιο αυτό παρουσιάζουμε τη μέθοδο που προτείνουμε για τη βελτίωση της ποιότητας εκπαίδευσης συναρτήσεων ταξινόμησης, με σκοπό τη βελτίωση της αναταξινόμησης των αποτελεσμάτων για εξατομίκευση [112, 109, 108, 107]. Η μέθοδος μας έχει δύο βασικές εναλλακτικές: (α) εκπαίδευση με βάση το περιεχόμενο και (β) εκπαίδευση με βάση τη συμπεριφορά αναζήτησης. Στην παρούσα έκθεση θα θα επικεντρωθούμε κυρίως στη δεύτερη, η οποία είναι και η πιο ενδιαφέρουσα.

4.1 Συμπεριφορά Αναζήτησης και Μηχανές Διανυσμάτων Στήριξης για Ταξινόμηση

Σε αυτή την ενότητα δείχνουμε με ένα παράδειγμα πώς μπορούμε να συσχετίσουμε τη συμπεριφορά αναζήτησης χρηστών με το μοντέλο εκπαίδευσης συναρτήσεων ταξινόμησης (εν προκειμένω τις Μηχανές Διανυσμάτων Στήριξης). Για ευκολία στην παρουσίαση, θεωρούμε ότι τα διανύσματα χαρακτηριστικών αποτελούνται από δύο μόνο χαρακτηριστικά. Έστω ότι το χαρακτηριστικό x αντιπροσωπεύει τη συχνότητα της λέξης "acm" στο κείμενο του αποτελέσματος και ότι το χαρακτηριστικό y αντιπροσωπεύει τη συχνότητα της λέξης "video" αντίστοιχα. Τα σκιασμένα σχήματα (Σχήμα 4.1) αντιπροσωπεύουν αποτελέσματα που έχουν προέλθει από αναζήτηση δημοσιεύσεων, ενώ τα κενά σχήματα αντιπροσωπεύουν αποτελέσματα από αναζήτηση κινητών τηλεφώνων. Τα τετράγωνα σχήματα αντιπροσωπεύουν *πολύ σχετικά*, τα τρίγωνα *μερικώς σχετικά* και οι κύκλοι *άσχετα* αποτελέσματα.

Στον παραπάνω χώρο χαρακτηριστικών, εκπαιδεύουμε 2 μοντέλα ταξινόμησης, τα οποία αντιπροσωπεύονται από τα διανύσματα βαρών $\mathbf{w}_1$ και $\mathbf{w}_3$. Αυτά τα διανύσματα αντιστοιχούν σε αναζητήσεις για δημοσιεύσεις και αναζητήσεις για κινητά τηλέφωνα. Η κλίση κάθε διανύσματος, δηλαδή η γωνία ανάμεσα στο διάνυσμα και σε έναν από τους άξονες, υποδεικνύει πόσο σημαντικό είναι κάθε ένα από τα χαρακτηριστικά εκπαίδευσης στη διαδικασία ταξινόμησης αποτελεσμάτων. Για παράδειγμα, η γωνία u_1 ανάμεσα στο διάνυσμα $\mathbf{w}_1$ και στον άξονα x είναι μικρότερη από τη γωνία v_1 ανάμεσα στο διάνυσμα και στον άξονα y. Αυτό σημαίνει ότι μία αλλαγή στην τιμή του χαρακτηριστικού x (συχνότητα της λέξης "acm") είναι πιο πιθανό να προκαλέσει αλλαγή στην κατάταξη του αποτελέσματος από ό,τι μία αλλαγή στην τιμή του



Σχήμα 4.1: Εκπαιδευμένα διανύσματα βάρους και υπερεπιφάνειες στον χώρο χαρακτηριστικών.

χαρακτηριστικού y (συχνότητα της λέξης "video"). Έτσι, για τη συγκεκριμένη εκπαίδευση πάνω σε δεδομένα εκπαίδευσης από αναζητήσεις δημοσιεύσεων, το χαρακτηριστικό x είναι πιο σημαντικό από το y για την ταξινόμηση των αποτελεσμάτων. Το αντίθετο ισχύει, όταν εκπαιδεύουμε το διάνυσμα $\mathbf{w}_3$ πάνω σε δεδομένα που προέρχονται από αναζητήσεις κινητών τηλεφώνων: το χαρακτηριστικό y είναι πιο σημαντικό από το x, όπως φαίνεται από την κλίση του $\mathbf{w}_3$.

Το παραπάνω παράδειγμα περιγράφει δύο διαφορετικές συμπεριφορές αναζήτησης, δηλαδή, διαφορετικά μοτίβα αναζήτησης που ακολουθούνται από χρήστες για διαφορετικές κατηγορίες αναζήτησης. Όπως βλέπουμε, οι συμπεριφορές αυτές δεν εκφράζονται μέσω του περιεχομένου (ερωτημάτων ή αποτελεσμάτων), αλλά μέσω του χώρου χαρακτηριστικών $\mathbf{X} \epsilon R^d$ που επιλέγεται να αναπαραστήσει τα δεδομένα εκπαίδευσης (ερωτήματα, αποτελέσματα και κρίσεις σχετικότητας). Έτσι, μπορούμε να εντοπίσουμε και να εκμεταλλευτούμε συμπεριφορές αναζήτησης, χρησιμοποιώντας την κατανομή των δεδομένων εκπαίδευσης στο χώρο χαρακτηριστικών. Στην επόμενη ενότητα περιγράφεται αναλυτικά η μέθοδός μας.

4.2 Εκπαίδευση Οδηγούμενη από τη Συμπεριφορά Αναζήτησης

Σε αυτή την εργασία προτείνουμε το ότι ο εντοπισμός και η ομαδοποίηση συμπεριφορών αναζήτησης μπορούν να ωφελήσουν την ποιότητα της διαδικασίας εκπαίδευσης συναρτήσεων ταξινόμησης. Αντί να εκπαιδεύουμε μία καθολική συνάρτηση ταξινόμησης ή διαφορετικές συναρτήσεις ταξινόμησης για κάθε χρήστη (ή ομάδα παρόμοιων χρηστών), προτείνουμε την

εκπαίδευση πολλαπλών συναρτήσεων ταξινόμησης, κάθε μία από τις οποίες αντιστοιχεί σε μία διαφορετική συμπεριφορά αναζήτησης. Οι συναρτήσεις αυτές εκπαιδεύονται συνεργατικά, δηλαδή χρησιμοποιώντας δεδομένα εκπαίδευσης από το σύνολο των χρηστών. Στη συνέχεια, οι συναρτήσεις αυτές συνδυάζονται κατάλληλα για να δώσουν μία τελική αναταξινόμηση των αποτελεσμάτων, λαμβάνοντας υπόψη την ομοιότητα κάθε ερωτήματος με τις διάφορες συμπεριφορές αναζήτησης. Παρακάτω, περιγράφεται λεπτομερώς η προσέγγισή μας.

4.2.1 Συσταδοποίηση Αποτελεσμάτων

Το πρώτο βήμα της μεθόδου μας είναι η διαμέριση του αρχικού σετ δεδομένων εκπαίδευσης σε ομάδες ερωτημάτων των οποίων τα μεταδεδομένα εκπαίδευσης (αποτελέσματα και κρίσεις σχετικότητας) 'αναμένεται' να εκπαιδεύσουν παρόμοιες συναρτήσεις ταξινόμησης και, συνεπώς, αντιστοιχούν σε παρόμοιες συμπεριφορές αναζήτησης. Ονομάζουμε αυτές τις ομάδες *συστάδες συμπεριφορών αναζήτησης*.

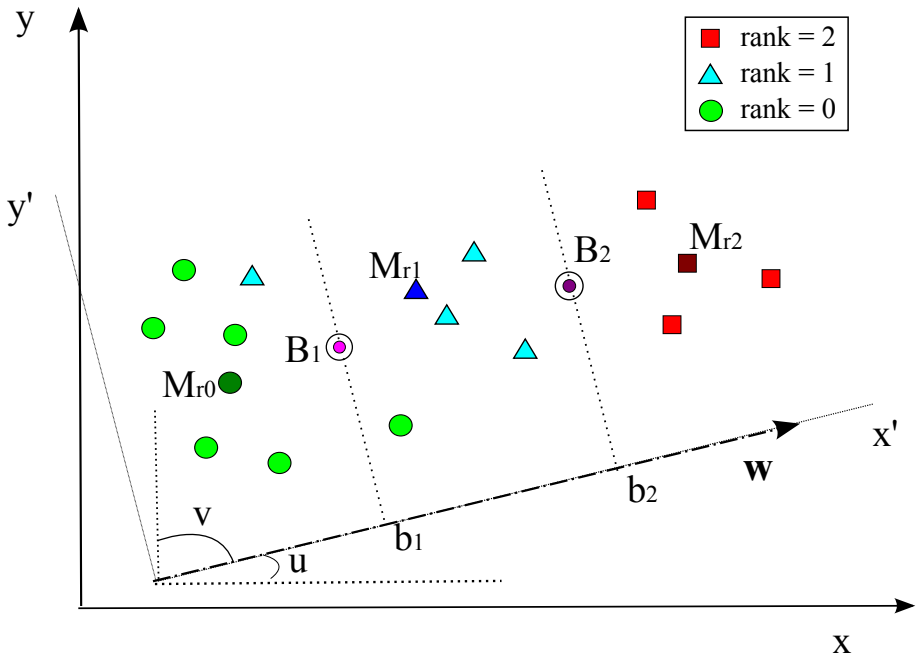
Μία απλοϊκή λύση για να παράξουμε τις παραπάνω ομάδες συμπεριφορών αναζήτησης είναι η ακόλουθη: Πρώτα, τρέχουμε τον αλγόριθμο Μηχανών Διανυσμάτων Στήριξης για Ταξινόμηση στα δεδομένα κάθε ερωτήματος *ξεχωριστά*, έτσι ώστε να εκπαιδεύσουμε ένα μοντέλο ταξινόμησης (δηλαδή ένα διάνυσμα βαρών $\mathbf{w}$)) *ανά ερώτημα*. Στη συνέχεια, εκτελούμε συσταδοποίηση στα διανύσματα, δηλαδή χρησιμοποιούμε τις διαστάσεις των $\mathbf{w}$ ως διαστάσεις συσταδοποίησης, για να παράγουμε συστάδες συμπεριφορών αναζήτησης. Αυτή η προσέγγισει εμφανίζει δύο προβλήματα. Πρώτον, τα δεδομένα εκπαίδευσης για κάθε εκπαιδευόμενο μοντέλο είναι πολύ λίγα για να παράγουν αξιόπιστα αποτελέσματα, αφού αποτελούνται από τα μεταδεδομένα μόνο ενός ερωτήματος κάθε φορά. Δεύτερον, η διακύμανση της παραμέτρου c (Εξίσωση 2.3) θα μπορούσε να δώσει πολύ διαφορετικά αποτελέσματα, για τα ίδια δεδομένα εκπαίδευσης, εξαιτίας, πάλι, του μικρού μεγέθους του σετ δεδομένων.

Στη λύση που δίνουμε, αντί να τρέχουμε ένα Διάνυσμα Μηχανών Στήριξης ανά ερώτημα, προσπαθούμε να προσεγγίσουμε το μοντέλο που πρόκειται να εκπαιδευτεί για τα δεδομένα κάθε ερωτήματος. Για αυτό το σκοπό, ορίζουμε 2 κατηγορίες διαστάσεων συσταδοποίησης, οι οποίες σχετίζονται με τα γεωμετρικά χαρακτηριστικά του μοντέλου εκπαίδευσης: τις *διαστάσεις κλίσης* (gradient dimensions), που αντιστοιχούν στις κλίσεις των διανυσμάτων $\mathbf{w}$ και τις *διαστάσεις περιθωρίου* (margin dimensions), που αντιστοιχούν στα σημεία όπου τα διανύσματα $\mathbf{w}$ 'κόβουν' τις υπερεπιφάνειες που χωρίζουν τις κλάσεις κατάταξης των αποτελεσμάτων (Ενότητα 2.1.2). Οι επόμενες υποενότητες περιγράφουν την κατασκευή των παραπάνω δύο διαστάσεων από τα δεδομένα εκπαίδευσης, καθώς και τη διαδικασία συσταδοποίησης που επιλέγεται ώστε να συμβαδίζει, πάλι, με τα γεωμετρικά χαρακτηριστικά του μοντέλου εκπαίδευσης.

Επιλογή Διαστάσεων Συσταδοποίησης

Όπως προείπαμε ορίζουμε δύο κατηγορίες διαστάσεων συσταδοποίησης: διαστάσεις κλίσης και διαστάσεις περιθωρίου.

Διαστάσεις κλίσης (f^g). Με αυτές τις διαστάσεις προσεγγίζουμε της κλίση του δια-



Σχήμα 4.2: Εξαγωγή διαστάσεων συσταδοποίησης στο χώρο χαρακτηριστικών.

νύσματος που εκπαιδεύεται από το μοντέλο Μηχανών Διανυσμάτων Στήριξης (SVM) για κάθε ερώτημα. Για να εξάγουμε αυτές τις διαστάσεις, εκμεταλλευόμαστε τα χαρακτηριστικά εκπαίδευσης του χώρου $\mathbf{X}\epsilon\mathbb{R}^d$, ο οποίος αναπαριστά ερωτήματα και τα μεταδεδομένα τους. Συγκεκριμένα, προσεγγίζουμε την κλίση του διανύσματος $\mathbf{w}$ για κάθε ζεύγος χαρακτηριστικών (x, y), όπου x, y χαρακτηριστικά του χώρου $\mathbf{X}$. Όπως φαίνεται στο Σχήμα 4.2, αυτό είναι ισοδύναμο με το να προσεγγίσουμε τη γωνία u.

Η προσέγγιση εκτελείται ως εξής. Για κάθε κλάση ταξινόμησης/σχετικότητας r, υπολογίζουμε τη μέση τιμή του κάθε ενός από τα χαρακτηριστικά (x, y) για το σύνολο των αποτελεσμάτων που ανήκουν στην κλάση. Έτσι, για παράδειγμα,

$$M^y_{r=2} = \frac{1}{m}\sum_{i=1}^{m} f^y_{i2} \tag{4.1}$$

είναι η y-συντεταγμένη του σημείου M_{r2} (Σχήμα 4.2), που είναι το κέντρο βάρους των m αποτελεσμάτων με κρίση σχετικότητας 2 (τετράγωνα σχήματα) στο δισδιάστατο χώρο χαρακτηριστικών xy. Το f^d_{ir} είναι η τιμή του χαρακτηριστικού d για το i-οστό αποτέλεσμα που ανήκει στην κλάση r.

Υπολογίζοντας την παραπάνω ποσότητα για όλα τα ζεύγη κλάσεων και για τα δύο χαρακτηριστικά x, y, μπορούμε πλέον να προσεγγίσουμε την εφαπτομένη της γωνίας u με τη βοήθεια του παρακάτω τύπου:

$$f^g_{(x,y),(a,b)} = \frac{M^y_{r=b} - M^y_{r=a} + \varepsilon}{M^x_{r=b} - M^x_{r=a} + \varepsilon} \tag{4.2}$$

όπου (a, b) είναι ζεύγη κλάσεων ταξινόμησης/σχετικότητας ($a, b \in \{0, 1, 2\}$ και $a \neq b$ στην περίπτωσή μας), (x, y) είναι ζεύγη χαρακτηριστικών, και το $\varepsilon = 10^{-9}$ χρησιμοποιείται

για να αποφευχθούν μηδενικές τιμές και διαιρέσεις με το μηδέν.

Κάθε ένας από τους όρους $f^g_{(x,y),(a,b)}$, οι οποίοι υπολογίζονται για όλα τα ζεύγη χαρακτηριστικών (x, y) και για όλα τα ζεύγη κλάσεων, χρησιμοποιείται ως μία διάσταση συσταδοποίησης. Έτσι, για ένα χώρο χαρακτηριστικών μεγέθους d και για ρ διακριτές κλάσεις ταξινόμησης/σχετικότητας, παράγονται $\frac{d(d-1)}{2} \cdot \frac{\rho(\rho-1)}{2}$ διαστάσεις κλίσης για συσταδοποίηση.

Παρατηρούμε, όμως, ότι, στις περισσότερες περιπτώσεις, από τα αποτελέσματα ενός ερωτήματος τα περισσότερα κρίνονται ως *άσχετα* και λιγότερα ως *μερικώς σχετικά ή πολύ σχετικά*. Έτσι, μπορούμε να επεκτείνουμε την Εξίσωση 4.2, ομαδοποιώντας όλες τις 'θετικές ' κλάσεις σχετικότητας ($r > 0$) σε μία κοινή κλάση, απέναντι στην κλάση των άσχετων ($r = 0$) αποτελεσμάτων:

$$f^g_{(x,y),(0,P)} = \frac{M^y_{r=P} - M^y_{r=0} + \varepsilon}{M^x_{r=P} - M^x_{r=0} + \varepsilon} \tag{4.3}$$

όπου $(0, P)$, $P \in \{1, 2\}$ είναι (μοναδικό) ζεύγος άσχετων και 'μη-άσχετων' αποτελεσμάτων (μερικώς σχετικών και πολύ σχετικών στην περίπτωσή μας). Αυτή η παραλλαγή παράγει $\frac{d(d-1)}{2}$ διαστάσεις κλίσης για συσταδοποίηση. Η εκδοχή αυτή παράγει ελαφρώς καλύτερα αποτελέσματα στα σετ δεδομένων όπου δοκιμάσαμε τις μεθόδους μας.

Διαστάσεις περιθωρίου (f^m). Με αυτές τις διαστάσεις προσεγγίζουμε τις θέσεις των σημείων b_i στο διάνυσμα βαρών $\mathbf{w}$. Αυτά τα σημεία ορίζουν κάθετες στο διάνυσμα υπερεπιφάνειες που διαχωρίζουν αποτελέσματα διαφορετικής κρίσης (κλάσης) σχετικότητας. Για να προσεγγίσουμε τις θέσεις των b_i, χρησιμοποιούμε τις προβολές των σημείων B_i στο $\mathbf{w}$, όπου B_i είναι το βαρύκεντρο των σημείων των αποτελεσμάτων δύο γειτονικών κλάσεων σχετικότητας. Οι συντεταγμένες των B_i υπολογίζονται χρησιμοποιώντας τα σημεία M (Εξίσωση 4.1), αφού, ουσιαστικά, τα B_i είναι τα βαρύκεντρα των M. Για παράδειγμα, στο Σχήμα 4.2, το

$$B^y_2 = \frac{M^y_{r=2} + M^y_{r=1}}{2} \tag{4.4}$$

είναι η y-συντεταγμένη του σημείου B_2, που είναι το βαρύκεντρο όλων των αποτελεσμάτων κλάσης 2 (τετράγωνα σχήματα) **και** κλάσης 1 (τρίγωνα σχήματα) στο δισδιάστατο χώρο χαρακτηριστικών xy.

Για να υπολογίσουμε την προβολή των σημείων B_i στο διάνυσμα $\mathbf{w}$, περιστρέφουμε τον δισδιάστατο χώρο xy, έτσι ώστε ένας από τους δύο άξονες να συμπέσει (να γίνει παράλληλος) με το $\mathbf{w}$. Άνευ βλάβης της γενικότητας, περιστρέφουμε τον χώρο κατά u μοίρες αντιωρολογιακά, όπου η ποσότητα $u = \arctan f^g$ προσεγγίζεται από τις διαστάσεις κλίσης. Στο νέο χώρο $x'y'$, το $\mathbf{w}$ είναι παράλληλο με τον άξονα x', οπότε $b^{x'}_i = B^{x'}_i$, αφού το ευθύγραμμο τμήμα b_iB_i είναι κάθετο στο $\mathbf{w}$.

Τελικά, οι διαστάσεις περιθωρίου, υπολογίζονται από τον τύπο:

$$f^m_{(x,y),(a,b)} = \frac{\frac{1}{m}\sum_{i=1}^{m} f^{x'}_{ib} + \frac{1}{n}\sum_{i=1}^{n} f^{x'}_{ia}}{2} \tag{4.5}$$

όπου τα (a, b) και (x, y) συμβολίζουν ζεύγη κλάσεων και χαρακτηριστικών αντίστοιχα, τα $f^{x'}_{ir} = f^x_{ir} \cos u - f^y_{ir} \sin u$ είναι οι τιμές των χαρακτηριστικών στον περιστραμμένο άξονα x' και $u = \arctan f^g$ είναι η γωνία περιστροφής.

Όμοια με τον υπολογισμό των διαστάσεων κλίσης, μπορούμε να επεκτείνουμε την Εξίσωση 4.6 ομαδοποιώντας όλες τις 'θετικές ' κλάσεις σε μία:

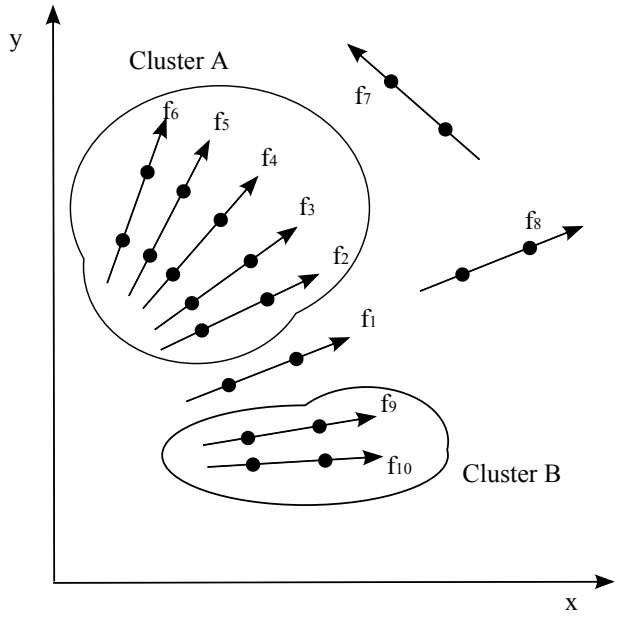
$$f^m_{(x,y),(0,P)} = \frac{\frac{1}{m}\sum_{i=1}^{m} f^{x'}_{iP} + \frac{1}{n}\sum_{i=1}^{n} f^{x'}_{i0}}{2} \tag{4.6}$$

όπου $(0, P)$ είναι το ζεύγος κλάσεων άσχετων και μη-άσχετων αποτελεσμάτων. Αυτή η παραλλαγή οδηγεί σε ένα σύνολο από $\frac{d(d-1)}{2}$ διαστάσεις περιθωρίου για συσταδοποίηση.

Με τη βοήθεια των παραπάνω διαστάσεων συσταδοποίησης, έχουμε, πλέον, τη δυνατότητα να εντοπίσουμε συμπεριφορές αναζήτησης, εκμεταλλευόμενοι τις κατανομές των αποτελεσμάτων (με διαθέσιμες κρίσεις σχετικότητας) στο χώρο χαρακτηριστικών των δεδομένων εκπαίδευσης. Το επόμενο βήμα είναι η επιλογή της κατάλληλης μεθοδολογίας συσταδοποίησης, η οποία θα εκμεταλλεύεται τις διαστάσεις αυτές για να ομαδοποιεί συμπεριφορές αναζήτησης.

Μεθοδολογία συσταδοποίησης

Για να καταλήξουμε στον καταλληλότερο αλγόριθμο συσταδοποίησης για το συγκεκριμένο πρόβλημα, εξετάζουμε τη γεωμετρική θεώρηση των αντικειμένων προς συσταδοποίηση. Στην περίπτωσή μας, τα αντικείμενα είναι ερωτήματα. Τα ερωτήματα αυτά, όμως, αντιπροσωπεύονται από τις κατανομές των σχετικών και άσχετων αποτελεσμάτων τους στο χώρο χαρακτηριστικών. Στην προηγούμενη υποενότητα μοντελοποιήσαμε αυτές τις κατανομές με διαστάσεις κλίσης (f^g) και περιθωρίου (f^m).



Σχήμα 4.3: Συσταδοποίηση στο χώρο χαρακτηριστικών.

Το Σχήμα 4.3 δίνει ένα παράδειγμα της λύσης που θέλουμε να πετύχουμε. Θεωρούμε ένα ζεύγος χαρακτηριστικών (x, y) για την ανάλυσή μας. Κάθε διάνυσμα f_i αντιπροσωπεύει ένα ερώτημα q_i. Η κλίση του διανύσματος αντιστοιχεί στη διάσταση κλίσης για συσταδοποίηση,

όσον αφορά το ζεύγος χαρακτηριστικών (x, y). Τα τονισμένα σημεία πάνω σε κάθε διάνυσμα αντιστοιχούν στις διαστάσεις περιθωρίου.

Στο παράδειγμά μας μία ιδανική λύση συσταδοποίησης θα τοποθετούσε τα ερωτήματα q_1 και q_7 σε διαφορετικές συστάδες, αφού οι κλίσεις των διανυσμάτων τους f_1 και f_7, διαφέρουν σημαντικά. Το ίδιο θα ίσχυε και για τα ερωτήματα q_1 και q_8, αφού, αν και τα διανύσματά τους έχουν παρόμοιες κλίσεις, οι θέσεις των σημείων που αντιστοιχούν στις διαστάσεις περιθωρίου διαφέρουν σημαντικά. Επίσης, θα προτιμούσαμε το ερώτημα q_1 να περιληφθεί στη συστάδα B, αντί για την A: Αν και το διάνυσμα f_2 είναι πιο όμοιο (όσον αφορά στην κλίση και στα σημεία περιθωρίου) με το διάνυσμα f_1 από ό,τι με τα διανύσματα f_9 και f_{10}, το κεντροειδές διάνυσμα της συστάδας B είναι πιο όμοιο με το f_1 από ό,τι το κεντροειδές της A. Συμπερασματικά, θέλουμε να παράγουμε *κεντρικοποιημένες συστάδες* (center-based clusters), δηλαδή συστάδες στις οποίες κάθε αντικείμενο είναι πλησιέστερο στο κεντροειδές της συστάδας του, παρά στο κεντροειδές οποιασδήποτε άλλης συστάδας [26].

Για να ικανοποιήσουμε τις παραπάνω προϋποθέσεις, επιλέγουμε (όπως και στην Ενότητα 3.1.1) έναν αλγόριθμο διαμερίσεων [23, 24]. Η συνάρτηση κριτήριο που οδηγεί σε κεντρικοποιημένες συστάδες στοχεύει στη μεγιστοποίηση της παρακάτω ποσότητας:

$$\sum_{r=1}^{k} \sum_{d_i \epsilon S_r} \cos(d_i, C_r)$$

όπου k είναι ο αριθμός των συστάδων, S_r είναι το σύνολο των αντικειμένων της συστάδας r, d_i είναι το i-οστό αντικείμενο της συστάδας, C_r είναι το κεντροειδές της συστάδας και $\cos(d_i, C_r)$ η cosine similarity μετρική ομοιότητας μεταξύ αντικειμένων της συστάδας και του κεντροειδούς.

4.2.2 Εκπαίδευση Συναρτήσεων Ταξινόμησης

Για κάθε συστάδα συμπεριφοράς αναζήτησης C_i που παράγεται, εκπαιδεύουμε μία ξεχωριστή συνάρτηση ταξινόμησης F_i, χρησιμοποιώντας τις Μηχανές Διανυσμάτων Στήριξης για Ταξινόμηση που βασίζονται στο μοντέλο των [2, 3]. Κάθε μία από αυτές τις συναρτήσεις εκπαιδεύεται πάνω στα δεδομένα που ανήκουν στην αντίστοιχη συστάδα C_i.

Ως χαρακτηριστικά για την εκπαίδευση χρησιμοποιούμε τα ίδια με το σετ δεδομένων για πειράματα LETOR[1]. Αυτά τα χαρακτηριστικά περιγράφονται στο [27] και, ενδεικτικά, αφορούν τιμές σκορ όπως TF, IDF, LMIR and BM25, θεωρώντας ως κείμενο τον τίτλο, την περίληψη και τη διεύθυνση του αποτελέσματος, καθώς και τιμές pagerank.

4.2.3 Αντιστοίχιση Συστάδων και Ερωτημάτων

Για κάθε συστάδα, εξάγουμε το κείμενο από όλα τα ερωτήματα που περιέχει ορίζοντας έτσι την *κειμενική αναπαράσταση* της συστάδας. Θεωρούμε αυτήν την αναπαράσταση κάθε συστάδας ως ένα ξεχωριστό 'έγγραφο'. Στη συνέχεια, ευρετηριάζουμε όλα αυτά τα έγγραφα χρησιμοποιώντας το Lucene[2], μία βιβλιοθήκη για ευρετηρίαση και αναζήτηση. Κάθε νέο ε-

[1] http://research.microsoft.com/en-us/um/beijing/ projects/letor/LETOR4.0/Data/Features_in_LETOR4.pdf
[2] http://lucene.apache.org/

ρώτημα, αποτελούμενο από τους όρους $t_1, t_2, \ldots, t_n$, το μετατρέπουμε στο εξής: $((t_1$ AND t_2 AND ... AND $t_n)$ OR t_1 OR t_2 ... OR t_n. Οι σύνδεσμοι OR χρησιμοποιούνται για την χαλάρωση του αρχικού ερωτήματος, έτσι ώστε να επιστραφούν από το ευρετήριο και έγγραφα-συστάδες που περιέχουν μερικούς από τους όρους του νέου ερωτήματος. Παράλληλα, αυτή η μετατροπή εξασφαλίζει ότι τα έγγραφα-συστάδες που περιέχουν όλους τους όρους θα πάρουν αρκετά μεγαλύτερο σκορ από τα υπόλοιπα.

Στη συνέχεια, χρησιμοποιούμε τη συνάρτηση σκορ του Lucene[3] για να πάρουμε μία λίστα από έγγραφα-συστάδες, μαζί με τα αντίστοιχα σκορ που υποδηλώνουν πόσο (κειμενικά) όμοιο είναι το περιεχόμενο κάθε συστάδας με το νέο ερώτημα. Επιλέγουμε τις k πιο όμοιες συστάδες που θα χρησιμοποιηθούν στη συνέχεια και απορρίπτουμε τις υπόλοιπες. Για τις επιλεγμένες συστάδες κανονικοποιούμε τα σκορ τους, έτσι ώστε να αθροίζονται στη μονάδα.

Στο τέλος, σε κάθε νέο ερώτημα q που τίθεται στο σύστημα ανατίθενται k βάρη w_i, $1 \leq i \leq k$, που αντιπροσωπεύουν το κανονικοποιημένο σκορ ομοιότητας (κειμενικού περιεχομένου) του ερωτήματος με τις k πιο όμοιες σε αυτό συστάδες.

4.2.4 Αναταξινόμηση αποτελεσμάτων

Όταν ένας χρήστης θέσει ένα νέο ερώτημα, η ομοιότητα w_i του ερωτήματος με κάθε συστάδα συμπεριφοράς αναζήτησης C_i υπολογίζεται με τον τρόπο που περιγράφεται στην προηγούμενη Υποενότητα. Στη συνέχεια, χρησιμοποιώντας κάθε ένα από τα μοντέλα ταξινόμησης F_i εκπαιδευμένα για τις αντίστοιχες συστάδες C_i, παράγουμε k διαφορετικές ταξινομήσεις R_i των αποτελεσμάτων του ερωτήματος, με rs_{ij} το σκορ του αποτελέσματος j σύμφωνα με το μοντέλο F_i. Σημειώνουμε ότι ο χρησιμοποιούμενος αλγόριθμος Μηχανών Διανυσμάτων Στήριξης δεν παρέχει συγκεκριμένες τιμές σκορ, παρά τιμές που καθορίζουν τις σχετικές θέσεις κατάταξης των αποτελεσμάτων. Για να παράγουμε τιμές σκορ, ορίζουμε πρώτα την ποσότητα:

$$rs_{ij} = M - r_{ij}$$

όπου M ο αριθμός των αποτελεσμάτων για το ερώτημα και r_{ij} η κατάταξη του αποτελέσματος j, σύμφωνα με το μοντέλο F_i. Τότε, το τελικό σκορ για κάθε αποτέλεσμα j δίνεται από τον τύπο:

$$score(j) = \sum_{i=1}^{k} w_i rs_{ij} \tag{4.7}$$

Δηλαδή, τελικά, συνδυάζουμε τις διαφορετικές ταξινομήσεις που προκύπτουν από μοντέλα που αντιστοιχούν στις πιο όμοιες με το ερώτημα συστάδες συμπεριφοράς αναζήτησης.

Ύστερα, όμως, από παρατήρηση πειραματικών αποτελεσμάτων, συμπεράναμε ότι η συνάρτηση ταξινόμησης που προτείνουμε (Εξίσωση 4.4.4) δίνει πολύ καλά αποτελέσματα σε υψηλές θέσεις ταξινόμησης, ενώ η βασική ταξινόμηση (δηλαδή αυτή που προκύπτει ύστερα από εκπαίδευση μίας κοινής συνάρτησης ταξινόμησης στο σύνολο των δεδομένων εκπαίδευσης) δουλεύει καλύτερα σε χαμηλότερες θέσεις της κατάταξης. Έτσι, ως τελική συνάρτηση

[3] http://lucene.apache.org/java/3_0_1/api/core/org/apache/ lucene/search/Similarity.html

ταξινόμησης, επιλέγουμε να ευνοούμε το δικό μας σκορ σε υψηλές θέσεις και το βασικό σκορ σε χαμηλότερες:

$$score(j) = \begin{cases} l \cdot \sum_{i=1}^{N} w_i rs_{ij} + (1-l) \cdot rs_j^b, \\ \text{for the top-a results according to the Eq. 4.4.4} \\ \\ (1-l) \cdot \sum_{i=1}^{N} w_i rs_{ij} + l \cdot rs_j^b, \\ \text{for the rest of the results} \end{cases} \tag{4.8}$$

όπου $l \in [0.5, 1]$ και rs_j^b το σκορ της βασικής συνάρτησης. Οι βέλτιστες τιμές των παραμέτρων α και l υπολογίζονται ύστερα από πειραματικές δοκιμές του συστήματος.

4.3 Πειραματική Μελέτη

Σε αυτήν την ενότητα παρουσιάζουμε την πειραματική αποτίμηση της μεθόδου μας και δείχνουμε την αποτελεσματικότητά της. Αρχικά παρουσιάζουμε το πειραματικό σετ δεδομένων και την απαραίτητη προεργασία που εκτελέσαμε σε αυτό. Στη συνέχεια, συγκρίνουμε τη μέθοδό μας με τη βασική μέθοδο σύγκρισης και αναλύουμε τη βελτίωση που επιτυγχάνουμε.

4.3.1 Σετ δεδομένων και προεργασία

Για τα πειράματά μας χρησιμοποιήσαμε την τελευταία έκδοση[4] του LETOR [27], που χρησιμοποιεί δεδομένα και αποτελέσματα από το TREC[5]. Το αρχικό σετ δεδομένων περιείχε 784 ερωτήματα μαζί με ένα σύνολο από 15211 αξιολογημένα αποτελέσματα. Η αξιολόγηση (κρίσεις σχετικότητας) έγινε χρησιμοποιώντας 3 κλάσεις σχετικότητας: 0, 1, 2, που αντιστοιχούν σε *άσχετα*, *μερικώς σχετικά* και *πολύ σχετικά* αποτελέσματα. Τα αποτελέσματα των ερωτημάτων αναπαρίστανται ως διανύσματα χαρακτηριστικών αποτελούμενα από τα χαρακτηριστικά που περιγράφονται στο [27]. Για καλύτερη αξιολόγηση, υπάρχουν διαθέσιμες 5 αναδιατάξεις του σετ δεδομένων, σε κάθε μία από τις οποίες διαφορετικά υποσύνολα του σετ αποτελούν τα (α) σετ εκπαίδευσης, (β) σετ επαλήθευσης και (γ) σετ αξιολόγησης.

Ύστερα από εξέταση του σετ δεδομένων παρατηρήσαμε ότι υπάρχουν ερωτήματα που έχουν μόνο 'άσχετα' αποτελέσματα (κρίσεις σχετικότητας 0). Αυτά τα ερωτήματα, προφανώς, δεν συνεισφέρουν στη φάση εκπαίδευσης, αφού, για κάθε ερώτημα, τουλάχιστον δύο διαφορετικές κλάσεις σχετικότητας απαιτούνται για την εκπαίδευση του μοντέλου. Έτσι, αφαιρέσαμε αυτά τα ερωτήματα, καταλήγοντας στις 5 διαμερίσεις που φαίνονται στον Πίνακα 4.1.

4.3.2 Αξιολόγηση μεθόδου

Σε αυτήν την ενότητα συγκρίνουμε τη μέθοδό μας (M) με τη βασική μέθοδο σύγκρισης (B), σύμφωνα με την οποία εκπαιδεύουμε ένα μόνο μοντέλο Μηχανών Διανυσμάτων Στήριξης

[4] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4dataset.aspx
[5] http://trec.nist.gov/

Folds	Training set	Validation set	Test set
Fold1	339	119	105
Fold2	353	105	105
Fold3	347	105	111
Fold4	330	111	122
Fold5	322	122	119

Πίνακας 4.1: Διαμέριση του σετ δεδομένωνLETOR: αριθμός ερωτημάτων που αντιστοιχούν σε κάθε σετ.

πάνω στο σύνολο των ερωτημάτων. Η σύγκριση των δύο προσεγγίσεων γίνεται με τη χρήση των μέτρων *Precision at position n (P@n)*, *Mean average precision (MAP)*, *Normalized discounted cumulative gain (NDCG)* και *Mean NDCG*.

Διαδικασία σύγκρισης

Η έκδοση του αλγορίθμου Μηχανών Διανυσμάτων Στήριξης που χρησιμοποιείται στο LETOR[6] απαιτεί τη ρύθμιση μόνο μίας παραμέτρου c. Αυτή η παράμετρος ρυθμίζει την ισορροπία μεταξύ των δύο όρων της Εξίσωσης 2.3 (Ενότητα 2.1.2). Το LETOR θέτει τις ακόλουθες τιμές για την παράμετρο: {0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10}.

Η μέθοδός μας εισάγει τέσσερις επιπλέον παραμέτρους: το N, δηλαδή τον αριθμό των συστάδων συμπεριφοράς αναζήτησης, το k, δηλαδή των αριθμό των συστάδων που χρησιμοποιούνται για αναταξινόμηση, το l, δηλαδή τη βαρύτητα του σκορ που προκύπτει από τη μέθοδό μας έναντι της βαρύτητας $(1-l)$ του σκορ της βασικής μεθόδου σύγκρισης B (Εξίσωση 4.8), και το α, δηλαδή το κατώφλι στο οποίο οι δύο παραπάνω βαρύτητες αντιστρέφονται.

Για κάθε διαμέριση του σετ δεδομένων ρυθμίσαμε το c στην τιμή που επιτυγχάνει τη μέγιστη MAP τιμή. Επίσης, τρέξαμε τη μέθοδό μας M σε κάθε διαμέριση, ώστε να ρυθμίσουμε διαδοχικά τα c, N, k, l και α. Τελικά, η διαδικασία ρύθμισης των παραμέτρων στόχευε στη βελτιστοποίηση των τιμών MAP, τόσο για τη βασική μέθοδο B, όσο και για την προτεινόμενη μέθοδο M. Τα αποτελέσματα που παρουσιάζονται στη συνέχεια επιτυγχάνονται θέτοντας ως εξής τις παραμέτρους της μεθόδου μας: $k = 3$, $l = 0.8$ και $\alpha = 4$. Οι τιμές των παραμέτρων N και c παρουσιάζονται, για κάθε διαμέριση ξεχωριστά, στον Πίνακα 4.2.

	Fold1		Fold2		Fold3		Fold4		Fold5	
Method	**N**	**c**	**N**	**c**	**N**	**c**	**N**	**c**	**N**	**c**
B	-	2	-	0.2	-	0.2	-	1	-	5
M	10	2	25	0.1	5	0.5	30	1	20	10

Πίνακας 4.2: Βέλτιστες τιμές των παραμέτρων N και c για κάθε διαμέριση

[6] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4baseline.aspx

Αξιολόγηση ακρίβειας

Ο Πίνακας 4.8 παρουσιάζει τις τιμές των *Mean Average Precision* και *Mean NDCG* για τη μέθοδό μας και τη βασική μέθοδο. Όπως βλέπουμε, η προτεινόμενη μέθοδος επιτυγχάνει καλύτερα αποτελέσματα από τη βασική, και για τα δύο μέτρα, στις περισσότερες περιπτώσεις, ενώ στις υπόλοιπες επιτυγχάνει ίδια αποτελέσματα. Η μέση τιμή του MAP στις διαμερίσεις είναι 2% υψηλότερη από τη βασική μέθοδο. Αυτή είναι μία σημαντική βελτίωση, αν λάβουμε υπόψη μας τη σύνθεση του συγκεκριμένου πειραματικού σετ δεδομένων που χρησιμοποιούμε. Σε τέτοιου τύπου σετ δεδομένων, ζητείται από απροσδιόριστο αριθμό χρηστών να αξιολογήσουν άμεσα τα αποτελέσματα συγκεκριμένων ερωτημάτων και όχι να πραγματοποιήσουν συνεδρίες αναζήτησης (search sessions) (και επιλογές αποτελεσμάτων) για ερωτήματα που τους ενδιαφέρουν, έτσι ώστε να αναπτυχθούν συγκεκριμένες συμπεριφορές αναζήτησης.

	MAP		**Mean NDCG**	
Folds	**B**	**M**	**B**	**M**
Fold1	0.68	0.70	0.70	0.70
Fold2	0.65	0.67	0.65	0.66
Fold3	0.64	0.65	0.66	0.66
Fold4	0.69	0.72	0.70	0.73
Fold5	0.66	0.67	0.68	0.68
Average	0.66	0.68	0.68	0.69

Πίνακας 4.3: Σύγκριση τιμών MAP και *Mean NDCG*

Ο Πίνακας 4.4 παρουσιάζει τις τιμές $P@n$ της μεθόδου μας, συγκρινόμενες με τις αντίστοιχες της βασικής μεθόδου, για κάθε διαμέριση. Όπως φαίνεται, η προτεινόμενη μέθοδος δίνει καλύτερες τιμές ακρίβειας για τις υψηλότερες θέσεις κατάταξης αποτελεσμάτων. Ειδικά για το $P@1$, η μέθοδός μας επιτυγχάνει μέγιστη αύξηση ακρίβειας 6% (4% μέση αύξηση). Αυτή η συμπεριφορά μπορεί να εξηγηθεί ως εξής: για έναν μεγάλο αριθμό ερωτημάτων, κάθε ερώτημα εμφανίζει υψηλή ομοιότητα με μία συγκεκριμένη συστάδα συμπεριφοράς αναζήτησης (και αρκετά μικρότερη ομοιότητα με τις υπόλοιπες συστάδες), κάτι που έχει ως αποτέλεσμα ένα μοντέλο που, για κάθε ερώτημα, ευνοεί τα πιο σχετικά του αποτελέσματα.

	Fold1		**Fold2**		**Fold3**		**Fold4**		**Fold5**		**Av. Folds**		**Max Increase**
Precision	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**M-B**
P@1	0.61	0.67	0.61	0.67	0.59	0.61	0.70	0.72	0.62	0.62	0.62	0.66	+0.06
P@2	0.60	0.60	0.52	0.54	0.53	0.54	0.67	0.67	0.58	0.58	0.58	0.58	+0.02
P@3	0.57	0.57	0.51	0.50	0.50	0.53	0.60	0.62	0.53	0.54	0.54	0.55	+0.03
P@4	0.54	0.53	0.48	0.48	0.48	0.49	0.56	0.57	0.49	0.50	0.51	0.51	+0.01
P@5	0.50	0.50	0.45	0.45	0.46	0.45	0.52	0.53	0.46	0.46	0.48	0.48	+0.01
P@6	0.48	0.47	0.43	0.43	0.43	0.43	0.49	0.49	0.42	0.42	0.45	0.45	-0.01
P@7	0.44	0.43	0.40	0.40	0.40	0.40	0.46	0.46	0.40	0.40	0.42	0.42	0.00
P@8	0.41	0.41	0.37	0.37	0.37	0.37	0.43	0.43	0.37	0.37	0.39	0.39	0.00
P@9	0.38	0.38	0.35	0.35	0.35	0.35	0.40	0.40	0.35	0.34	0.37	0.37	-0.01
P@10	0.36	0.36	0.33	0.33	0.33	0.33	0.38	0.38	0.33	0.33	0.34	0.34	0.00

Πίνακας 4.4: Σύγκριση τιμών $P@n$

Για χαμηλότερα στην κατάταξη αποτελέσματα, η μέθοδός μας έχει παρόμοια απόδοση με τη βασική μέθοδο. Αυτό συμβαίνει εξαιτίας του μικρού αριθμού (θετικών) κρίσεων σχετικότητας του σετ δεδομένων που περιορίζει τιμές τις ακρίβειας σε ορισμένο ύψος. Συγκεκριμένα,

ολόκληρο το σετ δεδομένων αποτελείται συνολικά από 563 ερωτήματα και 2932 θετικές κρίσεις σχετικότητας (2001 *μερικώς σχετικά* και 931 *πολύ σχετικά*). Έτσι, για κάθε ερώτημα υπάρχουν, κατά μέσο όρο, 5.2 θετικές κρίσεις σχετικότητας. Σε τέτοιου τύπου σετ δεδομένων, οι πιο ενδεικτικές τιμές απόδοσης μίας συνάρτησης ταξινόμησης είναι, προφανώς, οι τιμές ακρίβειας από $P@1$ έως $P@6$.

Ο Πίνακας 4.5 παρουσιάζει τις τιμές $NDCG@n$ για τις δύο προσεγγίσεις. Η συμπεριφορά είναι ανάλογη με τις τιμές του μέτρου $P@n$.

	Fold1		Fold2		Fold3		Fold4		Fold5		Av. Folds		Max Increase
NDCG	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**M-B**
NDCG@1	0.54	0.57	0.51	0.57	0.51	0.54	0.57	0.60	0.52	0.52	0.53	0.56	+0.06
NDCG@2	0.60	0.59	0.52	0.55	0.52	0.52	0.61	0.64	0.57	0.57	0.57	0.57	+0.03
NDCG@3	0.63	0.63	0.57	0.58	0.57	0.58	0.63	0.68	0.61	0.60	0.60	0.61	+0.05
NDCG@4	0.64	0.64	0.60	0.61	0.60	0.61	0.67	0.69	0.64	0.64	0.63	0.64	+0.02
NDCG@5	0.68	0.67	0.61	0.62	0.62	0.63	0.68	0.71	0.67	0.67	0.65	0.66	+0.03
NDCG@6	0.71	0.70	0.64	0.65	0.65	0.65	0.69	0.71	0.68	0.68	0.67	0.68	+0.02
NDCG@7	0.72	0.71	0.65	0.66	0.65	0.66	0.70	0.72	0.70	0.69	0.68	0.69	+0.02
NDCG@8	0.64	0.65	0.63	0.63	0.63	0.63	0.63	0.65	0.65	0.65	0.64	0.64	+0.02
NDCG@9	0.31	0.31	0.24	0.25	0.34	0.33	0.37	0.38	0.29	0.29	0.31	0.31	+0.01
NDCG@10	0.32	0.32	0.25	0.25	0.35	0.34	0.38	0.39	0.30	0.30	0.32	0.32	+0.01

Πίνακας 4.5: Σύγκριση τιμών $NDCG@n$

4.4 Μοντέλα ταξινόμησης βασιζόμενα στο σκοπό αναζήτησης

Σε αυτήν την ενότητα παρουσιάζουμε την επέκταση της προηγουμένως περιγεγραμμένης μεθόδου, κατά την οποία στηριζόμαστε απευθείας στα επιμέρους διανύσματα-μοντέλα κάθε ερωτήματος, προκειμένου να τα ομαδοποιήσουμε και να παράγουμε συστάδες ερωτημάτων που αντιπροσωπεύουν παρόμοιες συμπεριφορές-σκοπούς αναζήτησης.

4.4.1 Προαπαιτούμενες γνώσεις

Για να περιγράψουμε καλύτερα το μαθηματικό μοντέλο πίσω από την προτεινόμενη μέθοδο, ξαναπαρουσιάζουμε σύντομα κάποιες βασικές έννοιες των μηχανών διανυσμάτων στήριξης.

Έστω ότι έχουμε n ερωτήματα $q_1, \ldots, q_n$ από το ιστορικό αναζήτησης και τα top-m ανακτημένα έγγραφα $(x_1^{(q)}, y_1^{(q)}), \ldots, (x_m^{(q)}, y_m^{(q)})$, όπου $y_j^{(q)} = 1$ αν το $x_j^{(q)}$ έχει πατηθεί (επιλεγεί) και $y_j^{(q)} = 0$ αν όχι. Η ανάδραση του αν πατήθηκε ή όχι ένα αποτέλεσμα, συνεπάγεται μία μερική κατάταξη των εγγράφων-αποτελεσμάτων, έτσι ώστε να ισχύει:

$$x_i^{(q)} \text{ προτιμάται από το } x_j^{(q)} \;\Leftrightarrow\; y_i^{(q)} > y_j^{(q)}$$

Συγκεντρώνουμε όλες τις σχέσεις προτίμησης για κάθε ερώτημα q στο σύνολο $\mathcal{P}_q = \{(i,j) : y_i^{(q)} > y_j^{(q)}\}$, ακολουθώντας τη μέθοδο των [2, 3]. Με αυτόν τον τρόπο, είναι δυνατόν να προσαρμοστεί μία συνάρτηση ταξινόμησης $f : (q,x) \mapsto \mathbb{R}$ στις ανά ζεύγος προτιμήσεις $\mathcal{P} = \bigcup_q \mathcal{P}_q$. Σε αυτή τη δουλειά περιοριζόμαστε σε γραμμικά μοντέλα της μορφής

$$f(q,x) = \langle \vec{w}, \phi(q,x) \rangle,$$

όπου το $\phi(q, x)$ αναπαριστά μία συνδυαστική αναπαράσταση ερωτήματος -αποτελέσματος στον χώρο χαρακτηριστικών. Για απλούστευση των συμβολισμών, στο εξής θα χρησιμοποιούμε το $\phi(q, x) = x$. Ακολουθώντας μία προσέγγιση μεγάλων περιθωρίων, οδηγούμαστε στο ακόλουθο πρόβλημα βελτιστοποίησης ([48]):

$$\begin{aligned} \min_{\vec{w}, \xi_{ij} \geq 0} \quad & \langle \vec{w}, \vec{w} \rangle + \lambda \sum_{ij} \xi_{ij} \\ \text{έτσι ώστε} \quad & \forall (i,j) \in \mathcal{P} : \langle \vec{w}, x_i \rangle \geq \langle \vec{w}, x_j \rangle + 1 - \xi_{ij}, \end{aligned}$$

όπου το $\lambda > 0$ καθορίζει την βαρύτητα μεταξύ του κόστους μεγιστοποίησης περιθωρίου και ελαχιστοποίησης σφάλματος. Ο δεύτερος παράγοντας είναι το άθροισμα των επιμέρους σφαλμάτων-απωλειών ξ_{ij} και αποτελεί ένα πάνω φράγμα στην 0/1-απώλεια που προέρχεται από λανθασμένες σχέσεις προτίμησης. Οι περιορισμοί επιβάλλουν ότι $\langle \vec{w}, x_i \rangle > \langle \vec{w}, x_j \rangle$, όταν είναι δυνατόν και τιμωρούν αποκλίσεις από τη συνθήκη. Όταν βρεθούν οι βέλτιστες παράμετροι $\vec{w}^*$, χρησιμοποιούνται ως βοηθητικές εκτιμήσεις ώστε να προσεγγιστούν οι ταξινομήσεις εγγράφων για νέα ερωτήματα.

Στη συνέχεια, μετράμε την ποιότητα των συναρτήσεων αναταξινόμησης αξιολογώντας το $\hat{y}_i = f(q, x_i)$ για ένα νέο ερώτημα q. Έστω τα y_i και $\hat{y}_j$ ταξινομημένα και π μία μετάθεση πάνω στα $\{1, \ldots, m\}$, έτσι ώστε το $\pi(1)$ να αναπαριστά την υψηλότερα ταξινομημένη πρόβλεψη, δηλαδή, $\pi(1) = j : \hat{y}_j > \hat{y}_i \, \forall i \neq j$, και το $\pi(2)$ την αμέσως επόμενη, κτλ. Υπολογίζουμε την ακρίβεια στην θέση n, τη μέση ακρίβεια (mean average precision - MAP) και το κανονικοποιημένο αθροιστικό κέρδος (normalized discount cumulative gain - NDCG@n) ως εξής:

$$\begin{aligned} P@n &= \frac{\sum_{i=1}^{n} y_{\pi(i)}}{n} \\ MAP &= \frac{\sum_{i=1}^{n} P@i \times y_{\pi(i)}}{\sum_{i=1}^{n} y_i} \\ NDCG@n &= \frac{1}{Z_n} \sum_{i=1}^{n} \begin{cases} 2^{\pi(i)} - 1 & i = 1 \\ \frac{2^{\pi(i)}-1}{\log i} & i > 1 \end{cases} \end{aligned}$$

όπου το Z_n είναι μία σταθερά κανονικοποίησης που υπολογίζεται χρησιμοποιώντας την ταυτότητα $\pi(i) = i$, ώστε να εξασφαλιστεί ότι NDCG@$n = 1$ για μία ιδανική πρόβλεψη ταξινόμησης.

4.4.2 Συνδυαστική βελτιστοποίηση

Στόχος μας είναι η εκπαίδευση συναρτήσεων ταξινόμησης για *παρόμοια* ερωτήματα, όπου ο χαρακτηρισμός «παρόμοια» αναφέρεται στον υπολανθάνοντα σκοπό αναζήτησης του χρήστη. Η βασική ιδέα είναι η ακόλουθη: Η συμπεριφορά αναζήτησης και επιλογής αποτελεσμάτων από το χρήστη μάς δίνει μία, πιθανά «θορυβώδη», άποψη του λανθάνοντος σκοπού αναζήτησης. Ένα μοντέλο ταξινόμησης στοχεύει στην αναγνώριση αυτής της συμπεριφοράς και, σε ένα βαθμό, του ίδιου του σκοπού αναζήτησης. Για αυτό το λόγο, προτείνουμε τη συσταδοποίηση

τέτοιων μοντέλων ταξινόμησης, αφού παρόμοιες συναρτήσεις ταξινόμησης αντιστοιχούν σε παρόμοιες συμπεριφορές αναζήτησης και, επαγωγικά, σκοπούς αναζήτησης. Κάθε συστάδα αντιπροσωπεύεται, τελικά, από μία ξεχωριστή συνάρτηση ταξινόμησης, η οποία εκπαιδεύεται στα δεδομένα αναζήτησης των ερωτημάτων της συστάδας.

Στη συνέχεια, παρουσιάζουμε ένα πρόβλημα συνδυαστικής βελτιστοποίησης που λύνει, θεωρητικά, το πρόβλημα της προηγούμενης παραγράφου. Στόχος είναι μία συμπαγής διαδικασία συσταδοποίησης, η οποία θα οδηγεί σε μοντέλα ταξινόμησης που θα αναπαριστούν διαφορετικούς σκοπούς αναζήτησης. Αφού δεν υπάρχει κάποια βασική αλήθεια για τον έλεγχο της συσταδοποίησης, το τελικό σφάλμα ταξινόμησης των συναρτήσεων ταξινόμησης χρησιμεύει ως έμμεσο μέτρο αποτίμησης της αποτελεσματικότητας του σταδίου συσταδοποίησης. Δηλαδή, αν το ποσοστό σφαλμάτων μίας συνάρτησης αναταξινόμησης είναι μεγάλο, τότε τα ερωτήματα που περιέχονται στην αντίστοιχη συστάδα είναι πολύ ετερογενή, καθιστώντας αναποτελεσματική τη συστάδα. Για αυτό, σκοπός είναι να βρεθεί η κατάλληλη ομαδοποίηση (συσταδοποίηση) ώστε τα αντίστοιχα εκπαιδευόμενα μοντέλα πάνω στις συστάδες να λειτουργούν αποτελεσματικά. Η πιο λογική προσέγγιση για τη λύση του προβλήματος είναι η συνδυαστική βελτιστοποίηση των προβλημάτων της συσταδοποίησης και της εκπαίδευσης συναρτήσεων ταξινόμησης.

Έστω K ο επιθυμητός αριθμός παραγόμενων συστάδων. Στοχεύουμε να βρούμε (α) K μοντέλα ταξινόμησης, $\vec{w}_1, \ldots, \vec{w}_K$, ένα για κάθε συστάδα και (β) μία συσταδοποίηση $\vec{c}_1, \ldots, \vec{c}_K$, με $c_{kj} = 1$ αν το ερώτημα q_j ανήκει στη συστάδα k και $c_{kj} = 0$ διαφορετικά, το οποίο επιτυγχάνει βέλτιση προσαρμογή των μοντέλων ταξινόμησης. Το παραπάνω πρόβλημα κωδικοποιείται ως εξής:

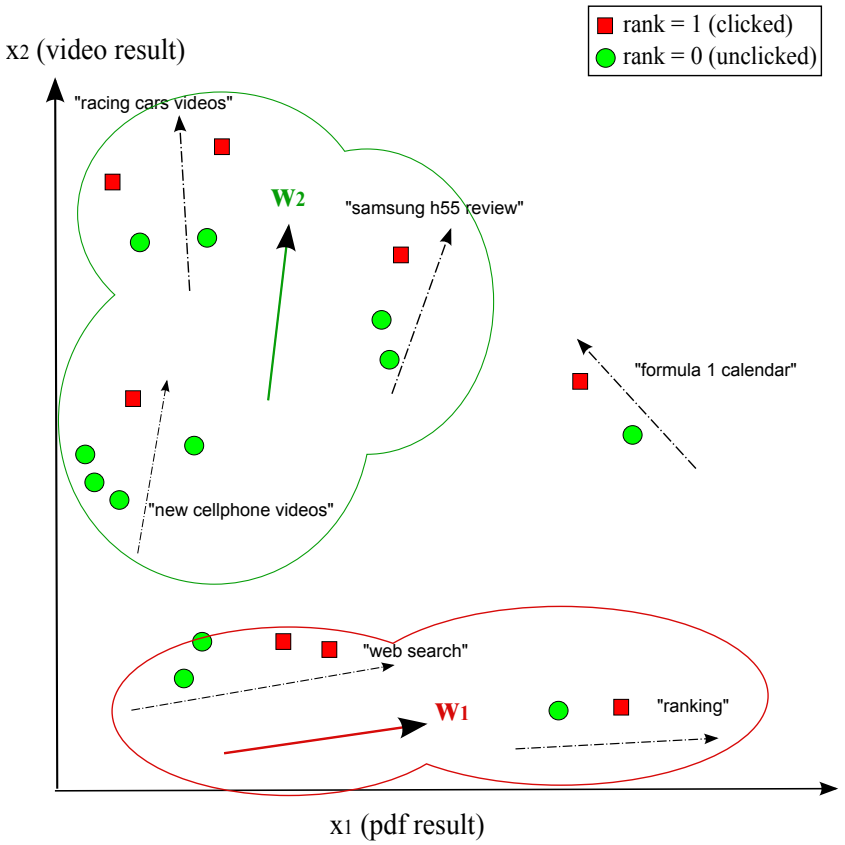
$$
\begin{aligned}
\min_{\vec{w}_k, \vec{c}_k, \xi_{ij}} \quad & \sum_{k=1}^{K} \left[\|\vec{w}_k\|^2 + \lambda_k \sum_{\ell=1}^{n} c_{k\ell} \sum_{(i,j)\in\mathcal{P}_{q_\ell}} \xi_{ij}^k \right] \\
\text{ς.τ.} \quad & \forall k, \forall (i,j) \in \mathcal{P}(k) : \langle \vec{w}_k, x_i \rangle \geq \langle \vec{w}_k, x_j \rangle + 1 - \xi_{ij}^k \\
& \forall k, \forall (i,j) \in \mathcal{P}(k) : \xi_{ij}^k \geq 0 \\
& \forall i, j, \ell : \; c_{ki} c_{kj} + c_{ki} c_{k\ell} \leq c_{kj} c_{k\ell} + 1 \\
& \forall k, \forall j : \; c_{kj} \in \{0, 1\}
\end{aligned}
\tag{4.9}
$$

όπου ορίζουμε το $\mathcal{P}(k) = \bigcup_{j:c_{kj}=1} \mathcal{P}_{q_j}$ ως την ένωση όλων των στοιχείων της συστάδας k, και παραμέτρους ρύθμισης βαρύτητας $\lambda_k > 0$.

Το παραπάνω πρόβλημα βελτιστοποίησης ομαδοποιεί παρόμοια ζεύγη ερωτημάτων - αποτελεσμάτων, όπου η ομοιότητα αναφέρεται στη θέση ομάδων των ζευγών που αναφέρονται στο ίδιο ερώτημα στον χώρο χαρακτηριστικών. Δηλαδή, σημασία δεν έχει η απόλυτη θέση των σημείων, αλλά η κατεύθυνση των διανυσμάτων που ορίζουν στο χώρο.

Η Εικόνα 4.4 δείχνει μία απλή δισδιάστατη αναπαράσταση του περιγεγραμμένου σεναρίου, επικεντρώνοντας σε pdf αποτελέσματα (διάσταση x_1) και βίντεο αποτελέσματα (διάσταση x_2). Διαφορετικά ερωτήματα (για παράδειγμα *racing cars videos*, *web search*) απεικονίζονται μέσω σχετικών-επιλεγμένων (κόκκινα τετράγωνα) και άσχετων-μη επιλεγμένων (πράσινοι κύκλοι) εγγράφων. Η απαιτούμενη εργασία είναι η ομαδοποίηση ερωτημάτων έτσι ώστε πα-

ρόμοιοι σκοποί αναζήτησης να βρίσκονται κοντά στο χώρο χαρακτηριστικών, με βάση κάποια συνάρτηση ομοιότητας, έτσι ώστε να βρεθούν στις ίδιες συστάδες. Η απόλυτη απόσταση μεταξύ των διανυσμάτων-συναρτήσεων ταξινόμησης είναι άνευ σημασίας, σε αντίθεση με τις διαφορές στην κατεύθυνσή τους, οι οποίες ποσοτικοποιούν τις διαφορές στη συμπεριφορά αναζήτησης: μοντέλα με μικρές διαφορές στη γωνία των διανυσμάτων τους αναμένεται να ευνοούν τους ίδιους τύπους αποτελεσμάτων. Για παράδειγμα, η συνάρτηση ταξινόμησης $\vec{w}_1$ ταξινομεί pdf αποτελέσματα υψηλότερα από τη $\vec{w}_2$ η οποία προφανώς ευνοεί βίντεο αποτελέσματα. Το πρόβλημα βελτιστοποίησης στοχεύει στην εύρεση της λύσης συσταδοποίησης και των συναρτήσεων ταξινόμησης $\vec{w}_i$ ταυτόχρονα.



Σχήμα 4.4: Απεικόνιση του προβλήματος. Δεδομένων ερωτημάτων και των αντίστοιχων (επιλεγμένων και μη) αποτελεσμάτων, το πρόβλημα βελτιστοποίησης στοχεύει στην απευθείας εύρεση συναρτήσεων ταξινόμησης $\vec{w}_1$, $\vec{w}_2$, και $\vec{w}_3$ του σκοπού αναζήτησης. Η προσεγγιστική λύση της Ενότητας 4.4.3 σπάει το πρόβλημα σε τρία βήματα: εκπαίδευση συνάρτησης ταξινόμησης για κάθε ερώτημα, συσταδοποίηση παρόμοιων ερωτημάτων και εκπαίδευση συνάρτησης ταξινόμησης για κάθε συστάδα.

Παρόλα αυτά, το παραπάνω πρόβλημα έχει κάποια ουσιώδη μειονεκτήματα. Πρώτον, η βελτιστοποίηση συνδυάζει πραγματικές και ακέραιες μεταβλητές, με αποτέλεσμα η απευθείας λύση του προβλήματος να είναι ακριβή και να είναι αναγκαία η χαλάρωση των δυαδικών μεταβλητών στο διάστημα $[0, 1]$, για να επιτευχθεί μία προσεγγιστική λύση. Δεύτερον, και κυρίως, ο αριθμός των τριγωνικών ανισοτήτων που θα εγγυηθούν μία κατάλληλη λύση στην

Πίνακας 4.6: Μοντελα ταξινομησης βασιζομενα στο σκοπο αναζητησης

Require: n queries q_j with preference relations $\mathcal{P}_{q_j}$

1: **for** $1 \leq j \leq n$ **do**
2: learn ranking function $\vec{w}_j$ for q_j using $\mathcal{P}_{q_j}$
3: **end for**
4: cluster $w_1, \ldots, w_n$
5: **for** $1 \leq k \leq K$ **do**
6: learn ranking function $\vec{w}_k$ using $\bigcup_{j:c_j=k} \mathcal{P}_{q_j}$
7: **end for**

Ensure: ranking models $\vec{w}_1, \ldots, \vec{w}_K$

Εξίσωση (4.9) είναι κυβικός ως προς τον αριθμό των ερωτημάτων και καθιστά το πρόβλημα ανέφικτο σε μεγάλες κλίμακες. Για αυτό, καταφεύγουμε σε μία αποδοτική προσεγγιστική λύση και προτείνουμε μία σειριακή μέθοδο που την υλοποιεί στην επόμενη ενότητα.

4.4.3 Εκπαίδευση του σκοπού αναζήτησης

Σε αυτήν την ενότητα παρουσιάζουμε την προτεινόμενη ακολουθιακή μέθοδο, η οποία προσεγγίζει το προηγουμένως περιγεγραμμένο πρόβλημα, καθιστώντας τη λύση του αποδοτική για μεγάλα μεγέθη δεδομένων. Η μέθοδός μας αποτελείται από τρία βήματα: Εκπαίδευση μοντέλων-συναρτήσεων ταξινόμησης ανά ερώτημα, συσταδοποίηση των αντιστοίχων διανυσμάτων των μοντέλων και εκαίδευση συναρτήσεων ταξινόμησης ανά συστάδα. Ο ψευδοκώδικας του αλγορίθμου δίνεται στον Πίνακα 4.6.

Εκπαίδευση μοντέλων ταξινόμησης ανά ερώτημα

Το αρχικό βήμα της διαδικασίας είναι η εκπαίδευση ενός μοντέλου ταξινόμησης ανά ερώτημα. Προκειμένου να το επιτύχουμε, λύνουμε το βασικό πρόβλημα SVM για ταξινόμηση (Ranking SVM) μόνο στα δεδομένα ανάδρασης χρήστη για κάθε ερώτημα (ερώτημα, λίστα αποτελεσμάτων, λίστα επιλεγμένων αποτελεσμάτων). Αναλογικά με την Ενότητα 4.4.1, το ℓ-οστό πρόβλημα βελτιστοποίησης μπορεί να λυθεί είτε με τετραγωνικό προγραμματισμό (quadratic programming) ή μεθόδους βασισμένες στην κλίση (online gradient-based) [2, 3, 48] και αναπαρίσταται ως εξής:

$$\begin{aligned} \min_{\vec{w}_\ell, \xi_{\ell ij} \geq 0} \quad & \langle \vec{w}_\ell, \vec{w}_\ell \rangle + \lambda \sum_{ij} \xi_{\ell ij} \\ \text{ς.τ.} \quad & \forall (i,j) \in \mathcal{P}_{q_\ell} : \langle \vec{w}_\ell, x_i \rangle \geq \langle \vec{w}_\ell, x_j \rangle + 1 - \xi_{\ell ij}. \end{aligned}$$

Γενικά, η παράμετρος ρύθμισης λ χρειάζεται να τεθεί κατάλληλα έτσι ώστε να παραχθούν τα βέλτιστα μοντέλα. Σε μεγάλης έκτασης εφαρμογές της μεθόδου, η πειραματική εύρεση μίας βέλτιστης τιμής για την παράμετρο δεν είναι εφικτή, εξαιτίας του μεγάλου όγκου δεδομένων.

Εμπειρικά στοιχεία, όμως, δείχνουν ότι για χαρακτηριστικά με δυαδικές τιμές ή τιμές στο διάστημα $[0, 1]$, η τιμή $\lambda \approx 1$ είναι μία λογική επιλογή. Έτσι, θέτουμε $\lambda = 1$ για τα αρχικά μοντέλα Ranking SVM, σημειώνοντας ότι ενδέχεται να υπάρχει περιθώριο για βελτίωση, με τη βελτιστοποίηση της παραμέτρου λ. Το αποτέλεσμα αυτού του βήματος είναι n συναρτήσεις ταξινόμησης $\vec{w}_1, \ldots, \vec{w}_n$, μία για κάθε ερώτημα.

Ομαδοποιώντας συναρτήσεις ταξινόμησης

Σκοπός του δευτέρου βήματος είναι η ομαδοποίηση παρόμοιων μοντέλων ταξινόμησης με στόχο την ομαδοποίηση παρόμοιων σκοπών αναζήτησης. Αφού οι απόλυτες θέσεις των διανυσμάτων είναι άνευ σημασίας και μόνο η κατεύθυνσή τους μετράει, τα διανύσματα κανονικοποιούνται με τη νόρμα ℓ_2 ως εξής: $\vec{w} \leftarrow \vec{w}/\|\vec{w}\|$, ώστε να βρίσκονται όλα στη μοναδιαία υπερσφαίρα. Η ομοιότητα μεταξύ δύο διανυσμάτων μπορεί τώρα να μετρηθεί μέσω της συνάρτησης συνημιτόνου, η οποία καταλήγει στο εσωτερικό γινόμενό τους, για κανονικοποιημένα διανύσματα:

$$\cos(\vec{w}, \vec{w}') = \langle \vec{w}, \vec{w}' \rangle.$$

Τα μοναδιαία διανύσματα συνήθως μοντελοποιούνται από μία κατανομή von Mises-Fisher ([47]), η οποία δίνεται από το:

$$p(\vec{x}|\vec{\mu}, \kappa) = Z_d(\kappa) \exp\{\kappa \langle \vec{\mu}, \vec{x} \rangle\}$$

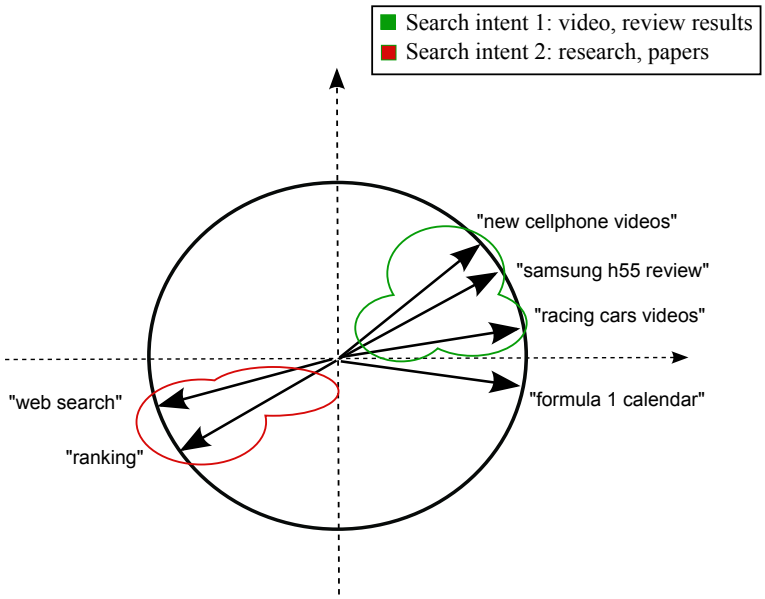
όπου $\|\vec{\mu} = 1\|$, $\kappa \geq 0$ και $d \geq 2$ και η συνάρτηση διαμέρισης $Z_d(\kappa) = \kappa^{d/2-1}/(2\pi)^{d/2} I_{d/2-1}(\kappa)$ όπου το $I_r(\cdot)$ δηλώνει την τροποποιημένη συνάρτηση Bessel πρώτου είδους και τάξης r. Εφαρμοζόμενο σε n συναρτήσεις ταξινόμησης $\vec{w}_1, \ldots, \vec{w}_n$, ένα αναμεμειγμένο μοντέλο από κατανομές von Mises-Fisher με K συστατικά (συστάδες) έχει πυκνότητα:

$$f(\vec{w}_i|\vec{\mu}_1, \ldots, \vec{\mu}_K, \vec{\kappa}) = \sum_{i=1}^{n} \alpha_{c_i} p(\vec{w}_i|\vec{\mu}_{c_i}, \kappa_{c_i})$$

$$\log p(\vec{w}_1, \ldots, \vec{w}_n, \vec{c}|\Theta) = \sum_{i=1}^{n} \log \left[\alpha_{c_i} p(\vec{w}_i|\vec{\mu}_{c_i}, \kappa_{c_i})\right]$$

όσον αφορά στις παραμέτρους $\Theta = \{(\vec{\mu}_k, \kappa_k)\}_{k=1}^{K}$ και τις λανθάνουσες μεταβλητές $\vec{c} = (c_1, \ldots, c_n)^\top$. Η διαδικασία EM (μεγιστοποίηση της προσδοκίας - expectation maximization) εκτιμά επαναληπτικά την κατανομή των λανθανουσών μεταβλητών $\hat{p}(c_i|\vec{w}_i, \Theta)$ και παραμέτρων $\hat{p}(\Theta|\vec{w}_1, \ldots, \vec{w}_n, \vec{c})$, θεωρώντας κάθε φορά τη μη εξεταζόμενη ποσότητα από τις δύο, ως σταθερή. Προκειμένου να επιτευχθεί αποδοτικός υπολογισμός και να παραχθούν αποτελέσματα με νόημα, εφαρμόζουμε την επονομαζόμενη ευριστική αυστηρής ανάθεσης (hard assignment heuristic) κατά τη βελτιστοποίηση, όπου η κατανομή των λανθανουσών μεταβλητών αντικαθίσταται από μία ψήφο πλειοψηφίας (majority vote), δηλαδή:

$$\hat{p}(c_i|\vec{w}_i, \Theta) = \begin{cases} 1 : c_i = \text{argmax}_k \, p(k|\vec{w}_i, \Theta) \\ 0 : \text{otherwise.} \end{cases} \tag{4.10}$$



Σχήμα 4.5: Εκπαιδευμένα, ανά ερώτημα, διανύσματα πάνω στη μοναδιαία υπερσφαίρα.

Εκπαίδευση μοντέλων ταξινόμησης για συστάδες

Δεδομένης της συσταδοποίησης που παράχθηκε μέσω των λανθανουσών μεταβλητών c_i στην προηγούμενη ενότητα, σε αυτό το βήμα εκπαιδεύουμε μία συνάρτηση ταξινόμησης για κάθε συστάδα. Η προσέγγιση είναι παρόμοια με την αρχική εκπαίδευση συναρτήσεων ανά ερώτημα, απλά σε αυτό το στάδιο, για κάθε εκπαίδευση, συμμετέχουν όλα τα ερωτήματα μίας συστάδας. Η βελτιστοποίηση για την k-οστή συστάδα μπορεί να λυθεί πάλι με τη μέθοδο Ranking SVM και δίνεται από το:

$$\begin{aligned} \min_{\vec{w}_k, \xi_{ij} \geq 0} \quad & \langle \vec{w}_k, \vec{w}_k \rangle + \lambda \sum_{ij} \xi_{ij} \\ \text{ς.τ.} \quad & \forall (i,j) \in \bigcup_{\ell : c_\ell = k} \mathcal{P}_{q_\ell} : \langle \vec{w}_k, x_i \rangle \geq \langle \vec{w}_k, x_j \rangle + 1 - \xi_{ij}. \end{aligned}$$

4.4.4 Εφαρμογή των εκπαιδευμένων μοντέλων

Αφού οι συναρτήσεις ταξινόμησης εκπαιδευτούν πάνω στις συστάδες, η μέθοδός μας μπορεί να εφαρμοστεί για αναταξινόμηση ανακτημένων αποτελεσμάτων για νέα ερωτήματα. Η μέθοδός μας στοχεύει στην επίτευξη ετερογένειας μεταξύ διαφορετικών σκοπών αναζήτησης, καθώς, το ίδιο ερώτημα μπορεί να καταλήξει σε περισσότερες από μία συστάδες αν, για παράδειγμα, οι χρήστες επιλέγουν διαφορετικούς τύπους αποτελέσματος για το ίδιο ερώτημα. Έτσι, ο σκοπός είναι η αντιστοίχιση ενός νέου ερωτήματος σε «παρόμοιες» συστάδες και η παραγωγή μίας συνδυαστικής ταξινόμησης αποτελεσμάτων, από τις επιμέρους ταξινομήσεις που παράγει το μοντέλο ταξινόμησης κάθε συστάδας.

Για το σκοπό αυτό, αναπαριστούμε τα προηγούμενα ερωτήματα, μαζί με τα επιλεγμένα αποτελέσματά τους ως ψευδο-έγγραφα, τα οποία ευρετηριάζονται από μία μηχανή αναζήτησης.

Στην υλοποίησή μας χρησιμοποιήσαμε το Lucene[7] IR engine, ενώ με τον ίδιο τρόπο ακριβώς θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε άλλη αντίστοιχη μηχανή αναζήτησης. Δεχόμενη ένα νέο ερώτημα q, η μηχανή βαθμολόγησης αποτελεσμάτων του Lucene χρησιμοποιείται για να ανακτήσει προηγούμενα ερωτήματα που είναι παρόμοια με το q.

Επιλέγουμε τα *top-u* πιο όμοια ερωτήματα και τις αντίστοιχες συστάδες στις οποίες ανήκουν. Με αυτόν τον τρόπο, δημιουργούμε μία σταθμισμένη αντιστοίχιση από το νέο ερώτημα στις συστάδες, ως εξής: Έστω v_j, $1 \leq j \leq u$, τα σκορ για τα *top-u* προηγούμενα ερωτήματα q_j. Τότε, αυτά κανονικοποιούνται με τη νόρμα ℓ_1 και μεταφράζονται σε σκορ συστάδων s_k, $1 \leq k \leq K$, τέτοια ώστε:

$$s_{qk} = \sum_{j:c_j=k} v_j \; / \; \sum_{i=1}^{u} v_i$$

όπου το c_j είναι η λανθάνουσα συμμετοχή στη συστάδα. Δηλαδή, αν μία συστάδα προκύπτει παραπάνω από μία φορά, τα σκορ της αθροίζονται. Λόγω της κανονικοποίησης, τα σκορ s_{qk} λειτουργούν ως πιθανότητες, ποσοτικοποιώντας την πιθανότητα ότι μία συστάδα k περιέχει το σκοπό αναζήτησης που εκφράζεται από το ερώτημα q.

Τελικά, η ταξινόμηση των εγγράφων για το ερώτημα q παράγεται σταθμίζοντας τη συνεισφορά κάθε συστάδας k με το αντίστοιχο σκορ της s_{qk}. Έστω ότι το r_{kj} δηλώνει την ταξινόμηση του j-οστού εγγράφου από τη συνάρτηση ταξινόμησης της συστάδας k. Τότε, το τελικό σκορ δίνεται από τη γραμμική στάθμιση των επιμέρους ταξινομήσεων r_{kj} με τη σημασία-σκορ της συστάδας s_{qk}, για το ερώτημα q:

$$score(q, j) = \sum_{k=1}^{K} s_{qk} r_{qkj}$$

Με άλλα λόγια, οι επιμέρους ταξινομήσεις εφαρμόζονται σε βαθμό ανάλογο με την ομοιότητα της αντίστοιχης συστάδας με το νέο ερώτημα.

4.5 Αξιολόγηση αποτελεσμάτων

Σε αυτήν την ενότητα αξιολογούμε την αποτελεσματικότητα της προτεινόμενης μεθόδου. Η Υποενότητα 4.5.1 παρουσιάζει το πειραματικό σύνολο δεδομένων και την προεπεξεργασία που πραγματοποιήθηκε. Η Υποενότητα 4.5.2 παρουσιάζει τις βασικές - ανταγωνιστικές μεθόδους με τις οποίες συγκρινόμαστε και η 4.5.3 παρουσιάζει τα αποτελέσματα. Συγκεκριμένα, αναλύουμε την ποιότητα της παραγόμενης συσταδοποίησης, την αποτελεσματικότητά της σε σχέση με το χρονικό διάστημα εκπαίδευσης - εφαρμογής της μεθόδου και, τέλος, συζητάμε τα συμπεράσματα.

4.5.1 Δεδομένα εκπαίδευσης και προεπεξεργασία

Για την πειραματική αξιολόγηση συγκεντρώσαμε δειγματοληπτικά ερωτήματα από το αρχείο καταγραφής ερωτημάτων της μηχανής αναζήτησης της Yahoo!. Από το δείγμα, απορρίψαμε

[7] http://lucene.apache.org/

Πίνακας 4.7: Κατηγορίες χαρακτηριστικών εκπαίδευσης

	Textual similarity features
4	Sum of TFs of query terms in result title—URL—text—all
4	Lucene score between query and result title—URL—text—all
	Result characteristics features
1	Result initial rank
4	Number of words in result title—url—text—all
1	Result URL length in characters
72	Result URL domain (boolean values)
83	Popular sites the result might belong to (boolean)
200	Top most frequent urls in the dataset
	Result special words features
10	Special words in result URL ("forum", "pdf", etc.)
10	Result site category (news, search, blog etc)
200	Top most frequent words in the dataset

ερωτήματα με λιγότερα από πέντε αποτελέσματα, ερωτήματα για τα οποία ο χρήστης δεν πάτησε (επέλεξε) κανένα αποτέλεσμα και ερωτήματα από χρήστες με λιγότερες από 100 αναζητήσεις συνολικά. Αυτό κατέληξε σε 76037 ερωτήματα από το αρχείο καταγραφής, τα οποία τέθηκαν από συνολικά 453 ξεχωριστούς χρήστες. Στη συνέχεια, χωρίσαμε το σύνολο δεδομένων (ερωτήματα μαζί με τα 10 πρώτα αποτελέσματα για το καθένα), χρονολογικά, σε 30053 (40%) ερωτήματα για εκπαίδευση του συστήματος και 45984 (60%) ερωτήματα ως σετ αξιολόγησης.

Η βασική αλήθεια δίνεται από τα πατήματα-επιλογές των χρηστών σε αποτελέσματα, με τη μορφή κρίσεων σχετικότητας ([2, 3]). Ειδικότερα, οι κρίσεις σχετικότητας ανατίθενται ως εξής: Αν ένα έγγραφο x_i έχει επιλεγεί, η κρίση σχετικότητας τίθεται $y_i = 1$. Μη επιλεγμένα έγγραφα που ταξινομούνται πιο ψηλα από επιλεγμένα, παίρνουν κρίση $y_j = 0$ η οποία επίσης τίθεται και στο μη επιλεγμένο αποτέλεσμα που βρίσκεται ακριβώς από κάτω (στην κατάταξη) από το τελευταίο επιλεγμένο. Αυτή η διαδικασία παρήγαγε ένα σύνολο από 96030 κρίσεις σχετικότητας για το σύνολο εκπαίδευσης και 144021 κρίσεις για το σύνολο αξιολόγησης. Τα παραπάνω μεγέθη δίνουν ένα μέσο όρο 3.2 κρίσεων σχετικότητας ανά ερώτημα στο πειραματικό σύνολο δεδομένων που χρησιμοποιήσαμε.

Τα ζεύγη ερωτήματος - αποτελέσματος αναπαρίστανται ως διανύσματα χαρακτηριστικών που περιγράφουν την ποιότητα του κάθε αποτελέσματος, όσον αφορά στο αντίστοιχο ερώτημα. Συχνά, τα χαρακτηριστικά αυτά βασίζονται στο περιεχόμενο και περιγράφουν εκφάνσεις κειμενικής ομοιότητας μεταξύ του ερωτήματος και του εγγράφου [27]. Άλλες κατηγορίες χαρακτηριστικών περιγράφουν πληροφορία υπερσύνδεσης εγγράφων-σελίδων, όπως ο page-rank [8], ή μεταδεδομένα του εγγράφου-σελίδας όπως το domain του url του ή η ταξινόμησή του από καθιερωμένες μηχανές αναζήτησης [2]. Τέλος, άλλα χαρακτηριστικά μπορεί να κωδικοποιούν στατιστική πληροφορία για τη συμπεριφορά των χρηστών, όπως, για παράδειγμα, την απόκλιση του χρόνου επισκόπησης ενός αποτελέσματος σε σχέση με μία μέση τιμή [9]. Το σύ-

νολο χαρακτηριστικών που χρησιμοποιούμε υλοποιεί διάφορες από τις παραπάνω προσεγγίσεις και παρουσιάζεται στον Πίνακα 4.7.

4.5.2 Βασικές μέθοδοι σύγκρισης

Συγκρίνουμε τη μέθοδό μας, η οποία συμβολίζεται ως *Intent* με τέσσερις εναλλακτικές μεθόδους για αναταξινόμηση αποτελεσμάτων: Πρώτα, θεωρούμε μία μοναδική/συνολική συνάρτηση ταξινόμησης Ranking SVM (*Single*) για όλους τους χρήστες, η οποία εκπαιδεύεται σε ολόκληρο το σύνολο εκπαίδευσης και χρησιμοποιείται για την αναταξινόμηση των αποτελεσμάτων όλων των ερωτημάτων αξιολόγησης. Δεύτερον, εκπαιδεύουμε μία συνάρτηση ταξινόμησης για κάθε χρήστη (*User*) αντιπροσωπεύοντας την ιδανική περίπτωση εξατομίκευσης, όπου μπορούμε να βρούμε, για κάθε χρήστη, ακριβώς τα ερωτήματα που έθεσε και το αντίστροφο. Αυτές οι δύο βασικές μέθοδοι αποτελούν το κάτω και πάνω όριο στην αποτελεσματικότητα μίας μεθόδου αναταξινόμησης αποτελεσμάτων. Σύμφωνα με το [5], τα βραχυπρόθεσμα και μακροπρόθεσμα ιστορικά αναζήτησης αιχμαλωτίζονται/αντιπροσωπεύονται αποτελεσματικά από εξατομικευμένα, εξειδικευμένα στους χρήστες μοντέλα και, για αυτό, περιμένουμε η βασική μέθοδος *User* να έχει τη μεγαλύτερη αποτελεσματικότητα, ενώ η *Single* αναμένεται να είναι πολύ απλοϊκή για να αιχμαλωτίσει ετερογενείς συμπεριφορές από τα δεδομένα.

Επιπλέον, εξετάζουμε την προσέγγιση *Content-1* η οποία ομαδοποιεί ερωτήματα του συνόλου εκπαίδευσης με βάση το κειμενικό περιεχόμενό τους και εκπαιδεύει ένα μοντέλο Ranking SVM για κάθε συστάδα, τα οποία μοντέλα τελικά συνδυάζονται για να αναταξινομήσουν αποτελέσματα των ερωτημάτων αξιολόγησης. Σημειώνουμε ότι, εκτός από τη λογική/διαδικασία συσταδοποίησης, η υπόλοιπη επεξεργαστική ροή της μεθόδου είναι ακριβώς η ίδια με της μεθόδου μας. Στο στάδιο συσταδοποίησης, τα ερωτήματα ομαδοποιούνται με βάση την κειμενική τους ομοιότητα, συνυπολογίζοντας το κείμενο από τα θετικά (επιλεγμένα) αποτελέσματά τους. Τέλος, εξετάζουμε μία παραλλαγή της μεθόδου θεματικού Ranking SVM [49] (*Content-2*). Σε αυτή τη μέθοδο η αναπαράσταση του ερωτήματος επεκτείνεται ώστε να περιλαμβάνει μέσες τιμές και διακυμάνσεις για κάθε χαρακτηριστικό, βασιζόμενες στα 5 πρώτα αποτελέσματα κάθε ερωτήματος. Σημειώνουμε ότι η συγκεκριμένη βασική μέθοδος δεν είναι ταυτόσημη με τη μέθοδο του [49], από την άποψη ότι χρησιμοποιούμε τη στάνταρ Ranking SVM μέθοδο για την επίλυση του προβλήματος βελτιστοποίησης.

Εξετάζοντας όλες τις παραπάνω βασικές μεθόδους προσπαθούμε να επιδείξουμε την αποτελεσματικότητα της εκπαίδευσης πολλαπλών μοντέλων ταξινόμησης βασιζόμενων στο σκοπό αναζήτησης, έναντι μοντέλων που βασίζονται κυρίως στο περιεχόμενο, και, φυσικά, έναντι του αφελούς και του ιδανικού μοντέλου, που αποτελούν το κάτω και άνω όριο. Για αυτό το λόγο, εξαιρώντας το αφελές και το ιδανικό μοντέλο, στις υπόλοιπες μεθόδους, η επεξεργαστική ροή πέραν του σταδίου της συσταδοποίησης, εφαρμόζεται ακριβώς με τον ίδιο τρόπο.

4.5.3 Αποτελεσματικότητα ταξινόμησης

Το πρώτο πείραμα μετράει την αποτελεσματικότητα των αλγορίθμων σε ένα στατικό περιβάλλον. Οι μετρικές αξιολόγησης που χρησιμοποιούμε είναι οι MAP, Precision@n, και

NDCG@n.

Τα αποτελέσματα για τη μετρική MAP παρουσιάζονται στον Πίνακα 4.8. Αναμενόμενα, η εκπαίδευση μοντέλων ταξινόμησης ανά χρήστη έχει την καλύτερη απόδοση, επιτυγχάνοντας αύξηση περίπου 14% στην ακρίβεια, σε σχέση με την αφελή προσέγγιση της εκπαίδευσης ενός συνολικού μοντέλου ταξινόμησης. Σημειώνουμε, βέβαια, ότι σε πραγματικές συνθήκες εφαρμογής της μεθόδου, η συγκεκριμένη μέθοδος θα απαιτούσε πλήρη διαχωρισμό των χρηστών και αντιστοίχιση με τα ερωτήματά τους, το οποίο παραμένει ακόμα ανοιχτό πρόβλημα.

Πίνακας 4.8: Mean average precision.

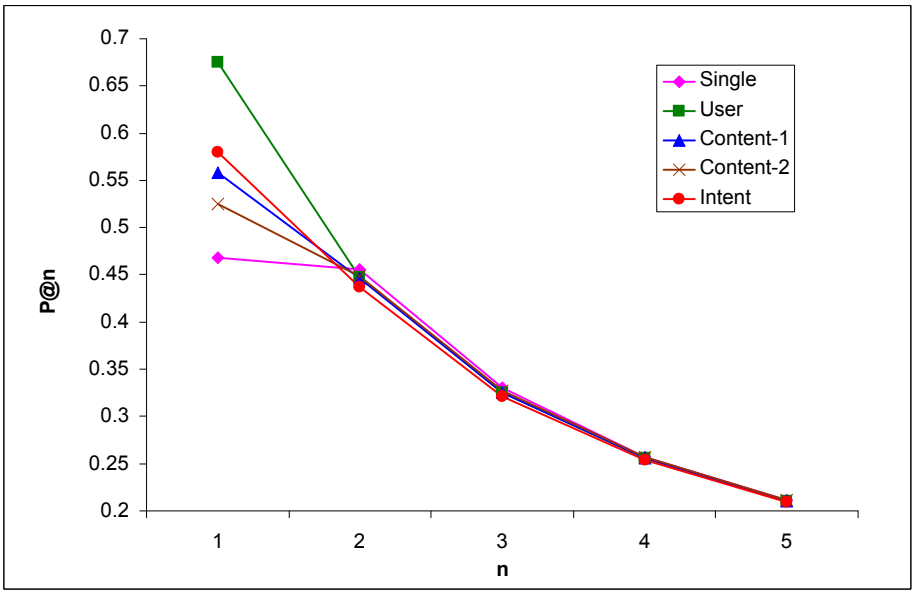
Method	**MAP**	**Increase**
Single	0.709	-
User	0.806	13.7%
Content-1	0.748	5.5%
Content-2	0.734	3.5%
Intent	0.754	6.3%

Αντιθέτως, οι μέθοδοι *Content-1*, *Content-2* και *Intent* είναι ανεξάρτητες από τους χρήστες, σχηματίζοντας ομάδες ερωτημάτων παρόμοιου περιεχομένου ή σκοπού, αντίστοιχα. Υπό αυτήν την έννοια είναι εφικτή η εφαρμογή τους σε συνθήκες πραγματικού κόσμου. Παρόλα αυτά διαφέρουν σημαντικά όσον αφορά στην αποτελεσματικότητα πρόβλεψής τους. Από τις τρεις μεθόδους, η *Content-2* είναι η πιο αδύναμη, παρόλο που αυξάνει την ακρίβεια σε σχέση με την αφελή βασική μέθοδο *Single* κατά 3.5%. Η *Content-1* επιτυγχάνει βελτίωση κατά 5.5% και η μέθοδός μας (*Intent*) κατά 6.3%.

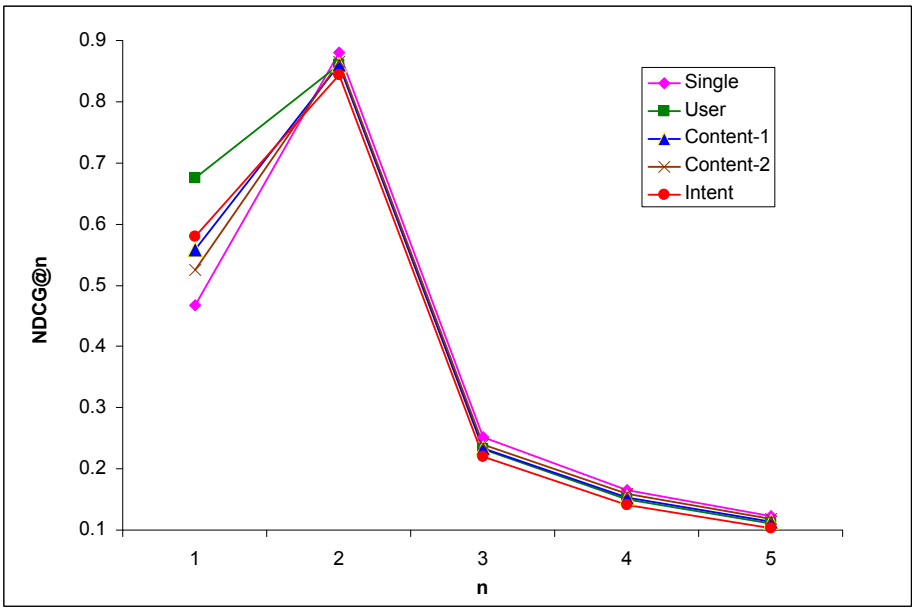
Παρόμοια εινόνα προκύπτει και από τη μετρική Precision@n η οποία παρουσιάζεται στην Εικόνα 4.6. Η απόδοση των μεθόδων ταυτίζεται για $n > 1$ εξαιτίας του μικρού αριθμού κρίσων σχετικότητας ανά ερώτημα (κατά μέσο όρο 3.2 ανά ερώτημα). Συγκεκριμένα, για τα 45984 ερωτήματα αξιολόγησης, υπάρχουν 51089 θετικές κρίσεις σχετικότητας (επιλογές αποτελεσμάτων από χρήστες), το οποίο μεταφράζεται σε περίπου 1.1 επιλογές αποτελεσμάτων ανά ερώτημα. Στην τιμή P@1, όμως, παρατηρούμε σημαντικές διαφορές στην απόδοση που επιβεβαιώνουν τις προηγούμενες παρατηρήσεις μας. Οι *Single* και *User* αποτελούν κάτω και άνω φράγματα, ενώ η *Intent* είναι πιο αποτελεσματική από τις *Content-1/2*. Οι γραφικές της Εικόνας 4.7 επιβεβαιώνουν τις παραπάνω παρατηρήσεις και για τη μετρική NDCG@n.

4.5.4 Ανάλυση

Για να αναλύσουμε καλύτερα τη φύση των μεθόδων που βασίζονται στο σκοπό έναντι αυτών που βασίζονται στο περιεχόμενο, εξετάζουμε ενδεικτικά λύσεις συσταδοποίησης που παράγονται από τις τρεις μεθόδους: *Intent*, *Content-1*, *Content-2*, στους Πίνακες 4.9, 4.10 και 4.11 αντίστοιχα. Διαλέξαμε, για όλες τις μεθόδους, συστάδες με ερωτήματα για τα οποία αποδίδουν καλά.



Σχήμα 4.6: Precision@n



Σχήμα 4.7: NDCG@n

Τα ποιοτικά αποτελέσματα είναι τα ακόλουθα. Πρώτον, οι μέθοδοι διαφέρουν σημαντικά όσον αφορά τον αριθμό συστάδων που παράγουν, έτσι ώστε να έχουν την καλύτερη απόδοση. Ενώ οι μέθοδοι που βασίζονται στο περιεχόμενο παράγουν από 20 (*Content-1*) έως 32 (*Content-2*) συστάδες, η μέθοδος *Intent* παράγει 75. Αν και συσταδοποιήσεις τέτοιου μεγέθους είναι δύσκολο να ερμηνευθούν, οι αριθμοί υποδεικνύουν ότι η μέθοδός μας εξειδικεύει περισσότερο τις παραγόμενες συστάδες, παράγοντας περισσότερες, με μικρότερο μέσο μέγεθος. Για την ακρίβεια, η *Intent* φαίνεται να αποδίδει καλά για πολλές εξειδικευμένες ανάγκες αναζήτησης, όπως φαίνεται στον Πίνακα 4.9. Το πρώτο σύνολο ερωτημάτων α-

Πίνακας 4.9: Παραδείγματα συσταδοποίησης για τη μέθοδο Intent.

Intent
1968 yamaha trailmaster 100 yl2 value
spendor s3 5 system
sonic video game 2011
peach fronted parakeet
85 mustang ignition module harness
owner of gold 39 s gym in wichita
cmmg complete 16 m4 lep ii gas piston upper
72 chevy fuel tank swap
closest city to labadee haiti
1985 mustang 302 engine ignition engine harness
2010 keystone fuzion fuzion model 322
artist lessons mountain painting
why is it important to follow the order of operation in algebra
who makes jet skis
why does spray paint come off
did you stay in a grand suite on freedom of the seas
where can i buy centrum materna in us
why is the order of operations for algebra
shooting a wedding without a flash
how much is the membership in makro philippines

ντιστοιχεί σε συστάδα που περιλαμβάνει ανάγκες αναζήτησης αποτελεσμάτων-κειμένων, ίσως εμπλουτισμένων με εικόνες, ενώ η δεύτερη συστάδα περιέχει ερωτήματα που αντιπροσωπεύουν εξειδικευμένες ερωτήσεις, που απαντούνται πιθανότατα από επίσης κειμενικά αποτελέσματα.

Αντίθετα, οι Πίνακες 4.10 και 4.11 δείχνουν παραδείγματα συστάδων για τις δύο κειμενικές μεθόδους. Ο πρώτος Πίνακας δείχνει δύο συστάδες για την *Content-1*. Ενώ η πρώτη συστάδα είναι παρόμοια με την αντίστοιχη της μεθόδου *Intent*, η δεύτερη είναι πάνω κάτω μία τυχαία συλλογή από ερωτήματα που αντιπροσωπεύουν ετερογενείς ανάγκες αναζήτησης. Τέλος, ο Πίνακας 4.11 δείχνει συστάδες για τη μέθοδο *Content-2*. Τα συμμετέχοντα ερωτήματα χαρακτηρίζονται από ομοιότητα όσον αφορά κοινούς όρους. Ο θόρυβος στις συστάδες μπορεί να εξηγηθεί από λέξεις κλειδιά που είναι βασικές για τη συστάδα (για παράδειγμα «picture»), που εμφανίζονται στα αποτελέσματα των ερωτημάτων αλλά όχι στα ίδια τα ερωτήματα.

4.5.5 Απόδοση ανά το χρόνο

Σε αυτήν την Ενότητα, προσομοιώνουμε μία πραγματική εφαρμογή των προσεγγίσεων ώστε να μελετήσουμε την εξέλιξή τους στο χρόνο. Για αυτό το σκοπό, χωρίσαμε το σύνολο αξιολόγησης χρονολογικά σε 10 διαδοχικά, ισόποσα υποσύνολα. Η Εικόνα 4.8 δείχνει την εξέλιξη της τιμής MAP ανά διαδοχικά χρονολογικά διαστήματα, *αθροιστικά*: η πρώτη τιμή αντιστοιχεί στην ακρίβεια για το πρώτο χρονικό διάστημα, η δεύτερη τιμή για τα δύο πρώτα

Πίνακας 4.10: Παραδείγματα συσταδοποίησης για τη μέθοδο Content-1.

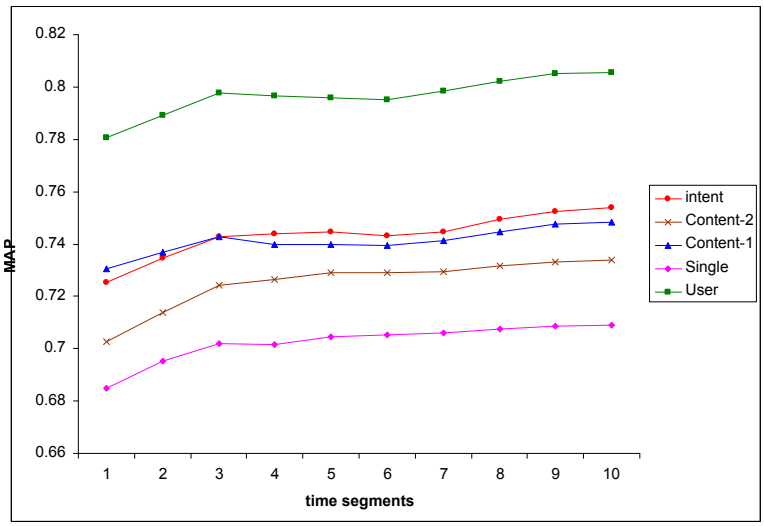
Content-1
austro diesel gmbh schwechat
skeleton reference of human muscle
mercury mountaineer 1998 california catalytic converters
65 mustang and dura spark need wiring diaphram
1985 mustang 302 electronic ignition module kit
double din dash facia for pt cruiser
after market stereo for 2010 pt cruiser
keilwerth tenor ex90
seiko ladies watch bracelet elegant
conn 37m tenor sax
san miguel beer philippines distributors in cavite
california pottery corn on the cob plater
new jersey animal shelters
rose garden portland oregon
best food to sell for profit
fbi national academy 2010 boston
passport renewal
new 2011 iphone 5
oprah wearing philip stein watches
top scottish baby names

Πίνακας 4.11: Παραδείγματα συσταδοποίησης για τη μέθοδο Content-2.

Content-2
mila kunis photo
marie osmond classical beauty doll margot
mickey mouse pictures
batman action figure power pack
lego star wars 2
fighter jets
dental office for sale in california
barrio indios puerto rico house rentals
tv shows solar power
logo design
hotel dei mellini rome
hotel suites in biscayne bay miami
gem kitchens and bath dublin

χρονικά διαστήματα και ούτω καθ' εξής.

Όπως έχει ήδη παρατηρηθεί από τα [5, 6], το εξατομικευμένο μοντέλο *User* σταθεροποιείται με την πάροδο του χρόνου. Αυτό έχει ως συνέπεια η αποτελεσματικότητά του να αυξάνεται ελαφρά με την πάροδο του χρόνου. Το ίδιο, όμως, ισχύει και για τις υπόλοιπες μεθόδους, κάτι που σημαίνει ότι, συνολικά, τα συγκεκριμένα μοντέλα έχουν σταθερή απόδοση καθώς τα ερωτήματα αξιολόγησης απομακρύνονται χρονικά από τα ερωτήματα εκπαίδευσης.



Σχήμα 4.8: MAP σε διαδοχικά χρονικά διαστήματα

4.5.6 Συζήτηση

Με μία πρώτη ανάγνωση όλες οι μέθοδοι, συμπεριλαμβανομένης και της προτεινόμενης, φαίνονται να ξεπερνώνται σε απόδοση από το καθαρά εξατομικευμένο μοντέλο *User*. Παρόλα αυτά, όπως προαναφέραμε, το συγκεκριμένο μοντέλο πρακτικά μπορεί να εφαρμοστεί μόνο σε ένα μικρό ποσοστό χρηστών/ερωτημάτων οι οποίοι είναι καταχωρημένοι και μπορούν να αντιστοιχηθούν ακριβώς στα ερωτήματά τους. Για αυτό το λόγο, η συγκεκριμένη μέθοδος αποτελεί μία ιδεατή αλλά όχι ρεαλιστική λύση. Ως εναλλακτικές λύσεις μπορούν να χρησιμοποιηθούν οι υπόλοιπες προσεγγίσεις, οι οποίες δεν βασίζονται σε πληροφορίες χρήστη. Από αυτές, η μέθοδος μας έχει σημαντικά καλύτερη απόδοση σε σχέση με τις μεθόδους που βασίζονται στο περιεχόμενο.

Στο δικό μας σενάριο, η αύξηση στις τιμές των μετρικών P@n και NDCG@n αντιπροσωπεύεται από σημαντική αύξηση στην τιμή P@1, δηλαδή η μέθοδός μας αποδίδει καλά στην ταξινόμηση σχετικών αποτελεσμάτων στην πρώτη θέση της κατάταξης. Αυτή η συμπεριφορά μπορεί να εξηγηθεί από το ίδιο το μοντέλο: ομαδοποιούμε ερωτήματα με παρόμοιο σκοπό αναζήτησης σε κοινές συστάδες, σε κάθε μία από τις οποίες αντιστοιχεί ένα διαφορετικό μοντέλο ταξινόμησης. Η τελική κατάταξη αποτελεσμάτων ενός νέου ερωτήματος προκύπτει συνδυάζοντας διάφορες επιμέρους κατατάξεις, από διαφορετικούς σκοπούς αναζήτησης. Οπότε, ακόμα κι αν το κειμενικό ταίριασμα ενός νέου ερωτήματος με συστάδες αποδειχθεί ανακριβές, για

παράδειγμα επειδή η κειμενική ομοιότητα δεν συνεπάγεται πάντα ομοιότητα στο σκοπό αναζήτησης, η τελική κατάταξη συνδυάζει ετερογενείς κατατάξεις και, κατά συνέπεια, σκοπούς αναζήτησης, ισοσταθμίζοντας πιθανή ανακρίβεια στο ταίριασμα.

Αν και η μέθοδός μας έχει ομοιότητες με την *Content-2*, έχει και ουσιαστικές διαφορές. Η *Content-2* αγνοεί τις διαφορές μεταξύ θετικών και αρνητικών αποτελεσμάτων, ενώ η *Intent*, αντιθέτως, τα λαμβάνει υπόψη της, βασίζοντας την εκπαίδευση των μοντέλων σε αυτές τις διαφορές. Η συσταδοποίηση της *Content-2* βασίζεται στον μέσο όρο των χαρακτηριστικών για τα top-k αποτελέσματα ενός ερωτήματος. Διαισθητικά, η *Content-2* συσταδοποιεί τα αποτελέσματα στο χώρο χαρακτηριστικών, ενώ η μέθοδός μας συσταδοποιεί τα ίδια τα μοντέλα αναζήτησης: αυτό το είδος «μετά-συσταδοποίησης» διατηρεί την συμπεριφορά αναζήτησης και επιτρέπει την ερμηνεία της ως μοντέλο ταξινόμησης βασισμένο στο σκοπό, κάτι που χάνεται από την *Content-2*.

Επιπλέον, η ανάλυση των υποσυνόλων δεδομένων αξιολόγησης, χωρισμένων σε χρονικά διαστήματα, στην Ενότητα 4.5.4 επιβεβαίωσε τις παρατηρήσεις των [5, 6] που υποστηρίζουν ότι η απόδοση των μοντέλων ταξινόμησης είναι συνήθως συνεπής στο πέρασμα του χρόνου.

Τέλος, το πιο ουσιαστικό συμπέρασμα της ανάλυσής μας είναι το εξής: σε κάθε περίπτωση κατά την οποία τα καθαρά εξατομικευμένα μοντέλα δεν έχουν εφαρμογή, μοντέλα βασισμένα στον σκοπό αναζήτησης μπορούν να τα αντικαταστήσουν με επαρκή και αποδοτικό τρόπο.

Κεφάλαιο 5

Προσαρμοστική Αναζήτηση σε Δεδομένα Σημασιολογικού Ιστού

5.1 Εξατομίκευση αναζήτησης σημασιολογικών δεδομένων

Σε αυτήν την ενότητα παρουσιάζουμε μία πρωταρχική μέθοδο που προτείνουμε για εξατομίκευση αναζήτησης με λέξεις κλειδιά, πάνω σε σημασιολογικά (RDF) δεδομένα [104]. Αρχικά, παρουσιάζουμε την προεπεξεργασία που πραγματοποιήσαμε σε δύο γνωστα σύνολα δεδομένων Netflix[1] και DBpedia[2], με σκοπό να παράγουμε ένα συνδυαστικό σετ δεδομένων, σε RDF μορφή, το οποίο θα είναι κατάλληλο για προσομοίωση της διαδικασίας σημασιολογικής αναζήτησης με λέξεις κλειδιά. Στη συνέχεια, παρουσιάζουμε τα χαρακτηριστικά εκπαίδευσης που ορίζουμε προκειμένου να εκπαιδεύσουμε συναρτήσεις αναταξινόμησης/εξατομίκευσης αποτελεσμάτων, καθώς και έναν πρωταρχικό αλγόριθμο εξατομίκευσης της αρχικά επιστρεφόμενης λίστας αποτελεσμάτων. Τέλος, παρουσιάζουμε μία πρώτη, ενθαρρυντική αξιολόγηση της προτεινόμενης μεθόδου.

5.1.1 Προεπεξεργασία και ανάλυση σετ δεδομένων

Προκειμένου να είναι δυνατή η εκπαίδευση μοντέλων εξατομίκευσης αναζήτησης είναι κρίσιμο να υπάρχουν διαθέσιμα δεδομένα για το ιστορικό αναζήτησης των χρηστών ή/και τις προτιμήσεις τους. Για να αποκτήσουμε δεδομένα ανάδρασης χρήστη, χρησιμοποιούμε το σετ δεδομένων της Netflix, το οποίο περιέχει βαθμολογίες χρηστών για ταινίες. Προκειμένου να εμπλουτίσουμε τη διαθέσιμη πληροφορία, εντοπίζουμε τις αντίστοιχες οντότητες (ταινίες για τις οποίες έχουμε βαθμολογίες) από την DBpedia, συλλέγοντας, επιπλέον, και περιφερειακή πληροφορία, όπως ηθοποιούς, σκηνοθέτες, είδος ταινίας, κ.α. Παρακάτω δίνουμε μία σύντομη περιγραφή των δύο συνόλων δεδομένων.

Η DBpedia είναι μία βάση γνώσης που βασίζεται στη συνεισφορά της κοινότητας για να εξάγει δομημένη πληροφορία από τη Wikipedia. Οι εγγραφές της παράγονται από τα Wikipedia infoboxes και αποθηκεύονται ως RDF τριπλέτες. Η έκδοση της DBpedia που

[1] http://www.netflixprize.com/community/viewtopic.php?id=68
[2] http://dbpedia.org/About

χρησιμοποιήσαμε χαρακτηρίζεται από τα ακόλουθα στατιστικά: περιγράφει περισσότερες από 3,5 εκατομμύρια οντότητες, από τις οποίες οι μισές είναι ταξονομήμενες κάτω από μία καλώς ορισμένη οντολογία. Περιλαμβάνει 416.000 πρόσωπα, 526.000 τοποθεσίες, 106.000 μουσικά άλμπουμ, 60.000 ταινίες, κτλ. Η οντολογία της αποτελείται από 320 κλάσεις, 750 ιδιότητες που δείχνουν σε οντότητες και 893 ιδιότητες που δείχνουν σε σταθερές τιμές. Από την άλλη πλευρά, το Netflix περιέχει περισσότερες από 100 εκατομμύρια αξιολογήσεις ταινιών στο διάστημα 1999-2005, με περίπου 480.189 χρήστες και 17.000 ταινίες.

Σκοπός της προεπεξεργασίας ήταν η παραγωγή ενός RDF συνόλου δεδομένων το οποίο θα περιείχε πληροφορία ανάδρασης χρήστη για αναζητούμενες οντότητες. Για αυτό το λόγο, συνενώνουμε τα δύο σύνολα πάνω στις κοινές ταινίες που περιέχουν. Προκειμένου ένα ταίριασμα να θεωρείται έγκυρο απαιτούμε ακριβές ταίριασμα των τίτλων των ταινιών (αφαιρουμένων των ειδικών χαρακτήρων) και της ημερομηνίας προβολής. Η παραπάνω διαδικασία μας έδωσε 5179 κοινές ταινίες πάνω στις οποίες συνενώθηκαν τα δύο σύνολα δεδομένων.

Ο Πίνακας 5.1 παρουσιάζει βασικές οντότητες της DBpedia πάνω στις οποίες θα στηριχθεί η αξιολόγηση της μεθόδου μας, καθώς και πώς αυτές οι οντότητες συσχετίζονται με τις ταινίες στο συνενωμένο σύνολο δεδομένων.

	actors	**directors**	**writers**
total	9666	2468	4945
appearing in more than 1 film	3706	997	1392
appearing in more than 5 films	990	169	140
appearing in more than 10 films	352	32	17
appearing in more than 15 films	121	5	6
appearing in more than 20 films	43	1	0

Πίνακας 5.1: Οντότητες του συνενωμένου συνόλου δεδομένων και στατιστικά πάνω στη σχέση τους με τη βασική οντότητα των ταινιών

Μαζί με την OWL οντολογία της DBpedia που περιγράφηκε παραπάνω, χρησιμοποιήσαμε τις κατηγορίες της Wikipedia, οι οποίες περιγράφονται με το λεξιλόγιο SKOS[3]. Μέσω των παραπάνω κατηγοριών μπορούμε να εξάγουμε πολύτιμη πληροφορία, όπως το είδος της ταινίας, θεματικές περιοχές, σκηνοθέτη, χρονολογία προβολής, κτλ. Όμοια, χρησιμοποιήσαμε την οντολογία YAGO[4], επίσης διαθέσιμη στο σύνολο της DBpedia. Ο Πίνακας 5.2 δίνει τον αριθμό SKOS και YAGO κατηγοριών που χαρακτηρίζουν διαφορετικούς αριθμούς ταινιών του συνόλου δεδομένων. Στην Ενότητα 5.1.2 περιγράφουμε πώς χρησιμοποιήσαμε τις παραπάνω οντολογίες ώστε να παράγουμε χαρακτηριστικά εκπαίδευσης.

	For more that # films								**Total**
For more that # films	**5**	**10**	**20**	**50**	**80**	**100**	**200**	**300**	**Total**
number of SKOS concepts	1158	718	324	166	93	63	22	3	4612
number of YAGO classes	732	482	279	112	67	47	24	6	2550

Πίνακας 5.2: Αριθμός συνολικών κατηγοριών SKOS και κλάσεων YAGO που χαρατκηρίζουν οντότητες ταινιών

[3]http://www.w3.org/TR/swbp-skos-core-spec/
[4]http://www.mpi-inf.mpg.de/yago-naga/yago/

Η επιλογή χρηστών πάνω στους οποίους βασίστηκε το τελικό πειραματικό σύνολο δεδομένων έγινε με βάση τα ακόλουθα κριτήρια:

- Αριθμός ταινιών που έχει βαθμολογήσει ο χρήστης, ώστε να εξασφαλιστεί ένα επαρκές μέγεθος για το σύνολο εκπαίδευσης.
- Ποσοστό βαθμολογημένων ταινιών στο Netflix για τις οποίες βρέθηκαν ταιριάσματα στη DBpedia, ως προς τις συνολικές ταινίες που έχει βαθμολογήσει ο χρήστης, ώστε να εξασφαλιστεί η συνέπεια του προφίλ του χρήστη, μετά τον περιορισμό των ταινιών που έχει βαθμολογήσει, λόγω συνένωσης των δύο συνόλων δεδομένων.
- Κατανομή των βαθμολογιών των χρηστών. Οι βαθμολογίες είναι ακέραιοι στο διάστημα $[1, 5]$. Χρήστες που δίνουν σταθερά πολύ χαμηλές ή πολύ υψηλές βαθμολογίες θεωρούνται ακραίες περιπτώσεις και απορρίπτονται.

Οι Πίνακες 5.3 και 5.4 δίνουν το τελικό σύνολο χρηστών που χρησιμοποιήθηκε, χωρισμένο σε δύο ομάδες: (α) A1-A14 χρήστες με πολλές βαθμολογίες ταινιών και (β) B1-B11 χρήστες με λίγες βαθμολογίες. Σημειώνουμε ότι η γραμμή *#ratings* αναφέρεται στον αριθμό ταινιών που έχουν βαθμολογήσει οι χρήστες, όσον αφορά το Netflix συνολικά. Αυτός ο αριθμός μειώνεται για κάθε χρήστη, μετά τη συνένωση των συνόλων.

user ID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
#ratings	2185	2206	1649	1770	1752	1726	1656	2092	2440	1935	1840	1675	1635	1890
mean rating	2.84	2.68	2.93	2.57	3.18	2.91	3.08	3.1	2.57	2.88	3.15	3.16	2.84	3.04
#(rating=5)	17	10	17	17	55	27	16	31	7	29	28	19	17	69
#(rating=4)	107	75	102	50	89	71	64	81	44	49	72	97	71	66
#(rating=3)	236	129	140	133	121	105	112	216	149	143	151	156	122	87
#(rating=2)	64	149	62	124	71	85	114	70	162	103	92	58	96	111
#(rating=1)	13	78	8	30	14	57	25	20	126	63	25	5	21	45

Πίνακας 5.3: Στατιστικά βαθμολογιών χρηστών - χρήστες με πολλές αξιολογήσεις

user ID	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11
#ratings	200	167	164	180	177	264	154	174	291	513	151
mean rating	3.61	3.23	3.48	3.29	2.42	3.13	3.06	3.07	3.05	3.63	3.2
#(rating=5)	3	0	5	0	0	0	3	2	0	6	2
#(rating=4)	8	11	10	12	6	8	6	14	11	31	9
#(rating=3)	25	12	7	12	12	43	12	10	40	59	15
#(rating=2)	2	7	6	8	8	1	6	7	7	6	4
#(rating=1)	2	3	4	4	9	0	3	1	0	0	0

Πίνακας 5.4: Στατιστικά βαθμολογιών χρηστών - χρήστες με λίγες αξιολογήσεις

5.1.2 Εκπαίδευση συναρτήσεων ταξινόμησης και εξατομίκευσης για RDF δεδομένα

Χρησιμοποιούμε το Ranking SVM μοντέλο για να εκπαιδεύσουμε συναρτήσεις αναταξινόμησης/εξατομίκευσης αποτελεσμάτων. Εξαιτίας των περιορισμών που τίθενται από τα περιορισμένα δεδομένα ανάδρασης χρηστών που διαθέτουμε, θεωρούμε ένα μεγάλο αριθμό από

χαρακτηριστικά εκπαίδευσης, έτσι ώστε να μπορούμε να αναπαραστήσουμε όσο το δυνατόν περισσότερη πληροφορία για τις εξεταζόμενες οντότητες. Επειδή το σύνολο δεδομένων μας δεν περιέχει ερωτήματα και τα αντίστοιχα (επιλεγμένα και μη) αποτελέσματα, μπορούμε να ορίσουμε μόνο χαρακτηριστικά εκπαίδευσης τα οποία είναι ανεξάρτητα από ερωτήματα. Επιπλέον, χειριζόμαστε όλες τις ταινίες που έχουν βαθμολογηθεί από έναν χρήστη ως αποτελέσματα ενός μοναδικού, υποθετικού ερωτήματος που αποτελείται από τον όρο «film» και θεωρούμε τις βαθμολογίες τους αντιπροσωπευτικές της επιθυμητής θέσης κατάταξης στην οποία θα ήθελε ο χρήστης να τις δει, για το συγκεκριμένο ερώτημα.

Τα χαρακτηριστικά που ορίζουμε για την εκπαίδευση του μοντέλου σχετίζονται με ιστορικό αναζήτησης του χρήστη, αλλά και με τη δομή του RDF γράφου του συνόλου δεδομένων. Οι οντότητες των ηθοποιών και των σκηνοθετών θεωρούνται στενά συνδεδεμένες με την άποψη ενός χρήστη για μία ταινία, παρόλο που η διαδικασία εκπαίδευσης λαμβάνει υπόψη μόνο οντότητες ταινιών. Για να εξασφαλίσουμε το παραπάνω, ορίζουμε χαρακτηριστικά εκπαίδευσης που μπορούν να εφαρμοστούν σε όλα τα είδη οντοτήτων. Στη συνέχεια, παρουσιάζουμε την κατηγοριοποίηση των χαρακτηριστικών που υλοποιούμε.

1. Ηθοποιοί, βασισμένοι στην ιδιότητα *starring*. Κάθε ηθοποιός αναπαρίσταται από ένα χαρακτηριστικό με boolean τιμή.

2. Σκηνοθέτες, βασισμένοι στην ιδιότητα *director*. Κάθε σκηνοθέτης αναπαρίσταται από ένα χαρακτηριστικό με boolean τιμή.

3. Κατηγορίες SKOS των ταινιών. Για μία ταινία το χαρακτηριστικό κάθε κατηγορίας παίρνει τιμή 1, αν η ταινία ανήκει στην κατηγορία, και 0 διαφορετικά. Για ηθοποιούς και σκηνοθέτες, η τιμή του χαρακτηριστικού είναι ο αριθμός των ταινιών στις οποίες συμμετείχαν και οι οποίες ανήκουν στην κατηγορία.

4. Κατηγορίες SKOS των ηθοποιών/σκηνοθετών. Για ηθοποιούς και σκηνοθέτες, το χαρακτηριστικό κάθε κατηγορίας παίρνει boolean τιμή ανάλογα με το αν σχετίζονται με την κατηγορία. Για ταινίες, η τιμή του χαρακτηριστικού είναι ο αριθμός των ηθοποιών/σκηνοθετών που συμμετέχουν στην ταινία και ανήκουν στην κατηγορία.

5. Είδος ταινίας από τη βάση γνώσης imdb[5]. Τα χαρακτηριστικά υπολογίζονται με τον ίδιο τρόπο με το 3.

6. Είδος ταινίας από το imdb και το SKOS συνδυαστικά. Για οντότητες ταινιών, το χαρακτηριστικά παίρνουν boolean τιμές ανάλογα με τον αν η ταινία χαρακτηρίζεται με το συγκεκριμένο είδος και ανήκει στην συγκεκριμένη SKOS κατηγορία. Για ηθοποιούς/σκηνοθέτες, αντίστοιχα, η τιμή του χαρακτηριστικού υπολογίζεται όπως στο 3.

7. Yago κλάσεις για τις ταινίες. Τα χαρακτηριστικά αντιπροσωπεύουν όλες τις κλάσεις της οντολογίας Yago στις οποίες ανήκουν οι ταινίες. Η τιμή του κάθε χαρακτηριστικού υπολογίζεται όπως στο 3.

[5]http://www.imdb.com/

8. Yago κλάσεις για τους ηθοποιούς/σκηνοθέτες. Τα χαρακτηριστικά αντιστοιχούν στις κλάσεις της συγκεκριμένης οντολογίας στις οποίες ανήκουν οι ηθοποιοί/σκηνοθέτες. Η τιμή του κάθε χαρακτηριστικού υπολογίζεται όπως στο 4.

9. Ο βαθμός εισερχόμενων και εξερχόμενων ιδιοτήτων από την κάθε οντότητα.

10. Συνολικός αριθμός ιδιοτήτων *starring* και *director* ιδιοτήτων που αφορούν την οντότητα.

Αρχικά, υλοποιούμε μία στοιχειώδη μηχανή αναζήτησης, ευρετηριάζοντας πόρους (οντότητες), κλάσεις, ιδιότητες και σταθερές τιμές από το σύνολο δεδομένων μας, χρησιμοποιώντας το εργαλείο Lucene[6]. Ορίζουμε ως ερώτημα μία διαχωριζόμενη από κόμματα ακολουθία όρων ή φράσεων, όπου τα κόμματα διαχωρίζουν όρους ή φράσεις που προσανατολίζονται σε διαφορετικά είδη οντοτήτων. Για παράδειγμα, το ερώτημα $Q = \{film, woody\ allen\}$ σημαίνει ότι αναζητούμε διαφορετικές οντότητες για τον όρο «film» και διαφορετικές για το «woody allen». Για καθένα από τους όρους/φράσεις, επιστρέφονται, με τη βοήθεια του Lucene, διάφορες οντότητες ως πιθανά αποτελέσματα για την ανάγκη αναζήτησης του χρήστη. Ο Αλγόριθμος 5.1.2 περιγράφει τη διαδικασία ανάκτησης και ενσωματώνει την εξατομικευμένη αναταξινόμηση των αποτελεσμάτων-οντοτήτων.

Algorithm 1 RDF results retrieval for keyword queries

for each keyword K_i **do**

Retrieve the full result list RL_i of size N_i and the respective scores S_{ij} for each result r_{ij}

Normalize scores S_{ij} in the interval $[0, 1]$

For each result r_{ij}, search whether the rest keywords K_k $k \neq i$ are found in its abstract textual description. If so, double its score.

For each result r_{ij} search whether it has been retrieved using the *label* property of the entity. If so, double its score.

Input the final result list into Ranking SVM and retrieve the personalized, re-ranked list PRL_i

Prune the result list of each keyword according to a given threshold of number of results.

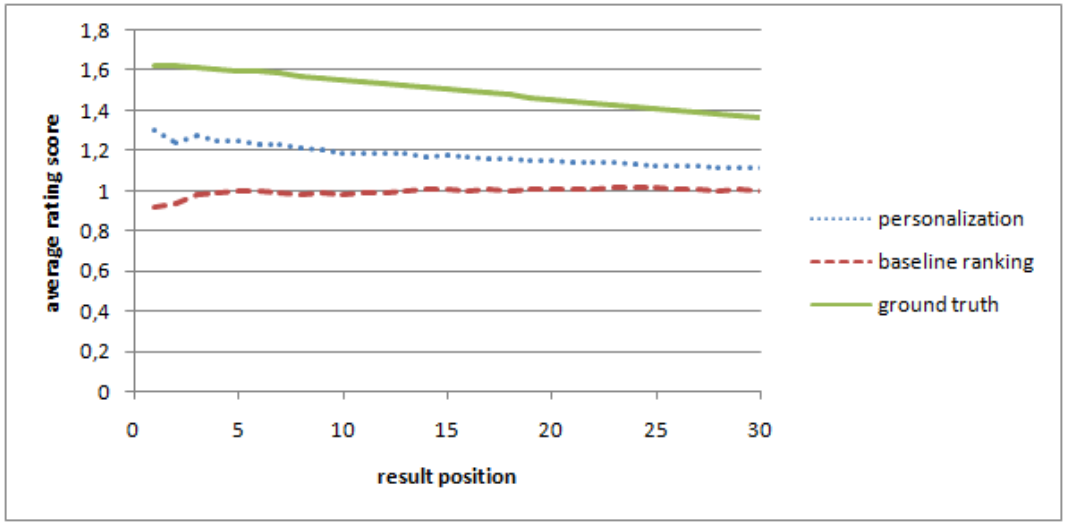
end for

5.1.3 Πειραματική αξιολόγηση

Σε αυτήν την ενότητα συγκρίνουμε την αποτελεσματικότητα της προσέγγισής μας με (α) τη βασική αλήθεια (ground truth) που δίνεται από τις βαθμολογήσεις των ταινιών στο σύνολο δεδομένων και αντιπροσωπεύει τη βέλτιστη λύση του προβλήματος και (β) τη βασική μέθοδο ταξινόμησης αποτελεσμάτων μέσω της μηχανής αναζήτησης που ορίζουμε, χωρίς εξατομίκευση. Για τα πειράματά μας θεωρήσαμε το 80% των διαθέσιμων δεδομένων ως σετ εκπαίδευσης

[6]http://lucene.apache.org/core/

και το υπόλοιπο 20% ως σετ αξιολόγησης. Σημειώνουμε ότι η βασική αλήθεια εξάγεται αποκλειστικά από το σετ αξιολόγησης, το οποίο δεν συμμετέχει στη διαδικασία εκπαίδευσης. Οι αξιολογηθέντες χρήστες παρουσιάστηκαν στον Πίνακα 5.3.



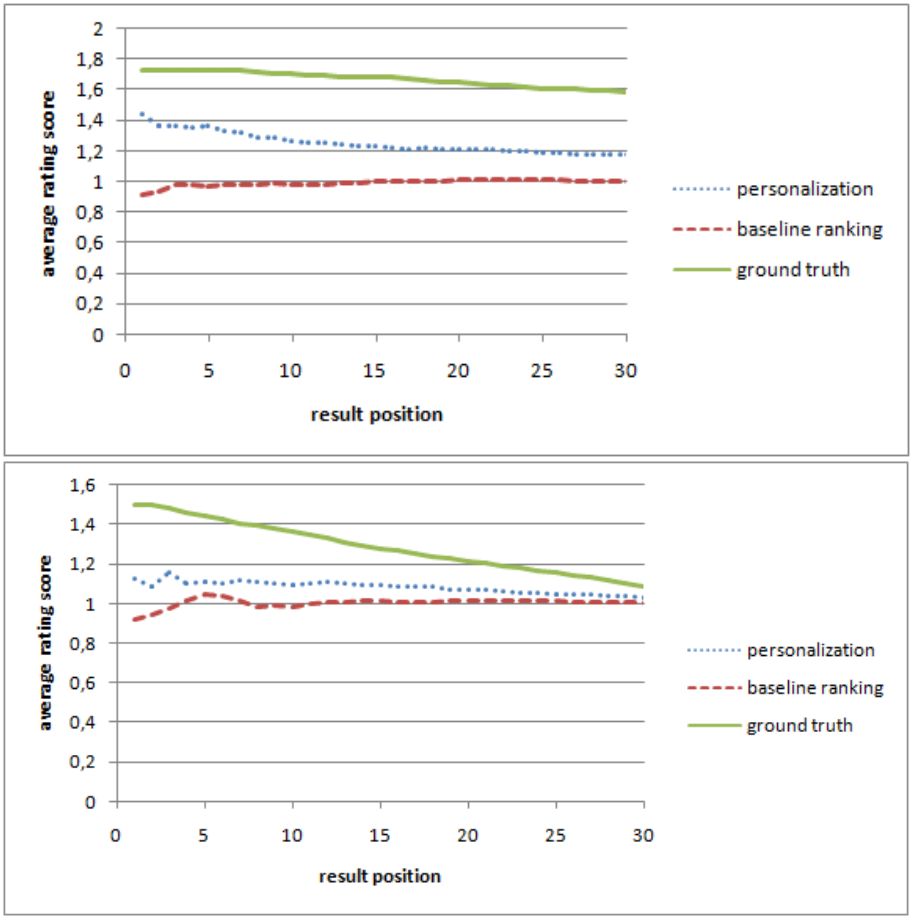
Σχήμα 5.1: Μέσα σκορ βαθμολόγησης για κάθε προσέγγιση

Η Εικόνα 5.1 παρουσιάζει τη μέση βαθμολογία των ταξινομημένων αποτελεσμάτων σε κάθε θέση της κατάταξης, για κάθε δοκιμαζόμενη προσέγγιση. Η γραφική *ground truth* αντιπροσωπεύει την ιδεατή κατάταξη η οποία θα τοποθετούσε σε ψηλότερες θέσεις αποτελέσματα με υψηλότερες βαθμολογίες χρηστών, ακολουθούμενα πάντα από αποτελέσματα με χαμηλότερη βαθμολογία. Η *personalization* αντιπροσωπεύει τη δική μας μέθοδο, ενώ η *baseline ranking* την βασική ταξινόμηση αποτελεσμάτων χωρίς εξατομίκευση.

Παρατηρούμε ότι η μέθοδός μας, καταρχήν, υπερισχύει της βασικής μεθόδου, επιστρέφοντας σε υψηλότερες θέσεις αποτελέσματα με υψηλότερες βαθμολογίες. Δεύτερον, η γραφική παράσταση της μεθόδου μας διαρκώς φθίνει, κάτι που σημαίνει ότι δεν υπάρχουν σημαντικά ακραία αποτελέσματα, δηλαδή πολλά αποτελέσματα χαμηλής βαθμολογίας σε υψηλές θέσεις, τα οποία θα προκαλούσαν αρχική κάθοδο και μετά άνοδο της γραφικής. Τέλος, αν και η βασική αλήθεια είναι σταθερά καλύτερη (όπως και αναμενόταν) από τη μέθοδό μας, οι δύο γραφικές παρουσιάζουν σχεδόν την ίδια συμπεριφορά, δηλαδή η κλίση της γραφικής της μεθόδου μας ακολουθεί την κλίση της βασικής αλήθειας.

Στην Εικόνα 5.2, παρουσιάζονται τα ίδια πειραματικά αποτελέσματα, αλλά ξεχωριστά για χρήστες με πολλές βαθμολογήσεις ταινιών (πάνω γραφική) και χρήστες με λίγες βαθμολογήσεις (κάτω). Παρατηρούμε ότι το μέγεθος του ιστορικού/προφίλ του χρήστη επηρεάζει σημαντικά την αποτελεσματικότητα του μοντέλου εκπαίδευσης. Για χρήστες με μεγάλη ανάδραση (βαθμολογίες ταινιών), η μέθοδος εξατομίκευσης έχει πολύ κοντινά αποτελέσματα με την ιδεατή μέθοδο, ειδικά στις υψηλότερες θέσεις κατάταξης, ενώ φθίνει και αρκετά πιο απότομα. Από την άλλη πλευρά, για χρήστες με λίγες βαθμολογήσεις, η μέθοδος εξατομίκευσης φθίνει λιγότερο απότομα και βρίσκεται αρκετά πιο κοντά στη βασική μέθοδο, αποδίδοντας,

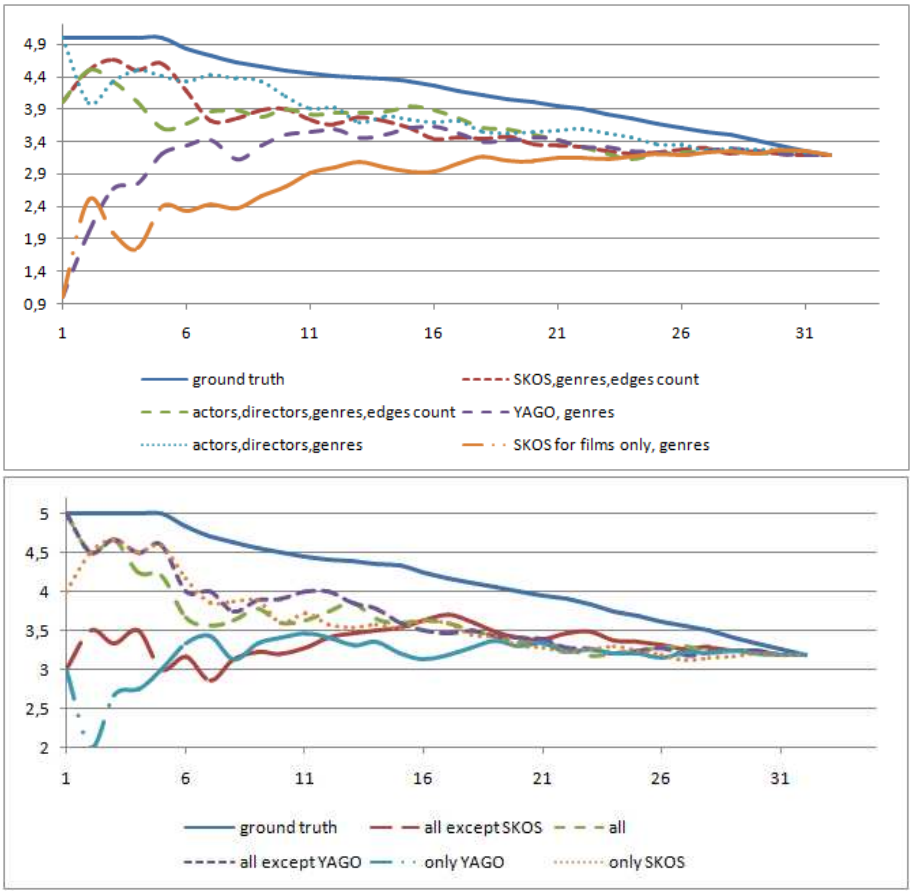
παρόλα αυτά, καλύτερα.



Σχήμα 5.2: Μέση βαθμολογία ανακτημένων αποτελεσμάτων, σε κάθε θέση κατάταξης, για «καλούς» (πάνω) και «κακούς» (κάτω) χρήστες.

Όπως προαναφέραμε, ένας μεγάλος αριθμός χαρακτηριστικών χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου. Αν και, συνολικά, τα χαρακτηριστικά αυτά φαίνεται να βελτιώνουν την απόδοση του μοντέλου, είναι ενδιαφέρον να εξετάσουμε την επιμέρους σημασία ομάδων χαρακτηριστικών ως προς την αποτελεσματικότητα της εξατομίκευσης αποτελεσμάτων. Στην Εικόνα 5.3 παρουσιάζεται η επίδραση των χρησιμοποιούμενων χαρακτηριστικών εκπαίδευσης στην αποτελεσματικότητα του μοντέλου (για ευκολία στην παρουσίαση, παρουσιάζουμε δύο διαφορετικές γραφικές, εξαιτίας του μεγάλου αριθμού διαφορετικών συνδυασμών χαρακτηριστικών που εξετάζονται). Σημειώνουμε ότι οι συγκεκριμένες γραφικές αφορούν έναν τυχαίο χρήστη (B3).

Παρατηρούμε ότι η αποτελεσματικότητα του μοντέλου διαφέρει σημαντικά για διαφορετικές ομάδες χαρακτηριστικών. Για παράδειγμα, τα χαρακτηριστικά που σχετίζονται με τις κλάσεις YAGO φαίνεται να έχουν αρνητική επίδραση. Από την άλλη πλευρά, χαρακτηριστικά που σχετίζονται με τις SKOS κατηγορίες φαίνεται να επιδρούν θετικά. Μέρος της μελλοντικής δουλειάς μας είναι μία πιο ολοκληρωμένη διαδικασία επιλογής χαρακτηριστικών, σε συσχέτιση και με τα ατομικά ιστορικά βαθμολόγησης των χρηστών.



Σχήμα 5.3: Επίδραση διαφορετικών χαρακτηριστικών στο μεντέλο εξατομίκευσης

Σημειώνουμε εδώ ότι τα παραπάνω αποτελέσματα αφορούν οντότητες ταινιών, πάνω στις οποίες έχουμε άμεσες βαθμολογήσεις από τους χρήστες. Παρόμοια πειράματα εξατομίκευσης ηθοποιών και σκηνοθετών δεν έδωσαν καλά αποτελέσματα οπότε παραλείπεται η παρουσίασή τους. Η βελτίωση της αποτελεσματικότητας του μοντέλου για τέτοιες οντότητες, για τις οποίες υπάρχει μόνο έμμεση ανάδραση χρήστη είναι κομμάτι της μελλοντικής μας δουλειάς.

5.1.4 Συμπεράσματα

Σε αυτήν την ενότητα παρουσιάστηκε μία αρχική μεθοδολογία για εξατομίκευση αποτελεσμάτων αναζήτησης με λέξεις κλειδιά σε σημασιολογικά (RDF) δεδομένα. Ορίσαμε μία σειρά από χαρακτηριστικά εκπαίδευσης, βασιζόμενοι και στις ιδιαιτερότητες των RDF δεδομένων και υιοθετήσαμε το μοντέλο Ranking SVM για να εκπαιδεύσουμε συναρτήσεις εξατομίκευσης. Για να μπορέσουμε να το κάνουμε αυτό, συνενώσαμε δεδομένα από δύο διαφορετικά σύνολα, (DBpedia, Netflix), παράγοντας ένα σετ δεδομένων κατάλληλο να «προσομοιώσει» δεδομένα ανάδρασης χρηστών σε αναζήτηση σημασιολογικών δεδομένων. Εφαρμόσαμε την προτεινόμενη μέθοδο με τη βοήθεια μίας βασικής μηχανής αναζήτησης και δείξαμε την αποτελεσματικότητα της μεθόδου μας.

Βασικό κομμάτι της μελλοντικής μας δουλειάς είναι η βελτίωση/επέκταση της τρέχουσας

μεθόδου ώστε να λειτουργεί πάνω σε πλήρη αποτελέσματα-γράφους και όχι απλά σε αυτόνομες οντότητες. Η πρόκληση εδώ είναι ο αποδοτικός συνδυασμός ξεχωριστών οντοτήτων ώστε να παράγονται αποτελέσματα-γράφοι με νόημα. Επιπλέον, σκοπεύουμε να βελτιώσουμε αδυναμίες της τρέχουσας μεθόδου, όπως επιλογή χαρακτηριστικών εκπαίδευσης και εξατομίκευση οντοτήτων για τις οποίες μόνο έμμεση ανάδραση χρήστη είναι διαθέσιμη.

5.2 Σημασιολογική επισημείωση και αναζήτηση

Σε αυτήν την ενότητα παρουσιάζουμε το αναπτυχθέν πλαίσιο σημασιολογικής επισημείωσης εγγράφων με χρήστη οντολογιών και υβριδικής αναζήτησης, τόσο με λέξεις κλειδιά, όσο και με περιήγηση σε ιεραρχίες οντολογιών (GoNTogle) [111, 110]. Η κύρια συνεισφορά του προτεινόμενου πλαισίου δίνεται παρακάτω:

1. Σχεδιάσαμε και υλοποιήσαμε ένα φιλικό προς το χρήστη πλαίσιο επισημείωσης εγγράφων που υποστηρίζει αρκετούς γνωστούς μορφότυπους κειμένων και προσφέρει προχωρημένες δυνατότητες αναζήτησης.

2. Το πλαίσιο ακολουθεί λογική βασισμένη σε εξυπηρετητή (server), όπου οι επισημειώσεις αποθηκεύονται σε ένα κεντρικό αποθετήριο, ξεχωριστά από το ίδιο το έγγραφο. Με αυτόν τον τρόπο, επιτυγχάνεται ένα συνεργατικό περιβάλλον, όπου κάθε χρήστης μπορεί να εκμεταλλευτεί τις επισημειώσεις των άλλων χρηστών, είτε για επισημείωση, είτε για αναζήτηση εγγράφων.

3. Προτείνουμε μία μέθοδο μηχανικής μάθησης για αυτοματοποιημένη επισημείωση κειμένων, με βάση μοντέλα εκπαιδευμένα πάνω στο ιστορικό επισημείωσης προηγούμενων κειμένων.

4. Εισάγουμε μία υβριδική μέθοδο αναζήτησης που συνδυάζει κλασσική αναζήτηση με λέξεις κλειδιά με σημασιολογική αναζήτηση.

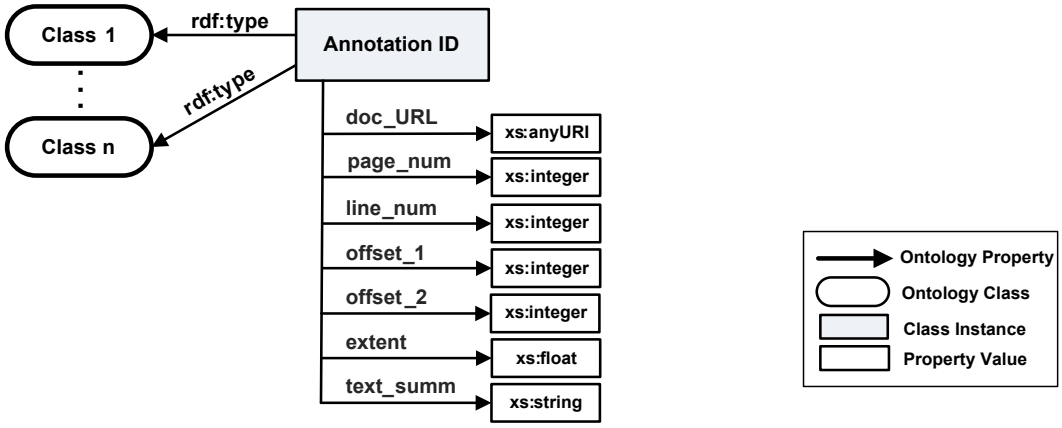
5. Παρουσιάζουμε μία μελέτη χρηστών, η οποία καταδεικνύει την αποτελεσματικότητα της αυτοματοποιημένης επισημείωσης. Επίσης, δείχνουμε ότι η προτεινόμενη υβριδική αναζήτηση υπερισχύει των απλών, επιμέρους τρόπων αναζήτησης (με λέξεις κλειδιά και με περιήγηση στην οντολογία), τόσο σε ακρίβεια, όσο και σε ανάκληση

5.2.1 Σημασιολογική επισημείωση

Το GoNTogle υποστηρίζει σημασιολογική επισημείωση για διάφορους, ευρέως χρησιμοποιούμενους μορφότυπους κειμένων (doc, pdf, txt, rtf, odt, sxw). Επιτρέπει την επισημείωση ολόκληρου του κειμένου ή κομματιών του, καθώς και χειροκίνητη ή ημιαυτόματη επισημείωση.

Το μοντέλο επισημείωσης είναι κοινό για όλους τους μορφότυπους κειμένων, παρόλο που διαφορετικοί μορφότυποι ενδέχεται να επιτρέπουν την εξαγωγή διαφορετικής μορφής μεταδεδομένων επισημείωσης. Οι επισημειώσεις αποθηκεύονται σε έναν κεντρικό εξυπηρετητή οντο-

λογίας, ανεξάρτητα από τα έγγραφα. Οι επισημειώσεις από κάθε είδος εγγράφου ορίζονται και οργανώνονται/αποθηκεύονται ακριβώς με τον ίδιο τρόπο. Κάθε επισημείωση κωδικοποιείται ως στιγμιότυπο μίας κλάσης της οντολογίας, μαζί με την μεταπληροφορία της επισημείωσης για το έγγραφο. Αυτή η μεταπληροφορία οργανώνεται με τη βοήθεια ενός ελάχιστου συνόλου από ιδιότητες που ορίζουμε, έτσι ώστε η επισημείωση να μπορεί να «αναπαραχθεί» και να αναπαρασταθεί κάθε φορά που φορτώνεται, πάνω στο κείμενο του εγγράφου. Οι ιδιότητες αυτές περιέχουν πληροφορία όπως: document URL, annotation offsets, page number, extent of annotation over the document κλπ.



Σχήμα 5.4: Μοντέλο επισημείωσης

Η Εικόνα 5.4 δείχνει το μοντέλο επισημείωσης του συστήματος. Οι επισημειώσεις είναι οντότητες που μπορεί να ανήκουν σε μία ή περισσότερες κλάσεις. Μέσω των ιδιοτήτων που χαρακτηρίζουν τις οντότητες επισημείωσης, κωδικοποιούμε όλη την απαραίτητη μεταπληροφορία. Η ιδιότητα *doc_URL* αποθηκεύει το URL του εγγράφου. Οι ιδιότητες *page_num* και *line_num* περιέχουν τον αριθμό σελίδας και γραμμής μέσα στη σελίδα, στις οποίες ξεκινά η επισημείωση. Η $offset\_1$ αντιστοιχεί στη θέση της αρχής της επισημείωσης από την αρχή του εγγράφου, ενώ η $offset\_2$ στη θέση του τέλους της επισημείωσης από την αρχή του εγγράφου. Η *extent* αποθηκεύει το μέγεθος της επισημείωσης και η *text_sum* μία μικρή περίληψη του κειμένου της επισημείωσης, για λόγους παρουσίασης στη γραφική διεπιφάνεια.

Αυτόματη σημασιολογική επισημείωση

Στη συνέχεια, παρουσιάζουμε το μοντέλο εκμάθησης που εφαρμόζουμε για αυτοματοποιημένη επισημείωση εγγράφων. Προτείνουμε μία μέθοδο βασισμένη στον αλγόριθμο kNN (k Nearest Neighbor) [50] που εκμεταλλεύεται το ιστορικό επισημείωσης των χρηστών για να προτείνει αυτόματα επισημειώσεις για νέα έγγραφα. Το σύνολο εκπαίδευσης του αλγορίθμου αποτελείται από προηγούμενες επισημειώσεις χρηστών. Συγκεκριμένα, όταν πραγματοποιείται μία επισημείωση, το αντίστοιχο κείμενο εξάγεται και ευρετηριοποιείται σε ένα ανεστραμμένο ευρετήριο (inverted index). Μαζί με την κειμενική πληροφορία, στο ευρετήριο κρατείται και πληροφορία σχετικά με τις κλάσεις επισημείωσης.

Algorithm 2 Annotation Suggestion Algorithm

Input: Selected text st, index I

Output: Suggested class cl_i, suggested class score Scr_{cli}

```
for each annotated text at in I do
    calculate ts_{st,at}
end for
Insert the k most similar annotated texts in S
for each at in S do
    for for each class cl annotating at do
        Scr_cl = Scr_cl + (w1 * ts_{st,at}) * (w2 * e_{cl,at})
    end for
end for
return cl_i, Scr_cli
```

Ο Αλγόριθμος 5.2.1 παρουσιάζει τη διαδικασία προτάσεων επισημείωσης. Είσοδος είναι το επιλεγμένο κείμενο st και το ευρετήριο I. Βασιζόμενοι στο σκορ ομοιότητας $ts_{st,at}$ μεταξύ του επιλεγμένου κειμένου st και οποιουδήποτε ευρετηριασμένου κειμένου at ήδη υπάρχουσας επισημείωσης, κρατάμε τις k πιο όμοιες επισημειώσεις στο σύνολο S (γραμμές 1-4). Στη συνέχεια, για κάθε επισημείωση του συνόλου S, εξάγουμε τις κλάσεις που της αντιστοιχούν. Σε κάθε κλάση cl ανατίθεται ένα σκορ Scr_{cl} που συνδυάζει: (α) την κειμενική ομοιότητα μεταξύ του κειμένου της επιλεγμένης και της ήδη υπάρχουσας επισημείωσης και (β) ένα σκορ $e_{cl,at}$ που αντιπροσωπεύει την έκταση κατά την οποία η κλάση cl καλύπτει ένα επισημειωμένο κείμενο at (γραμμή 7). Το $e_{cl,at}$ ορίζεται ως ο λόγος των όρων που είναι επισημειωμένοι με την κλάση cl, προς το συνολικό αριθμό όρων του επισημειωμένου κειμένου at:

$$e_{cl,at} = \frac{number\_of\_tokens\_of\_cl\_annotations\_over\_at}{number\_of\_tokens\_in\_at} \tag{5.1}$$

Τα βάρη $w1$ και $w2$ χρησιμοποιούνται για να σταθμίσουν τη βαρύτητα των δύο σκορ. Τελικά, μία ταξινομημένη λίστα από προτεινόμενες κλάσεις cl_i και τα αντίστοιχα σκορ τους Scr_{cli} παρουσιάζεται στο χρήστη.

5.2.2 Αναζήτηση

Στη συνέχεια, παρουσιάζουμε τις δυνατότητες αναζήτησης του πλαισίου. Αρχικά, ορίζουμε τυπικά τους υποστηριζόμενους τύπους αναζήτησης και έπειτα αναλύουμε τις, βασιζόμενες στην οντολογία, προχωρημένες λειτουργίες αναζήτησης. Ο ακόλουθος πίνακας επεξηγεί τα σύμβολα που χρησιμοποιούμε.

Τύποι αναζήτησης

Κατηγοριοποιούμε τις βασικές δυνατότητες του πλαισίου σε τρεις τύπους: (α) Αναζήτηση με λέξεις κλειδιά, (β) Σημασιολογική αναζήτηση και (γ) Υβριδική αναζήτηση:

Symbol	Notation
qkey	Keyword query, consisting of search termt1, t2,...tm
Skey(qkey)	Keyword-based search
RSkey	Keyword-based search result set
Scrkey(qkey,d)	Keyword-based similarity score
qsem	Semantic query, consisting of search classes cl1, cl,...cln
Ssem(qsem)	Semantic-based search
RSsem	Semantic-based search result set
Scrsem(qsem,d)	Semantic-based similarity score
Shybr(qsem,qkey)	Hybrid search
RShybr	Hybrid search result set
Scrhybr(qsem,qkey,d)	Hybrid similarity score

Πίνακας 5.5: Επεξήγηση συμβόλων

Αναζήτηση με λέξεις κλειδιά. Ο χρήστης παρέχει τους όρους αναζήτησης και το σύστημα ανακτά σχετικά έγγραφα βασισμένο μόνο στην κειμενική ομοιότητα όρων αναζήτησης/κειμένων. Υιοθετούμε τη συνάρτηση βαθμολόγησης - ταξινόμησης αποτελεσμάτων της μηχανής αναζήτησης Lucene. Έστω ένα ερώτημα $q_{key} = t_1, t_2, ...t_m$, όπου t_i οι όροι (λέξεις κλειδιά) του ερωτήματος. Συμβολίζουμε τα αποτελέσματα της παραπάνω αναζήτησης ως ένα ταξινομημένο σύνολο RS_{key} από πλειάδες $< d, Scr_{key}(q_{key}, d) >$, όπου d όλα τα έγγραφα που ανακτήθηκαν, με σκορ $Scr_{key}(q_{key}, d)$.

Σημασιολογική αναζήτηση. Ο χρήστης περιηγείται στις κλάσεις της οντολογίας και επικεντρώνει την αναζήτηση σε μία ή περισσότερες από αυτές. Ομοίως με την κλασσική αναζήτηση, ορίζουμε ένα σημασιολογικό ερώτημα ως $q_{sem} = cl_1, cl_2, ...cl_n$, όπου cl_i οι κλάσεις αναζήτησης. Ορίζουμε τα αποτελέσματα της παραπάνω αναζήτησης ως ένα ταξινομημένο σύνολο RS_{sem} από πλειάδες $< d, Scr_{sem}(q_{sem}, d) >$, όπου d όλα τα έγγραφα που έχουν χαρακτηριστεί με τουλάχιστον μία από τις κλάσεις του ερωτήματος, με σημασιολογικό σκορ $Scr_{sem}(q_{sem}, d)$. Προκειμένου να ορίσουμε το σημασιολογικό σκορ, εξετάζουμε την έκταση της κάλυψης ενός εγγράφου από κάθε κλάση, δηλαδή το λόγο των όρων ενός εγγράφου που χαρακτηρίζονται με την κλάση, προς το συνολικό αριθμό όρων του εγγράφου. Το τελικό σημασιολογικό σκορ ορίζεται ως:

$$Scr_{sem}(q_{sem}, d) = \sum_{i=1}^{n} \frac{ss_{cli,d}}{n} \tag{5.2}$$

όπου $ss_{cli,d} = \frac{number\_of\_tokens\_of\_cl_i\_annotations\_over\_d}{number\_of\_tokens\_in\_d}$, n οι κλάσεις που συμμετέχουν στο σημασιολογικό ερώτημα και $ss_{cli,d}$ ο βαθμός έκτασης κάθε κλάσης cl_i σε ένα έγγραφο d.

Υβριδική αναζήτηση. Ο χρήστης συνδυάζει αναζήτηση με λέξεις κλειδιά και περιήγηση/επιλογή κλάσεων της οντολογίας, έχοντας τη δυνατότητα να διαλέξει αν το τελικό αποτέλεσμα θα είναι η τομή ή η ένωση των αποτελεσμάτων. Και στις δύο περιπτώσεις, το τελικό σκορ είναι ένα σταθμισμένο άθροισμα των δύο επιμέρους σκορ:

$$Scr_{hybr}(q_{sem}, q_{key}, d) = Scr_{sem}(q_{sem}, d) * w_{sem} + Scr_{key}(q_{key}, d) * w_{key}$$

Προχωρημένες δυνατότητες αναζήτησης

Σε αυτήν την υποενότητα παρουσιάζουμε λειτουργίες αναζήτησης που δύνανται να επιλετεστούν, αφού έχει προηγηθεί μία αρχική αναζήτηση:

- Εύρεση σχετικών εγγράφων. Ξεκινώντας από ένα έγγραφο-αποτέλεσμα d, ο χρήστης μπορεί να ψάξει όλα τα έγγραφα που έχουν επισημειωθεί με οποιαδήποτε από τις κλάσεις που χαρακτηρίζουν το έγγραφο.

- Εύρεση παρόμοιων εγγράφων. Η συγκεκριμένη είναι παραλλαγή της προηγούμενης δυνατότητας. Ξεκινώντας από ένα έγγραφο-αποτέλεσμα d, ο χρήστης μπορεί να ψάξει όλα τα έγγραφα που έχουν επισημειωθεί με οποιαδήποτε από τις κλάσεις που χαρακτηρίζουν το έγγραφο και, παράλληλα, ανήκουν ήδη στα αποτελέσματα της προηγούμενης αναζήτησης.

- Εύρεση επόμενης «γενιάς» εγγράφων. Η συγκεκριμένη δυνατότητα επιτρέπει τη μετάδοση της αναζήτησης σε χαμηλότερα επίπεδα της οντολογίας, αναζητώντας έγγραφα σε υποκλάσεις της αρχικής κλάσης αναζήτησης. Η λειτουργικότητα αυτή βρίσκει εφαρμογή, όταν η αρχική αναζήτηση είναι πολύ γενική.

- Εύρεση προηγούμενης «γενιάς» εγγράφων. Είναι η αντίστροφη λειτουργικότητα της παραπάνω. Η συγκεκριμένη δυνατότητα επιτρέπει τη μετάδοση της αναζήτησης σε υψηλότερα επίπεδα της οντολογίας, αναζητώντας έγγραφα σε υπερκλάσεις της αρχικής κλάσης αναζήτησης. Η λειτουργικότητα αυτή βρίσκει εφαρμογή, όταν η αρχική αναζήτηση είναι πολύ ειδική.

- Αναζήτηση εγγύτητας. Η συγκεκριμένη λειτουργία λαμβάνει υπόψη το επίπεδο των κλάσεων στην ιεραρχία της οντολογίας, μεταδίδοντας την αναζήτηση σε υποκλάσεις της αρχικής κλάσης.

5.2.3 Επισκόπηση συστήματος

Αρχιτεκτονική συστήματος

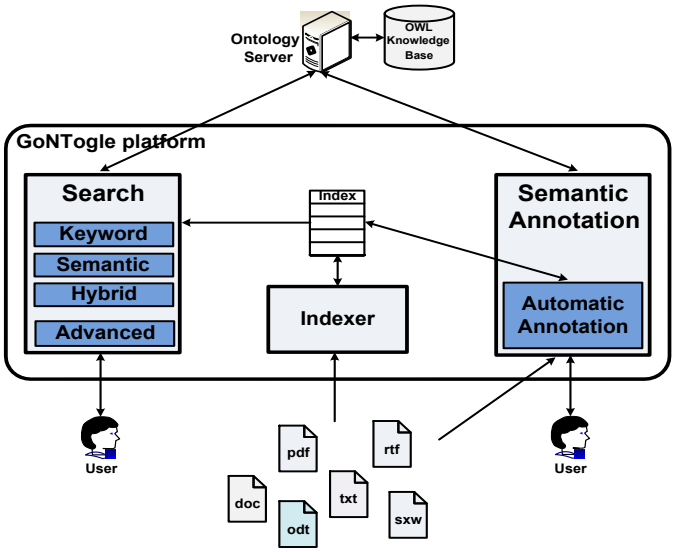
Χάρη στην κεντρικοποιημένη αρχιτεκτονική, το GoNTogle εξασφαλίζει ένα συνεργατικό περιβάλλον στους χρήστες. Οι επισημειώσεις είναι ορατές και επαναχρησιμοποιήσιμες από διάφορες ομάδες χρηστών. Η Εικόνα 5.5 παρουσιάζει την αρχιτεκτονική του συστήματος, το οποίο χωρίζεται σε τέσσερα υποσυστήματα:

1. Το Semantic Annotation Υποσύστημα, το οποίο αποτελείται από τις εξής μονάδες: (i) Document Viewer, (ii) Ontology Viewer και (iii) Annotation Editor.

2. Το Ontology Server Υποσύστημα, το οποίο αποθηκεύει την οντολογία και τις επισημειώσεις. Αποτελείται από δύο μονάδες: (i) Ontology Manager και (ii) Ontology Knowledge Base.

3. Το Indexing Υποσύστημα, υπεύθυνο για την ευρετηρίαση των εγγράφων.

4. Το Search Υποσύστημα, με το οποίο επιτελείται αναζήτηση των εγγράφων



Σχήμα 5.5: Αρχιτεκτονική συστήματος

Λειτουργία συστήματος

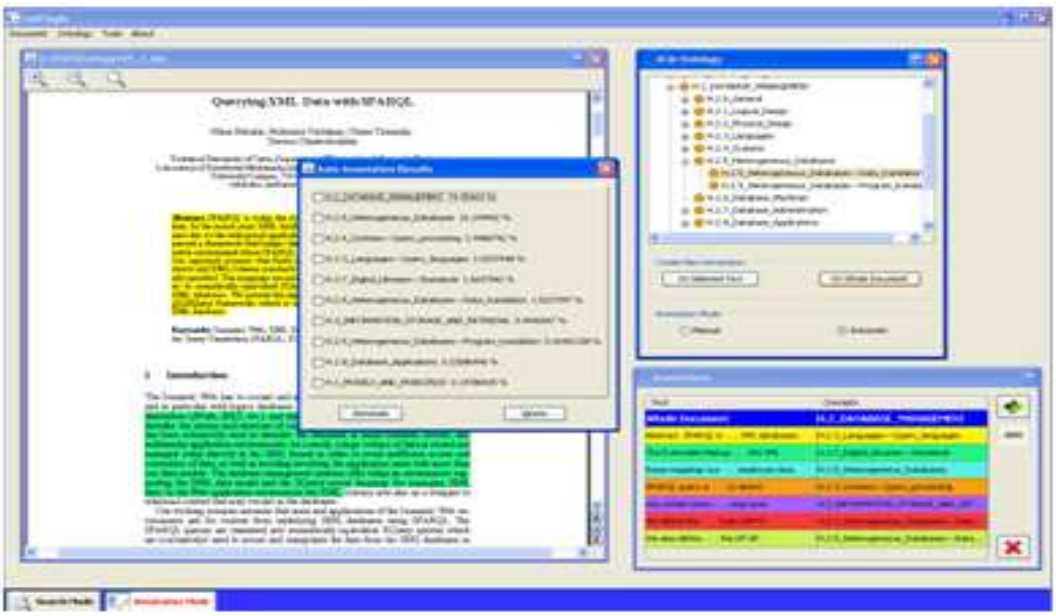
Η Εικόνα 5.6 παρουσιάζει μία οθόνη της εφαρμογής, που αφορά τη σημασιολογική επισημείωση εγγράφων. Ο χρήστης μπορεί να ανοίξει ένα έγγραφο στον Document Viewer. Επιπλέον, μπορεί να φορτώσει μία οντολογία μέσω του Ontology Viewer, επιλέγοντας μία ή περισσότερες κλάσεις για να χαρακτηρίσει όλο το έγγραφο ή την επιλεγμένη περιοχή κειμένου. Για κάθε επισημείωση εμφανίζεται στον Annotation Editor μία εγγραφή που αντιστοιχεί σε μία αποθηκευμένη στον Ontology Server επισημείωση. Ο χρήστης μπορεί να προσθαφαιρέσει κλάσεις σε μία επισημείωση ή να τη διαγράψει εντελώς. Με την επιλογή μίας επισημείωσης από τη λίστα του Annotation Editor, η εμφανιζόμενη περιοχή του εγγράφου μετακυλίεται στην αντίστοιχη επισημειωμένη περιοχή, η οποία επισημαίνεται με ειδικό χρωματισμό.

5.2.4 Πειραματική αξιολόγηση

Σε αυτήν την Ενότητα παρουσιάζεται η πειραματική αξιολόγηση (α) του μηχανισμού αυτόματης επισημείωσης εγγράφων και (β) της αποτελεσματικότητας της υβριδικής αναζήτησης εγγράφων, συγκρινόμενης με τις επιμέρους: αναζήτηση με λέξεις κλειδιά και σημασιολογική αναζήτηση.

Αυτόματη επισημείωση

Προκείμενου να αποτιμήσουμε την αποτελεσματικότητα του αλγορίθμου παραγωγής προτάσεων επισημείωσης, πραγματοποιούμε μία μελέτη χρηστών, κατά την οποία μετράμε την ακρίβεια



Σχήμα 5.6: Γραφική διεπιφάνεια συστήματος

σε κάθε θέση κατάταξης (P@n) και την ανάκληση (Recall).

Διαμόρφωση πειράματος

Αρχικά μετατρέψαμε την ταξονομία της ACM[7] σε OWL οντολογία. Η οντολογία που προέκυψε είχε τέσσερα επίπεδα και 1463 κλάσεις.

Στη συνέχεια ζητήσαμε από 15 χρήστες να συμμετέχουν στην πειραματική μελέτη. Κάθε χρήστης επέλεξε δύο περιοχές ερευνητικού ενδιαφέροντος και, για κάθε περιοχή, επέλεξε δέκα δημοσιεύσεις τις οποίες είχε διαβάσει. Προκειμένου να εκπαιδεύσουμε το σύστημα, ζητήσαμε από κάθε χρήστη να επισημειώσει κομμάτια ή όλο το κείμενο για 12 από τις συνολικά 20 δημοσιεύσεις που του αναλογούσαν, με τουλάχιστον μία κλάση της χρησιμοποιούμενης οντολογίας. Αφού κάθε χρήστης εκτέλεσε τις επισημειώσεις, οι οποίες απετέλεσαν τη φάση εκπαίδευσης της πειραματικής διαδικασίας, ζητήσαμε από τον καθένα να αξιολογήσει τις αυτόματες προτάσεις κλάσεων του συστήματος, για κάθε μία από τις 8 εναπομείνασες δημοσιεύσεις. Πριν επισκοπήσει τις προτάσεις του συστήματος, ζητήθηκε από το χρήστη να σημειώσει κάποιες κλάσεις που περίμενε να του προτείνει το σύστημα, δηλαδή πολύ σχετικές κλάσεις με τη δημοσίευση, σύμφωνα με την άποψη του χρήστη. Στη συνέχεια, ο χρήστης σημείωσε ποιες από τις προτάσεις του συστήματος ήταν σωστές, ακόμα κι αν αυτός δεν είχε σκεφτεί από πριν τις αντίστοιχες κλάσεις. Στη συνέχεια, υπολογίσαμε τις τιμές ακρίβειας και ανάκλησης για κάθε χρήστη ξεχωριστά, καθώς και μέσες τιμές. Επιπλέον, με τη μετρική *απρόσμενων κλάσεων* ($UVCS$) μετράμε το μέσο αριθμό κλάσεων που δεν είχε προβλέψει ο χρήστης, αλλά προτάθηκαν σωστά από το σύστημα.

Πειραματικά αποτελέσματα

Ο Πίνακας 5.6 παρουσιάζει, για κάθε χρήστη, τις μέσες τιμές ακρίβειας στη θέση n (P@n)

[7] http://www.acm.org/about/class/2012

πάνω στις 8 αποτιμημένες δημοσιεύσεις. Ο Πίνακας 5.7 παρουσιάζει τις αντίστοιχες τιμές ανάκλησης και απρόσμενων κλάσεων. Περιορίζουμε την ανάλυσή μας στις 5 πρώτες θέσεις, αφού στο συγκεκριμένο σενάριο δεν περιμένουμε μία δημοσίευση να αφορά πολύ περισσότερες θεματικές περιοχές.

Παρατηρούμε ότι η μέθοδός μας επιτυγχάνει πολύ υψηλές τιμές ακρίβειας και ανάκλησης, με μέση ανάκληση 93%. Οι χαμηλές τιμές που παρατηρούνται για την ακρίβεια στις θέσεις 4 και 5 (P@4, P@5) δικαιολογούνται από το γεγονός ότι, σε αρκετές περιπτώσεις, δεν υπήρχαν περισσότερες από 3 κλάσεις που να χαρακτηρίζουν μία δημοσίευση. Τέλος, φαίνεται από τη μετρική $UVCS$ ότι το σύστημα καθοδηγεί τους χρήστες στην εύρεση νέων κλάσεων επισημείωσης, τι οποίες δεν είχαν προηγουμένως σκεφτεί.

User	**P@1**	**P@2**	**P@3**	**P@4**	**P@5**
1	0.82	0.79	0.79	0.75	0.68
2	1	0.94	0.8	0.65	0.6
3	0.8	0.8	0.7	0.7	0.76
4	1	1	0.8	0.84	0.8
5	1	0.9	0.9	0.82	0.81
6	0.8	0.9	0.73	0.7	0.64
7	1	1	0.93	0.85	0.84
8	0.93	1	0.73	0.71	0.69
9	0.9	0.9	0.87	0.8	0.76
10	0.91	0.87	0.8	0.75	0.71
11	1	1	0.87	0.84	0.78
12	0.8	0.77	0.72	0.7	0.66
13	0.95	0.92	0.83	0.75	0.68
14	1	0.9	0.87	0.8	0.76
15	0.8	0.8	0.73	0.65	0.56
Avg	0.91	0.9	0.81	0.75	0.72

Πίνακας 5.6: Μέση Ακρίβεια στη θέση n για κάθε χρήστη

User	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Avg**
Recall	0.8	0.92	0.98	0.97	0.98	1	0.97	0.82	1	0.89	0.88	0.95	0.87	0.95	1	0.93
UVCS	0.4	0.2	0.2	0.4	0.4	1.2	0.2	0.2	0.2	1	0.8	0.65	0.4	1.6	0	0.52

Πίνακας 5.7: Ανάκληση και τιμή $UVCS$ για κάθε χρήστη

Αναζήτηση

Σε αυτήν την ενότητα αξιολογούμε την αποτελεσματικότητα των τύπων αναζήτησης που προσφέρονται από το πλαίσιο του GoNTogle. Η σύγκριση γίνεται με τη βοήθεια των μετρικών της ακρίβειας στη θέση n (Precision@n), ανάκλησης (Recall) και F-measure. Σε κάθε περίπτωση, η προτεινόμενη υβριδική μέθοδος αναζήτησης επιτυγχάνει καλύτερα αποτελέσματα από τις επιμέρους μεθόδους αναζήτησης με λέξεις κλειδιά και σημασιολογικής αναζήτησης.

Τα βάρη που χρησιμοποιούνται για την υβριδική αναζήτηση παίρνουν τις ακόλουθες τιμές: $w_{sem} = 0.7, w_{key} = 0.3$ για τον τελεστή AND και $w_{sem} = 0.6, w_{key} = 0.4$ για τον τελεστή OR μετά από δοκιμαστική βελτιστοποίησή τους. Διαισθητικά, αυτές οι τιμές καταδεικνύουν ότι,

στο σενάριό μας, η σημασιολογική αναζήτηση είναι ελαφρά πιο σημαντική από την αναζήτηση με λέξεις κλειδιά.

Το πειραματικό σύνολο δεδομένων αποτελείται από 300 χειροκίνητα και αυτόματα επισημειωμένα ερευνητικά κείμενα από το προηγούμενο πείραμα. Πρώτα, δημιουργήσαμε μία δεξαμενή από λέξεις κλειδιά που χαρακτηρίζουν/απαντώνται στα κείμενα και επιλέξαμε τυχαία 10 από αυτές, ώστε να τις θεωρήσουμε ως ερωτήματα. Σημειώνουμε ότι τα ερωτήματα μπορεί να περιέχουν περισσότερες από μία λέξεις. Επιπλέον, αντιστοιχίζουμε τα ερωτήματα λέξεων κλειδιών σε σημασιολογικά ερωτήματα, επιλέγοντας τις κατάλληλες (παρόμοιες) κλάσεις της οντολογίας. Με αυτόν τον τρόπο, μπορούμε να εφαρμόσουμε τους δύο τύπους αναζήτησης για τα ίδια ερωτήματα. Επιπλέον, μπορούμε να συνδυάσουμε τους δύο τύπους, σταθμίζοντας τα επιμέρους σκορ που προκύπτουν από τον κάθε τύπο, παράγοντας έτσι αποτελέσματα υβριδικής αναζήτησης. Ο συνδυασμός μπορεί να πραγματοποιηθεί τόσο με τελεστή AND (τομή των επιμέρους αποτελεσμάτων) όσο και με τελεστή OR (ένωση των επιμέρους αποτελεσμάτων). Ο Πίνακας 5.9 παρουσιάζει τα ερωτήματα λέξεων κλειδιών και τα αντίστοιχα σημασιολογικά ερωτήματα.

ID	**Keywords**	**Classes**
qkey1	knowledge discovery and privacy	K.4.1 [Public Policy Issues]: Privacy
qkey2	stream mining	H.2.8 [Database Applications]: Data mining
qkey3	RDF indexing	H.3.1 [Content Analysis and Indexing]: Indexing methods
qkey4	spatial databases	H.2.8 [Database Applications]: Spatial databases and GIS
qkey5	clustering	H.3.3 [Information Search and Retrieval]: Clustering
qkey6	spatial access	H.2.2 [Physical Design]: Access Methods
qkey7	query language	H.2.3 [Languages]: Query languages
qkey8	data model	H.2.1 [Logical Design]: Data models
qkey9	XML interoperability	D.2.12 [Interoperability]
qkey10	information integration	H.2.5 [Heterogeneous Databases]

Πίνακας 5.8: Ερωτήματα λέξεων κλειδιών και αντίστοιχα σημασιολογικά ερωτήματα

Για κάθε ερώτημα μετράμε την ποιότητα της ανάκτησης χρησιμοποιώντας την ακρίβεια στη θέση n και την ανάκληση (Precision@n, Recall). Όπως αναφέραμε παραπάνω, αξιολογούμε τέσσερις τύπους αναζήτησης: (α) αναζήτηση με λέξεις κλειδιά (qkey), (β) σημασιολογική αναζήτηση (qsem), (γ) υβριδική με τελεστή AND (hybrA) και (δ) υβριδική με τελεστή OR (hybrO). Τέλος, για κάθε τύπο αναζήτησης εξετάζουμε συνολικά τέσσερις μετρικές: Precision@n, Recall, F-measure, Precision-Recall curve.

Ο Πίνακας 5.9 παρουσιάζει τις μέσες τιμές των μετρικών Precision@n, Recall, F-measure πάνω σε όλα τα ερωτήματα. Σημειώνουμε ότι τα περισσότερα ερωτήματα στην υβριδική αναζήτηση με τελεστή AND δεν επιστρέφουν πάνω από $5-6$ αποτελέσματα (όπως μπορεί να φανεί και από τον Πίνακα 5.10). Κατά συνέπεια, η ακρίβεια για αυτόν τον τύπο αναζήτησης υπολογίζεται μόνο για τις θέσεις 1 έως 5, για όλα τα ερωτήματα.

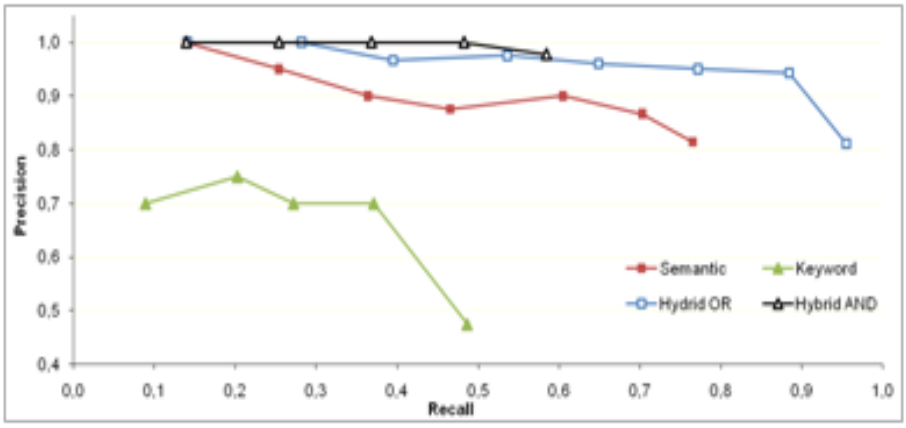
Ακρίβεια. Όπως παρατηρούμε στον Πίνακα 5.9, η υβριδική αναζήτηση (και για τους δύο τελεστές) αποδίδει καλύτερα από τις επιμέρους αναζητήσεις, σε κάθε θέση κατάταξης, με τον τύπο qhybrA να επιτυγχάνει ελαφρά καλύτερες τιμές για τις θέσεις 4 και 5. Επιπλέον, παρατηρούμε ότι η ακρίβεια της αναζήτησης με λέξεις κλειδιά μειώνεται δραστικά μετά τη

θέση 4, ενώ η ακρίβεια των υπολοίπων τύπων αρχίζει να μειώνεται μετά από τη θέση 6. Η υβριδική αναζήτηση, συγκρινόμενη με την αναζήτηση με λέξεις κλειδιά, επιτυγχάνει μέγιστη αύξηση κατά 100% στη θέση 7 και ελάχιστη αύξηση 33.3% στη θέση 2. Συγκρινόμενη με τη σημασιολογική αναζήτηση, επιτυγχάνει μέγιστη αύξηση κατά 17.2% στη θέση 10 και ελάχιστη αύξηση 0% στη θέση 1.

	P@1	**P@2**	**P@3**	**P@4**	**P@5**	**P@6**	**P@7**	**P@8**	**P@9**	**P@10**	**Recall**	**F-measure**
qkey	0.7	0.75	0.7	0.7	0.6	0.52	0.47	0.48	0.44	0.43	0.55	0.48
qsem	1	0.95	0.9	0.88	0.9	0.87	0.81	0.76	0.7	0.64	0.84	0.73
qhybrA	1	1	1	1	0.98	-	-	-	-	-	0.66	0.73
qhybrO	1	1	0.97	0.98	0.96	0.95	0.94	0.89	0.81	0.75	0.98	0.85

Πίνακας 5.9: Μέσες τιμές των μετρικών Precision@n, Recall, F-measure για όλα τα ερωτήματα, για τέσσερις διαφορετικές εκδοχές του κάθε ερωτήματος

Ανάκληση. Όπως μπορούμε να δούμε, ο τύπος qhybrO αποδίδει καλύτερα από τις επιμέρους μεθόδους, επιτυγχάνοντας ανάκληση κοντά στο 100% (98%). Ο qhybrA αποδίδει ανάμεσα στην αναζήτηση με λέξεις κλειδιά και στη σημασιολογική. Αυτό οφείλεται στο ότι ο συγκεκριμένος τύπος είναι πολύ περιοριστικός, επιστρέφοντας αρκετά λιγότερα έγγραφα από τους υπόλοιπους τύπους, κάτι το οποίο επηρεάζει αρνητικά την ανάκληση.
F-measure. Και σε αυτή τη μετρική, η υβριδική αναζήτηση ξεπερνά τους επιμέρους τύπους αναζήτησης. Συγκρίνοντας την υβριδική με τελεστή OR με αναζήτηση με λέξεις κλειδιά και με σημασιολογική, επιτυγχάνει αύξηση 77% και 16.4% αντίστοιχα. Συγκρίνοντας την υβριδική με τελεστή AND με αναζήτηση με λέξεις κλειδιά και με σημασιολογική, επιτυγχάνει αύξηση 52% και 0% αντίστοιχα.



Σχήμα 5.7: Καμπύλη ακρίβειας -ανάκλησης για το σύνολο των ερωτημάτων

Καμπύλη ακρίβειας - ανάκλησης. Η Εικόνα 5.7 παρουσιάζει τις καμπύλες ακρίβειας - ανάκλησης στο σύνολο των ερωτημάτων. Βλέπουμε ότι και οι δύο τύποι υβριδικής αναζήτησης έχουν πολύ σταθερή συμπεριφορά, επιτυγχάνοντας υψηλή ακρίβεια (κοντά στο 100%) ακόμα και για τιμές ανάκλησης μεγαλύτερες του 80%. Η υβριδική qhybrA παρουσιάζεται

ελαφρά καλύτερη της qhybrO για τιμές ανάκλησης μικρότερες του 60%. Οι επιμέρους τύποι αναζήτησης συμπεριφέρονται χειρότερα, με αισθητά χειρότερη συμπεριφορά της καμπύλης αναζήτησης με λέξεις κλειδιά.

Ο Πίνακας 5.10 παρουσιάζει, για κάθε ερώτημα τις τιμές ακρίβειας στη θέση n και ανάκλησης. Οι επιμέρους αυτές τιμές επιβεβαιώνουν τις παραπάνω παρατηρήσεις μας: οι δύο υβριδικοί τύποι αναζήτησης ξεπερνούν σε αποτελεσματικότητα τους επιμέρους τύπους αναζήτησης, επιτυγχάνοντας, σε αρκετές περιπτώσεις, πολύ υψηλές τιμές ακρίβειας και ανάκλησης ταυτόχρονα, ενώ η πιο φτωχή συμπεριφορά παρουσιάζεται από την αναζήτηση με λέξεις κλειδιά.

	Query 1				Query 2				Query 3				Query 4			
	$q_{kw}1$	$q_{sem}1$	$q_{hybrA}1$	$q_{hybrO}1$	$q_{kw}2$	$q_{sem}2$	$q_{hybrA}2$	$q_{hybrO}2$	$q_{kw}3$	$q_{sem}3$	$q_{hybrA}3$	$q_{hybrO}3$	$q_{kw}4$	$q_{sem}4$	$q_{hybrA}4$	$q_{hybrO}4$
P@1	1.00	1.00	1.00	1.00	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@2	1.00	1.00	1.00	1.00	0.50	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@3	1.00	1.00	1.00	1.00	0.33	0.67	1.00	0.67	0.67	0.67	1.00	1.00	1.00	0.67	1.00	1.00
P@4	1.00	1.00	1.00	1.00	0.50	0.50	-	0.75	0.75	0.75	1.00	1.00	1.00	0.75	1.00	1.00
P@5	1.00	1.00	1.00	1.00	0.40	0.60	-	0.60	0.60	0.80	1.00	1.00	1.00	0.80	1.00	1.00
P@6	0.83	1.00	-	1.00	0.33	0.67	-	0.67	0.50	0.67	-	0.83	0.83	0.83	1.00	1.00
P@7	0.71	1.00	-	1.00	0.29	0.57	-	0.57	0.43	0.71	-	0.86	0.71	0.71	-	1.00
P@8	0.63	1.00	-	1.00	0.25	0.50	-	0.50	0.38	0.63	-	0.75	0.75	0.75	-	1.00
P@9	0.56	1.00	-	1.00	0.22	0.44	-	0.44	0.33	0.56	-	0.67	0.67	0.78	-	0.89
P@10	0.50	0.90	-	1.00	0.20	0.40	-	0.40	0.40	0.50	-	0.60	0.60	0.80	-	0.80
Recall	0.45	0.82	0.45	0.91	0.50	1.00	0.25	1.00	0.67	0.83	0.83	1.00	0.75	1.00	0.75	1.00

	Query 5				Query 6				Query 7				Query 8			
	$q_{kw}5$	$q_{sem}5$	$q_{hybrA}5$	$q_{hybrO}5$	$q_{kw}6$	$q_{sem}6$	$q_{hybrA}6$	$q_{hybrO}6$	$q_{kw}7$	$q_{sem}7$	$q_{hybrA}7$	$q_{hybrO}7$	$q_{kw}8$	$q_{sem}8$	$q_{hybrA}8$	$q_{hybrO}8$
P@1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.00	1.00	1.00	0	1.00	1.00	1.00
P@2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.00	1.00	1.00	0	1.00	1.00	1.00
P@3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	1.00	1.00	1.00	0	1.00	1.00	1.00
P@4	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	0.25	1.00	1.00	1.00	0	1.00	1.00	1.00
P@5	0.80	0.80	1.00	1.00	0.80	1.00	0.80	1.00	0.20	1.00	1.00	1.00	0	1.00	1.00	1.00
P@6	0.67	0.83	-	1.00	0.67	0.83	0.67	1.00	0.33	1.00	1.00	1.00	0	1.00	1.00	1.00
P@7	0.57	0.71	-	1.00	0.71	0.71	0.57	1.00	0.29	1.00	1.00	1.00	0.14	1.00	-	1.00
P@8	0.63	0.63	-	0.88	0.75	0.63	0.50	1.00	0.38	1.00	1.00	1.00	0.13	0.88	-	0.88
P@9	0.56	0.56	-	0.78	0.67	0.56	0.44	0.89	0.44	0.89	1.00	1.00	0.11	0.78	-	0.78
P@10	0.50	0.50	-	0.70	0.60	0.50	0.40	0.80	0.50	0.80	0.90	1.00	0.10	0.70	-	0.70
Recall	0.63	0.63	0.63	0.88	0.67	0.63	0.50	1.00	0.50	0.80	0.90	1.00	0.14	0.70	0.89	1.00

	Query 9				Query 10			
	$q_{kw}9$	$q_{sem}9$	$q_{hybrA}9$	$q_{hybrO}9$	$q_{kw}10$	$q_{sem}10$	$q_{hybrA}10$	$q_{hybrO}10$
P@1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
P@3	1.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00
P@4	0.75	1.00	1.00	1.00	0.75	1.00	1.00	1.00
P@5	0.60	1.00	1.00	1.00	0.60	1.00	1.00	1.00
P@6	0.50	1.00	-	1.00	0.50	0.83	1.00	1.00
P@7	0.43	0.86	-	1.00	0.43	0.86	-	1.00
P@8	0.38	0.75	-	0.88	0.50	0.88	-	1.00
P@9	0.33	0.67	-	0.78	0.56	0.78	-	0.89
P@10	0.30	0.60	-	0.70	0.60	0.70	-	0.80
Recall	0.43	0.86	0.71	1.00	0.75	0.88	0.75	1.00

Πίνακας 5.10: Τιμές των μετρικών Precision@n, Recall για κάθε ερώτημα

5.2.5 Συμπεράσματα

Σε αυτήν την ενότητα, παρουσιάσαμε το GoNTogle, ένα πλαίσιο σημασιολογικής επισημείωσης και ανάκτησης εγγράφων, το οποίο συνδυάζει τεχνολογίες Ανάκτησης Πληροφορίας και Σημασιολογικού ιστού. Το GoNTogle υποστηρίζει χειροκίνητη και αυτόματη επισημείωση εγγράφων, χρησιμοποιώντας οντολογίες εννοιών. Η λειτουργικότητα αυτόματης επισημείωσης βασίζεται σε έναν υποκείμενο μηχανισμό εκμάθησης, που εκμεταλλεύεται κειμενική πληροφορία και πληροφορία επισημειώσεων (κλάσεις της οντολογίας).

Προκειμένου να ξεπεράσουμε τα μειονεκτήματα της κλασικής αναζήτησης με λέξεις κλει-

διά (πολυσημία νοημάτων, συνωνυμία) και της σημασιολογικής αναζήτησης (ανυπαρξία επισημειώσεων με κλάσεις της οντολογίας), προτείνουμε μία υβριδική μέθοδο αναζήτησης η οποία συνδυάζει τις δύο παραπάνω επιμέρους κατηγορίες αναζήτησης. Επιπλέον, ορίζουμε προχωρημένες λειτουργίες σημασιολογικής αναζήτησης, που διευκολύνουν το χρήστη, όταν θέλει να επεκτείνει την αναζήτηση σε ήδη ανακτημένα αποτελέσματα.

Επίσης, πραγματοποιήσαμε μία μελέτη χρηστών η οποία καταδεικνύει την αποτελεσματικότητα της μεθόδου αυτόματης επισημείωσης εγγράφων, καθώς και των μεθόδων υβριδικής αναζήτησης. Επόμενα βήματα βελτίωσης της συγκεκριμένης δουλειάς θα μπορούσαν να περιλαμβάνουν τα εξής: (α) Εκμετάλλευση της διαδικασίας συλλογιστικής σε οντολογίες για τη βελτίωση της αναζήτησης, (β) Ενσωμάτωση σημασιολογικών προσεγγίσεων επεξεργασίας φυσικής γλώσσας, (γ) Ενσωμάτωση υποστήριξης ελεύθερων επισημειώσεων (tagging) ταυτόχρονα με δομημένες σε οντολογίες επισημειώσεις και (δ) βελτίωση της λειτουργικότητας προβολής εγγράφων και επισημείωσης του υλοποιημένου εργαλείου.

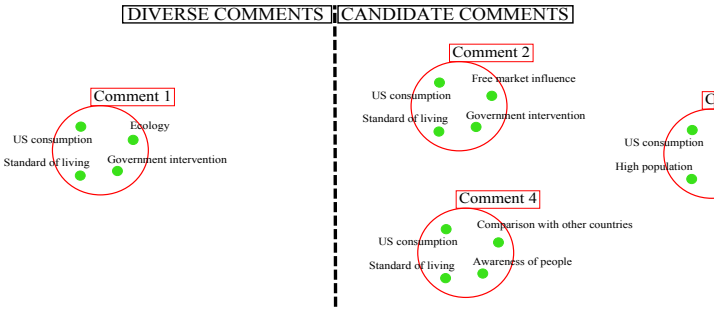
Κεφάλαιο 6

Διαφοροποιημένη Ανάκτηση Σχολίων Χρηστών σε Κοινωνικά Δίκτυα

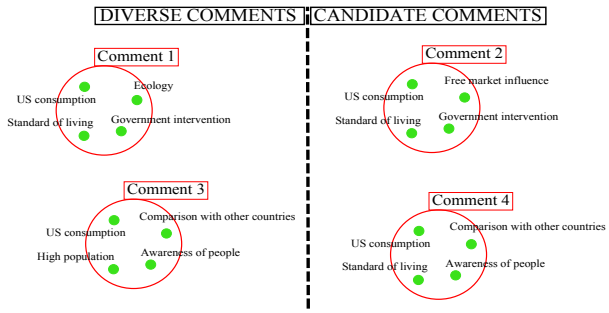
Σε αυτήν την ενότητα παρουσιάζουμε τις μεθόδους που προτείνουμε για διαφοροποίηση σχολίων χρηστών σε ειδησεογραφικά άρθρα και γενικότερα σε δεδομένα κοινωνικών δικτύων [106]. Υποστηρίζουμε ότι, αν και η διαφοροποίηση βασισμένη μόνο στο κειμενικό περιεχόμενο των στοιχείων μπορεί να επαρκεί στο σενάριο της αναζήτησης εγγράφων με λέξεις κλειδιά, όπως αποδεικνύεται από τη βιβλιογραφία, δεν είναι αρκετό, όταν πρόκειται για διαφοροποίηση σχολίων χρηστών. Για αυτό το λόγο, στο πλαίσιο που προτείνουμε, ορίζουμε εξειδικευμένα κριτήρια διαφοροποίησης που λαμβάνουν υπόψη τα χαρακτηριστικά των σχολίων, με σκοπό να παράγουμε τις αντίστοιχες διαστάσεις διαφοροποίησης με τη μορφή διανυσμάτων χαρακτηριστικών. Τα κριτήρια που ορίζουμε αντιστοιχούν στα εξής: κειμενική ομοιότητα, συναίσθημα που εκφράζεται στα σχόλια, ονοματικές οντότητες, ποιότητα γραφής των σχολίων, καθώς και συνδυασμοί από τα παραπάνω. Στη συνέχεια, εφαρμόζουμε αλγόριθμους διαφοροποίησης στα σχόλια, χρησιμοποιώντας τα παραπάνω κριτήρια. Το αποτέλεσμα της παραπάνω διαδικασίας είναι ένα υποσύνολο των αρχικών σχολίων που περιέχει ετερογενή σχόλια, που αντιπροσωπεύουν διαφορετικές εκφάνσεις του προς εξέταση άρθρου, διαφορετικά είδη συνασισθημάτων, διαφορετική ποιότητα γραφής, κτλ. Πραγματοποιούμε πειραματική αξιολόγηση των μεθόδων μας, δείχνοντας ότι τα κριτήρια που προτείνουμε επιφέρουν διακριτά πιο ετερογενή υποσύνολα διαφοροποιημένων σχολίων σε σύγκριση με τη βασική μέθοδο διαφοροποίησης μόνο με το κειμενικό περιεχόμενο. Τέλος, παρουσιάζουμε μία πρωτότυπη εφαρμογή που υλοποιεί το παραπάνω πλαίσιο.

6.1 Εισαγωγή

Στη συνέχεια παρουσιάζουμε ένα παράδειγμα που καταδεικνύει τη χρησιμότητα, αλλά και, διαισθητικά, τη λογική των μεθόδων που προτείνουμε. Για την περιγραφή μας, αλλά και για την αξιολόγηση των μεθόδων, όπως θα δούμε στη συνέχεια, υιοθετούμε την έννοια των



Σχήμα 6.1: Παράδειγμα διαφοροποίησης βασιζόμενο στις μονάδες πληροφορίας - Φάση 1.



Σχήμα 6.2: Παράδειγμα διαφοροποίησης βασιζόμενο στις μονάδες πληροφορίας - Φάση 2.

μονάδων πληροφορίας (information nuggets), που ορίζεται στο [95] και την επεκτείνουμε στο δικό μας σενάριο. Εν συντομία, θεωρούμε ως μονάδα πληροφορίας για σχόλια χρηστών *κάθε έννοια ή θεματική κατηγορία ή υποκατηγορία που σχετίζεται με το κυρίως άρθρο ή κάθε σχετική άποψη/συναίσθημα ή επέκταση των παραπάνω εννοιών και κατηγοριών, τα οποία μπορεί να εντοπιστούν, είτε στο κυρίως άρθρο, είτε στα σχόλια του.* Με βάση το παραπάνω, αντικειμενικός στόχος της διαδικασίας της διαφοροποίησης είναι να συγκεντρώσει ένα υποσύνολο σχολίων το οποίο θα περιέχει όσο το δυνατόν περισσότερες και πιο ετερογενείς μονάδες πληροφορίας που να αφορούν το κυρίως άρθρο.

Στην Εικόνα 6.1, παρουσιάζονται τέσσερα σχόλια σχετικά μέ ένα άρθρο για την «κατανάλωση στις Η.Π.Α.» - «US consumption». Θεωρούμε, χάριν ευκολίας της παρουσίασης, ότι κάθε σχόλιο περιέχει τέσσερις διαφορετικές μονάδες πληροφορίας σχετιζόμενες με το άρθρο. Έστω ότι το σχόλιο 1 έχει ήδη επιλεγεί ως το πρώτο αποτέλεσμα του διαφοροποιημένου συνόλου αποτελεσμάτων. Σε αυτήν την περίπτωση, στόχος της διαδικασίας διαφοροποίησης είναι να επιλέξει ένα από τα εναπομείναντα υποψήφια σχόλια, μεγιστοποιώντας την ετερογένεια των μονάδων πληροφορίας στο σύνολο αποτελεσμάτων που θα προκύψει. Από τα υποψήφια σχόλια, το 3 έχει 3/4 μονάδες πληροφορίας διαφορετικές από το ήδη επιλεγμένο σχόλιο 1, όντας, έτσι, το πιο «διαφορετικό» σχόλιο από το 1. Το σχόλιο 2 έχει 1/4 μονάδες διαφορετικές από το 1, ενώ το σχόλιο 4 έχει 2/4 διαφορετικές μονάδες από το 1. Έτσι, το σχόλιο 3 είναι αυτό που επιλέγεται ως το επόμενο αποτέλεσμα (βλέπε Εικόνα 6.2).

Ενώ στην εικόνα 6.1 τα υποψήφια σχόλια συγκρίνονται μόνο με το σχόλιο 1, τα πράγματα αλλάζουν στην Εικόνα 6.2. Εδώ, λόγω του ότι το σύνολο αποτελεσμάτων περιέχει τα σχόλια 1

και 3, το πιο «διαφορετικό» υποψήφιο σχόλιο είναι το 2, αφού περιέχει τη μονάδα «Freemarket influence» η οποία δεν περιέχεται ούτε στο σχόλιο 1, ούτε στο 3. Από την άλλη πλευρά, όλες οι μονάδες του σχολίου 4 περιέχονται στα σχόλια του συνόλου αποτελεσμάτων. Οπότε, το σχόλιο 2 είναι αυτό που επιλέγεται ως το επόμενο αποτέλεσμα.

Το παραπάνω παράδειγμα καταδεικνύει πώς η διαφοροποίηση/ετερογένεια ενός συνόλου σχολίων όσον αφορά σε ένα άρθρο, μπορεί να ποσοτικοποιηθεί και να αξιολογηθεί μέσω των μονάδων πληροφορίας που εξάγονται από το άρθρο και τα σχόλιά του. Παρόλα αυτά, η εργασία της αναγνώρισης εννοιών, θεμάτων και γνωμών, είναι πολύ δύσκολη, ακόμα και όταν γίνεται χειροκίνητα από κάποιον κριτή. Ακόμα και δημοφιλή εργαλεία εξαγωγής οντοτήτων, θεμάτων ή συναισθημάτων δεν μπορούν να αναγνωρίσουν όλες τις έννοιες που περιέχονται σε ένα κομμάτι κειμένου ή να αναγνωρίσουν με συνέπεια την ίδια οντότητα, η οποία περιέχεται σε διαφορετικά κομμάτια κειμένου, πιθανόν με ελαφρά διαφορετικές μορφές. Για παράδειγμα, όταν οι χρήστες χρησιμοποιούν εκφράσεις όπως «Republicans», «conservatives» ή «Bush's governance», ανάλογα με τα συμφραζόμενα του άρθρου, υπάρχει περίπτωση να αναφέρονται ακριβώς στην ίδια έννοια. Αυτό είναι πρακτικά αδύνατο να αναγνωρισθεί από μία εφαρμογή. Η κύρια συνεισφορά της δουλειάς που περιγράφεται σε αυτήν την ενότητα είναι τα εξειδικευμένα σε σχόλια χρηστών κριτήρια διαφοροποίησης που προτείνουμε, τα οποία προσπαθούν, με έμμεσο αλλά αυτοματοποιημένο τρόπο, να αιχμαλωτίσουν τις διαφορές των εννοιών ανά τα σχόλια. Τα κριτήρια που προτείνουμε αφορούν τα ακόλουθα χαρακτηριστικά των σχολίων:

- Κειμενικό περιεχόμενο. Αυτό είναι το βασικό κριτήριο διαφοροποίησης που χρησιμοποιείται και στη βιβλιογραφία όσον αφορά στη διαφοροποίηση αποτελεσμάτων αναζήτησης. Ο στόχος είναι η ανάκτηση σχολίων με όσο το δυνατόν πιο ετερογενές κειμενικό περιεχόμενο.

- Συναίσθημα. Θεωρούμε το συναίσθημα που εκφράζουν οι χρήστες στα σχόλιά τους. Μετράμε το συναίσθημα σε μία 9-βάθμια κλίμακα, που αναπαριστά αρνητικό, ουδέτερο και θετικό συναίσθημα. Στόχος είναι η συλλογή σχολίων που καλύπτουν ένα ομοιογενές φάσμα συναισθημάτων.

- Ονοματικές οντότητες. Θεωρούμε τις ονοματικές οντότητες (Πρόσωπα, Οργανισμοί, Τοποθεσίες) που εντοπίζονται στο άρθρο και στα σχόλια. Στη συνέχεια, εξετάζουμε ποιες από αυτές τις ονοματικές οντότητες αναφέρονται στο κάθε σχόλιο. Ξανά, στόχος είναι το διαφοροποιημένο σύνολο σχολίων να περιέχει όσο το δυνατόν περισσότερες/διαφορετικές ονοματικές οντότητες.

- Συναίσθημα γύρω από ονοματικές οντότητες. Για κάθε ονοματική οντότητα ενός άρθρου θεωρούμε ένα παράθυρο λέξεων γύρω από αυτήν και εξάγουμε το συναίσθημα που εκφράζεται σε αυτήν την περιοχή. Με αυτόν τον τρόπο επικεντρώνουμε την αναγνώριση συναισθήματος μόνο στις ονοματικές οντότητες.

- Ποιότητα γραφής σχολίου. Η ποιότητα του σχολίου μετράται σε μία 7-βάθμια κλίμακα και εξαρτάται από το πόσο ευανάγνωστο είναι το σχόλιο. Στόχος είναι η συλλογή σχολίων διαφορετικής ποιότητας.

- Αθροιστική ποιότητα γραφής σχολίου. Θεωρούμε, για κάθε χρήστη, όλα τα σχόλια που έχει ποστάρει συνολικά για όλα τα άρθρα και παράγουμε μία μέση ποιότητα γραφής για κάθε χρήστη. Τότε, κάθε σχόλιο αντιπροσωπεύεται από την ποιότητα γραφής του αντίστοιχου χρήστη.

Διεξάγουμε μία αναλυτική πειραματική ανάλυση που δείχνει την αποτελεσματικότητα των μεθόδων μας, ενάντια στη βασική μέθοδο της διαφοροποίησης σχολίων μόνο με βάση το κειμενικό περιεχόμενο. Η αξιολόγηση γίνεται με τη βοήθεια τριών μετρικών αξιολόγησης που ορίζουμε, με σκοπό να ποσοτικοποιήσουμε τον αριθμό των μονάδων πληροφορίας (συνολικών ή διακριτών) που καταμετρώνται σε κάθε διαφοροποιημένο σύνολο-αποτελέσμα, καθώς επίσης και την ομοιογένεια των μονάδων πληροφορίας στα διαφορετικά σύνολα σχολίων. Σχεδόν όλες οι παραλλαγές της μεθόδου μας αποδίδουν καλύτερα από τη βασική μέθοδο όσον αφορά στην κάλυψη διαφορετικών μονάδων πληροφορίας, με την καλύτερη παραλλαγή να έχει στατιστικά σημαντικά (statistically significant) καλύτερη αποτελεσματικότητα.

Συνολικά, η συνεισφορά της περιγραφόμενης δουλειάς είναι η ακόλουθη: (α) Ορίζουμε εξειδικευμένα σε σχόλια χρηστών κριτήρια διαφοροποίησης, (β) Προτείνουμε μία παραλλαγή ευριστικού αλγόριθμου διαφοροποίησης που αποδίδει πολύ κοντά στον βέλτιστο δοκιμαζόμενο αλγόριθμο, (γ) Επεκτείνουμε την έννοια των μονάδων πληροφορίας στο σενάριο των άρθρων/σχολίων, ορίζοντας διαισθητικές μετρικές για να αξιολογήσουμε την αποτελεσματικότητα των μεθόδων, (δ) Πραγματοποιούμε μία αναλυτική αξιολόγηση αποδεικνύοντας την αποτελεσματικότητα των μεθόδων μας και (ε) Υλοποιούμε τις μεθόδους μας σε ένα πρωτότυπο σύστημα που τρέχει πάνω από ένα δημοσίως διαθέσιμο σύνολο δεδομένων από άρθρα και σχόλια.

6.2 Διαδικασία διαφοροποίησης σχολίων

Σε αυτήν την ενότητα πρώτα ορίζουμε το πρόβλημα που λύνουμε. Στη συνέχεια παρουσιάζουμε αναλυτικά τα προτεινόμενα κριτήρια διαφοροποίησης και περιγράφουμε την υλοποίηση τεσσάρων αλγορίθμων διαφοροποίησης που εφαρμόζουμε. Τέλος, βασιζόμενοι στα ορισμένα κριτήρια, ορίζουμε τις συναρτήσεις ομοιότητας (τόσο για ομοιότητα όσο και για ετερογένεια στοιχείων) που χρησιμοποιούνται από τους αλγόριθμους.

6.2.1 Ορισμός προβλήματος

Προσπαθούμε να λύσουμε το πρόβλημα της ανάκτησης k-ετερογενών σχολίων χρηστών για ένα άρθρο. Συγκεκριμένα, το πρόβλημα ορίζεται ως εξής:

Ορισμός 6.2 (Διαφοροποίηση σχολίων). *Έστω A ένα άρθρο και N ένα σύνολο από σχόλια στο άρθρο. Βρες ένα υποσύνολο $S \subset N$ σχολίων που μεγιστοποιεί μία συνάρτηση στόχο f η οποία ποσοτικοποιεί την ετερογένεια των σχολίων στο S.*

Οι πιο πρόσφατες εργασίες στη διαφοροποίηση καταπιάνονται με τη διαφοροποίηση αποτελεσμάτων αναζήτησης, όσον αφορά το αντίστοιχο ερώτημα. Το δικό μας σενάριο διαφοροποίησης έχει αρκετές ομοιότητες με τη διαφοροποίηση αποτελεσμάτων αναζήτησης, για

παράδειγμα, το γεγονός ότι, και στις δύο περιπτώσεις, τα στοιχεία προς διαφοροποίηση έχουν κειμενικές περιγραφές. Επίσης, και στις δύο περιπτώσεις, πέρα από την ετερογένεια, πρέπει να ληφθεί υπόψη και η ομοιότητα των προς διαφοροποίηση στοιχείων (αποτελέσματα/σχόλια) με τον αντίστοιχο βασικό πόρο (ερώτημα/άρθρο). Παρόλα αυτά, υπάρχουν και ουσιώδεις διαφορές που επιβάλλουν την ανάγκη ανάλυσης και επέκτασης/προσαρμογής των αλγορίθμων/κριτηρίων διαφοροποίησης, ειδικά για το σενάριο της διαφοροποίησης σχολίων. Στη συνέχεια αναλύουμε εν συντομία αυτές τις διαφορές που θεωρούμε πιο κρίσιμες:

- **Βασικός πόρος.** Τα ερωτήματα είναι σύντομα σε έκταση και, τις περισσότερες φορές, αντιπροσωπεύουν μία ή λίγες ανάγκες αναζήτησης, οι οποίες είναι στενά συνδεδεμένες με τις λιγοστές λέξεις του ερωτήματος. Έτσι, στο σενάριο της αναζήτησης η διαφοροποίηση στοχεύει κυρίως στο διαχωρισμό διαφορετικών εκφάνσεων-εννοιών των λέξεων-όρων του ερωτήματος και στην παρουσίαση αποτελεσμάτων που καλύπτουν καλύτερα αυτές τις εκφάνσεις. Από την άλλη πλευρά, ένα άρθρο περιέχει πολύ περισσότερο κείμενο. Συγκεκριμένα, αποτελείται από μία ολοκληρωμένη περιγραφή ενός ή περισσοτέρων θεμάτων, που μπορεί να αναφέρονται σε επιμέρους θέματα. Έτσι, δεν υπάρχει ένας περιορισμένος αριθμός από έννοιες προς διαφοροποίηση, όπως στην αναζήτηση αποτελεσμάτων. Επίσης, οι οντότητες προς διαφοροποίηση μπορεί να μην είναι συμπαγείς έννοιες, αλλά να περιέχουν υπο-έννοιες, όπως στο παράδειγμα των Εικόνων 6.1 και 6.2, της Ενότητας 6.1.

- **Στοιχεία προς διαφοροποίηση.** Υποθέτουμε με ασφάλεια ότι τα περισσότερα αποτελέσματα αναζήτησης που επιστρέφονται από καθιερωμένες μηχανές αναζήτησης είναι σε κάποιο βαθμό σχετικά με το αντίστοιχο ερώτημα. Τα περισσότερα από αυτά αναμένεται να περιέχουν καλά δομημένο κείμενο που περιγράφει με σαφήνεια ένα ή περισσότερα θέματα που άπτονται του ερωτήματος, αφού η ποιότητα αυτών των αποτελεσμάτων έχει εξακριβωθεί από τους μηχανισμούς ταξινόμησης των αντίστοιχων μηχανών αναζήτησης. Από την άλλη πλευρά, τα σχόλια χρηστών περιέχουν πολύ λιγότερο κείμενο και μπορεί να έχουν πολύ ετερογενή ποιότητα (έλλειψη σημείων στίξης, συντομογραφίες ή αργκώ, κτλ.). Επίσης, μερικά σχόλια μπορεί να είναι απαντήσεις ή συνέχειες προηγούμενων σχολίων.

- **Άποψη.** Τις περισσότερες φορές τα αποτελέσματα αναζήτησης περιέχουν περιγραφές εννοιών που σχετίζονται με τους όρους του ερωτήματος. Από την άλλη πλευρά, τις περισσότερες φορές τα σχόλια εκφράζουν, σε διαφορετικό βαθμό, απόψεις και συναισθήματα σχετικά με τις προς συζήτηση έννοιες.

- **Ονοματικές οντότητες.** Στο σενάριο των άρθρων/σχολίων, οι ονοματικές οντότητες αναμένεται να παίζουν σημαντικό ρόλο. Είναι αναμενόμενο αρκετές από τις έννοιες που περιγράφονται σε ένα άρθρο να σχετίζονται με μία ή περισσότερες ονοματικές οντότητες, καθώς επίσης και οι χρήστες να αναφέρονται σε ονοματικές οντότητες, όταν σχολιάζουν ένα θέμα.

6.2.2 Κριτήρια διαφοροποίησης

Στη συνέχεια περιγράφουμε τα κριτήρια που ορίζουμε για διαφοροποίηση σχολίων χρηστών σε άρθρα.

Κειμενικό περιεχόμενο.

Θεωρούμε την κειμενική περιγραφή του σχολίου, που είναι το βασικό κριτήριο διαφοροποίησης στις περισσότερες δουλειές που καταπιάνονται με τη διαφοροποίηση. Η σημασία του κειμένου στη διαφοροποίηση είναι προφανής. Για κάθε σχόλιο, κατασκευάζουμε το διάνυσμα όρων (term vector) του, με κάθε στοιχείο του διανύσματος να αντιστοιχεί σε κάθε διακριτό όρο που υπάρχει σε όλο το σώμα κειμένων που περιέχει τα άρθρα και τα σχόλια. Η τιμή του κάθε στοιχείου του διανύσματος υπολογίζεται κανονικοποιώντας τη συχνότητα του αντίστοιχου όρου μέσα στο σχόλιο, με βάση το συνολικό αριθμό όρων του σχολίου.

Συναίσθημα.

Εξετάζουμε το συναίσθημα που εκφράζεται από τους χρήστες μέσω των σχολίων τους. Θεωρούμε ότι το συναίσθημα (θετικό, αρνητικό ή ουδέτερο) είναι παράγοντας διαφοροποίησης, αφού εκφράζει τις απόψεις των χρηστών πάνω στα θέματα του άρθρου. Υπό αυτήν την έννοια, η παραγωγή ενός συνόλου σχολίων που καλύπτουν διαφορετικές διαβαθμίσεις συναισθήματος και, κατά προτίμηση, με ομοιόμορφο τρόπο, ευνοεί την ετερογένεια.

Ορίζουμε εννιά κλάσεις συναισθήματος μέσα στο διάστημα $[-4, 4]$, με -4 να υποδηλώνει πολύ αρνητικό συναίσθημα, 4 πολύ θετικό και 0 ουδέτερο συναίσθημα. Η παραπάνω σύμβαση κληρονομείται από το εργαλείο που χρησιμοποιήσαμε για εξαγωγή συναισθήματος (Sentistrength - [103]). Παρόλα αυτά μπορεί να εφαρμοστεί οποιοσδήποτε αριθμός από κλάσεις/διαβαθμίσεις συναισθήματος, χωρίς να αλλάξει η λογική του κριτηρίου. Για κάθε σχόλιο, ορίζουμε δύο διαφορετικούς τρόπους εξαγωγής συναισθήματος από την κειμενική περιγραφή του:

- **Μέγιστο/ελάχιστο συναίσθημα**. Εξετάζουμε συνολικά το κείμενο του σχολίου. Από αυτό εξάγουμε το μέγιστο θετικό και το ελάχιστο αρνητικό συναίσθημα.
- **Μέσο συναίσθημα**. Εξετάζουμε κάθε πρόταση του σχολίου ξεχωριστά, εξάγοντας τις αντίστοιχες τιμές θετικού και αρνητικού συναισθήματος. Στη συνέχεια, παίρνουμε τη μέση τιμή θετικού και αρνητικού συναισθήματος πάνω στις προτάσεις του σχολίου.

Η εξαγωγή συναισθήματος βασίζεται σε συγκεκριμένες λέξεις που εντοπίζονται στο κείμενο και αντιπροσωπεύουν θετικό ή αρνητικό συναίσθημα. Προτείνουμε τους παραπάνω δύο διαφορετικούς τρόπους εξαγωγής συναισθήματος, έτσι ώστε να μπορούμε να αναγνωρίσουμε διαφορετικές όψεις του εκφραζόμενου συναισθήματος. Για παράδειγμα, ένα σχόλιο μπορεί να περιέχει μόνο μία πρόταση με πολύ θετικό συναίσθημα σχετιζόμενο με ένα θέμα του άρθρου. Από την άλλη πλευρά, το υπόλοιπο σχόλιο μπορεί να είναι συνολικά αρνητικό προς το σύνολο

του άρθρου. Με τη διάκριση που προτείνουμε, μπορούμε να εντοπίσουμε αυτή τη διαφορά, μέσω της εξαγωγής μέσου συναισθήματος.

Αφού πραγματοποιηθεί η εξαγωγή συναισθήματος, για κάθε σχόλιο κατασκευάζουμε δύο διανύσματα μεγέθους εννιά χαρακτηριστικών, όπου το κάθε χαρακτηριστικό αντιστοιχεί σε μία βαθμίδα συναισθήματος. Κάθε χαρακτηριστικό μπορεί να πάρει τιμή 1 ή 0, ανάλογα με το αν το αντίστοιχο συναίσθημα έχει εντοπιστεί ή όχι.

Ονοματικές οντότητες.

Εξετάζουμε τις ονοματικές οντότητες που υπάρχουν στο κείμενο του άρθρου και των σχολίων. Ξεχωρίζουμε τρεις κατηγορίες: Πρόσωπα, Οργανισμούς και Τοποθεσίες. Θεωρούμε τις ονοματικές οντότητες σημαντικές για τη διαφοροποίηση, αφού πολλές φορές τα ειδησεογραφικά άρθρα περιστρέφονται γύρω από αυτές. Ακόμα και όταν ένα άρθρο αφορά περιστατικά ή καταστάσεις, συνήθως εμπλέκονται πρόσωπα ή τοποθεσίες. Με δεδομένο αυτό, είναι σημαντικό ένα διαφοροποιημένο σύνολο σχολίων να περιλαμβάνει όσο το δυνατόν περισσότερες διαφορετικές ονοματικές οντότητες.

Για κάθε μία από τις τρεις παραπάνω κατηγορίες ονοματικών οντοτήτων, ορίζουμε ξεχωριστά διανύσματα χαρακτηριστικών, με κάθε χαρακτηριστικό να αντιστοιχεί σε μία διακριτή ονοματική οντότητα που έχει βρεθεί στο άρθρο ή στα σχόλιά του. Για κάθε σχόλιο, οι τιμές του κάθε χαρακτηριστικού του διανύσματος είναι οι συχνότητες των αντιστοίχων ονοματικών οντοτήτων μέσα σε αυτό. Επιπλέον, θεωρούμε ένα συναθροιστικό διάνυσμα που περιέχει όλες τις ονοματικές οντότητες, από όλες τις κατηγορίες, ως χαρακτηριστικά. Έτσι, συνολικά, ορίζουμε τέσσερα διανύσματα που αντιπροσωπεύουν τις ονοματικές οντότητες για κάθε σχόλιο.

Συναίσθημα γύρω από ονοματικές οντότητες.

Εξετάζουμε το συναίσθημα που εντοπίζεται κοντά σε ονοματικές οντότητες. Για κάθε ονοματική οντότητα θεωρούμε ένα παράθυρο ± 5 λέξεων γύρω από αυτήν και εξάγουμε το συναίσθημα μόνο για τη συγκεκριμένη περιοχή. Σκοπός είναι η εκλέπτυνση της εξαγωγής συναισθήματος, επικεντρώνοντας στις ονοματικές οντότητες, οι οποίες αναμένεται να είναι πιο σημαντικές από τους υπόλοιπους όρους του κειμένου του σχολίου.

Στη συνέχεια, θεωρούμε το διάνυσμα ονοματικών οντοτήτων και, επεκτείνοντάς το, ορίζουμε ένα νέο διάνυσμα που στη θέση κάθε χαρακτηριστικού ονοματικής οντότητας θεωρεί εννιά χαρακτηριστικά, ένα για κάθε βαθμίδα συναισθήματος.

Ποιότητα γραφής σχολίου.

Θεωρούμε ότι η ποιότητα γραφής σχολίου είναι παράγοντας διαφοροποίησης, αφού εκφράζει την καταληπτότητα του ίδιου του σχολίου. Έτσι, είναι σημαντικό ένα σύνολο από σχόλια να περιλαμβάνει διαφορετικά επίπεδα ποιότητας. Η καταληπτότητα ενός κειμένου εκφράζεται μέσω μίας βαθμολόγησης που προκύπτει από αντίστοιχη φόρμουλα που λαμβάνει υπόψη ποιοτικά χαρακτηριστικά ενός κειμένου, όπως μήκος λέξεων και προτάσεων. Η πιο διαδεδομένη

από αυτές τις φόρμουλες είναι η Flesch Reading Ease Score[1], η οποία συνδυάζει μέσο αριθμό συλλαβών ανά λέξη και μέσο μήκος πρότασης για να παράγει μία βαθμολογία καταληπτότητας. Συγκεκριμένα, παράγει ένα σκορ στο διάστημα $[0, 100]$, με υψηλότερες τιμές να υποδεικνύουν ευκολότερα κείμενα. Σε αυτό το διάστημα ορίζουμε επτά κλάσεις αναγνωσιμότητας. Για κάθε σχόλιο εφαρμόζουμε τη φόρμουλα και αναθέτουμε μία κλάση αναγνωσιμότητας σε αυτό. Στη συνέχεια, κατασκευάζουμε ένα διάνυσμα επτά χαρακτηριστικών και αναθέτουμε τιμή 1 στο χαρακτηριστικό που αντιπροσωπεύει την αναγνωσιμότητα του σχολίου και 0 στα υπόλοιπα χαρακτηριστικά.

Συναθροιστική ποιότητα γραφής σχολίου.

Θεωρούμε κάθε σχόλιο χρήστη ξεχωριστά και εξάγουμε το αντίστοιχο σκορ καταληπτότητας. Στη συνέχεια, παίρνουμε το μέσο όρο καταληπτότητας για όλα τα σχόλια, για τον κάθε χρήστη, και χαρακτηρίζουμε κάθε σχόλιο του χρήστη με το παραπάνω μέσο σκορ και το αντίστοιχο διάνυσμα καταληπτότητας.

6.2.3 Ερμηνεία των συναρτήσεων-στόχων και αλγόριθμοι διαφοροποίησης

Παρακάτω περιγράφουμε τους τέσσερις ευριστικούς αλγόριθμους διαφοροποίησης που υλοποιήσαμε στην τρέχουσα δουλειά. Από αυτούς, τρεις έχουν παρουσιαστεί σε προηγούμενη εργασία [96] (*MAXSUM1, MAXMIN, MONO-OBJECTIVE*), ενώ ο τέταρτος είναι μία παραλλαγή άπληστου αλγορίθμου που προτείνουμε, η οποία υλοποιεί τη λογική *Max-Sum*.

Η λογική της συνάρτησης στόχου *Max-Sum* είναι η μεγιστοποίηση όλων των ανά δύο αποστάσεων μεταξύ όλων των στοιχείων του διαφοροποιημένου συνόλου S. Οι Αλγόριθμοι 3 και 4 δίνουν δύο προσεγγιστικές λύσεις για την παραπάνω συνάρτηση στόχο.

Ο Αλγόριθμος 3 παρουσιάζεται στο [96]. Ο αλγόριθμος, σε κάθε βήμα, εξετάζει τις ανά δύο αποστάσεις μεταξύ των υποψήφιων στοιχείων και επιλέγει το ζεύγος με τη μεγαλύτερη μεταξύ του απόσταση, εισάγοντας τα αντίστοιχα δύο στοιχεία στο διαφοροποιημένο σύνολο S. Δηλαδή, ο αλγόριθμος χωρίζει το σύνολο των στοιχείων σε διαφοροποιημένα και υποψήφια και πραγματοποιεί τους υπολογισμούς του, σε κάθε βήμα, μόνο με βάση τα υποψήφια στοιχεία.

Algorithm 3 Produce diverse set of comments with MAXSUM1

Input: Set of candidate comments T, size of diverse set k

Output: Set of diverse comments S

$S = \emptyset$

for $i = 1 \rightarrow \lfloor \frac{k}{2} \rfloor$ **do**

Find $(u, v) = argmax_{x,y \in T} d(x, y)$

Set $S = S \cup \{u, v\}$

Set $T = T \setminus \{u, v\}$

end for

If k is odd, add an arbitrary document to s

[1] http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php

Στην τρέχουσα εργασία, υλοποιούμε, επίσης, τον αλγόριθμο 4 για να προσεγγίσουμε τη συνάρτηση στόχο *Max-Sum*. Σημειώνουμε ότι παραλλαγές της γενικής λογικής του αλγορίθμου μπορεί να έχουν προταθεί και σε άλλες δουλειές, για παράδειγμα στο [97]. Παρόλα αυτά, μπορεί να υπάρχουν διαφορές στην αρχικοποίηση του αλγορίθμου, καθώς και στον ακριβή ορισμό της συνάρτησης απόστασης - ομοιότητας (distance-similarity function). Ο συγκεκριμένος αλγόριθμος αρχικοποιεί το διαφοροποιημένο σύνολο αποτελεσμάτων S εισάγοντας στην αρχή το υποψήφιο σχόλιο με τη μεγαλύτερη κειμενική ομοιότητα με το άρθρο. Στη συνέχεια, σε κάθε βήμα, επιλέγει το υποψήφιο σχόλιο με την μεγαλύτερη απόσταση από το κεντροειδές (centroid) του συνόλου S, δηλαδή το υποψήφιο σχόλιο που μεγιστοποιεί την ακόλουθη ποσότητα:

$$d(u, S) = \frac{1}{|S|} \sum_{x \in S} d(u, x) \tag{6.1}$$

Αν και ο αλγόριθμος στοχεύει στη μεγιστοποίηση της ίδιας συνάρτησης στόχου, η κύρια διαφορά του με τον αλγόριθμο 3 είναι ότι, σε κάθε βήμα, εξετάζει αποστάσεις μεταξύ υποψηφίων και ήδη επιλεγμένων σχολίων.

Algorithm 4 Produce diverse set of comments with MAXSUM2

Input: Set of candidate comments T, size of diverse set k

Output: Set of diverse comments S

$S = \emptyset$

Find the most relevant comment u and set $S = \{u\}$

For any $x \in T \setminus S$, define $d_{MAX}(x, S) = d(x, c_s)$ where c_s the centroid of the comments contained in S

while $|S| < k$ **do**

FIND $u = argmax_{x \in T} d_{MAX}(x, S)$

Set $S = S \cup \{u\}$

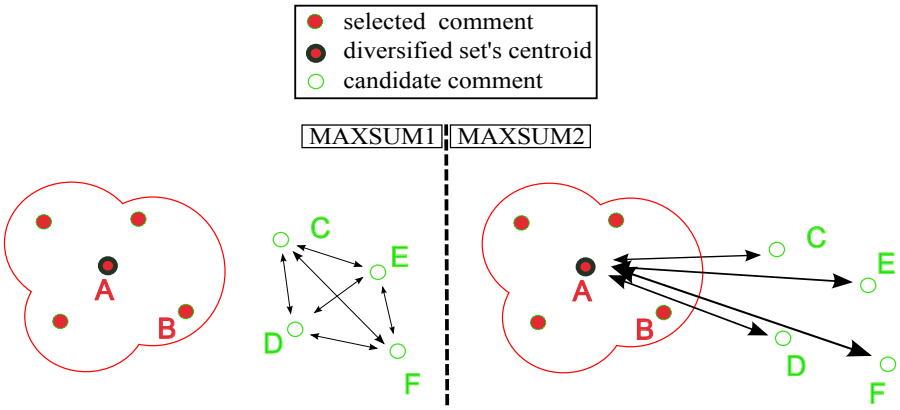
Set $T = T \setminus \{u\}$

Update c_s

end while

Επιδεικνύουμε με γραφικό τρόπο τη λογική των δύο αλγορίθμων στην Εικόνα 6.3. Έστω η γενική περίπτωση όπου ένας αριθμός σχολίων έχει ήδη επιλεγεί και εισαχθεί στο διαφοροποιημένο σύνολο S (επιλεγμένα σχόλια). Τα υποψήφια σχόλια αναπαρίστανται ως κενοί κύκλοι, ενώ τα ήδη επιλεγμένα ως γραμμοσκιασμένοι κύκλοι. Ο Αλγόριθμος *MAXSUM1* θα εξετάσει όλες τις ανά δύο αποστάσεις μεταξύ των υποψήφιων σχολίων και θα επιλέξει τα σχόλια C και F, που είναι τα πιο απόμακρα μεταξύ τους. Ο Αλγόριθμος *MAXSUM2* θα συγκρίνει όλα τα υποψήφια σχόλια με το σχόλιο A (κεντροειδές του συνόλου S) και θα επιλέξει το πιο απόμακρο υποψήφιο, δηλαδή το F. Σημειώνουμε ότι, στο επόμενο βήμα, το κεντροειδές του S επανυπολογίζεται, οπότε οι αποστάσεις των υποψήφιων σχολίων από το νέο κεντροειδές A' θα αλλάξουν.

Η λογική της συνάρτησης Max-Min είναι η μεγιστοποίηση της απόστασης μεταξύ των



Σχήμα 6.3: Αλγόριθμοι διαφοροποίησης 3 και 4

δύο πιο κοντινών σημείων μέσα στο διαφοροποιημένο, τελικό σύνολο σχολίων S. Τέλος, η συνάρτηση στόχος Mono-objective στοχεύει στην ταυτόχρονη μεγιστοποίηση τόσο της ομοιότητας ενός στοιχείου με το βασικό πόρο (σχόλιο με άρθρο), όσο και της απόστασης μεταξύ των στοιχείων-σχολίων.

Ο Αλγόριθμος 5 προσεγγίζει τη συνάρτηση στόχο Max-Min. Ο αλγόριθμος (που παρουσιάστηκε στο [96]) αρχικοποιεί το διαφοροποιημένο σύνολο όπως ακριβώς και ο Αλγόριθμος 3. Στη συνέχεια, σε κάθε βήμα, για κάθε υποψήφιο σχόλιο, βρίσκει το κοντινότερο του σχόλιο από το σύνολο S και υπολογίζει την απόστασή τους d_{MIN}. Το υποψήφιο σχόλιο με τη μέγιστη απόσταση d_{MIN} επιλέγεται για εισαγωγή στο S. Σημειώνουμε ότι, στην υλοποίησή μας, αλλάξαμε τη διαδικασία αρχικοποίησης, έτσι ώστε να ταυτίζεται με τον αλγόριθμο 2.

Algorithm 5 Produce diverse set of comments with MAXMIN

Input: Set of candidate comments T, size of diverse set k

Output: Set of diverse comments S

$S = \emptyset$

Find the most relevant comment u and set $S = \{u\}$

For any $x \in T \setminus S$, define $d_{MIN}(x, S) = \min_{u \in S} d(x, u)$

while $|S| < k$ **do**

FIND $u = argmax_{x \in T} d_{MIN}(x, S)$

Set $S = S \cup \{u\}$

Set $T = T \setminus \{u\}$

end while

Τέλος, ο Αλγόριθμος 6 προσεγγίζει τη συνάρτηση στόχο Mono-Objective. Ο αλγόριθμος, κατά την αρχικοποίηση, υπολογίζει ένα σκορ απόστασης για κάθε υποψήφιο σχόλιο. Το σκορ αυτό σταθμίζει την ομοιότητα του σχολίου με το άρθρο και την απόστασή του από όλα τα υπόλοιπα σχόλια. Τα παραπάνω σκορ μένουν σταθερά κατά την υπόλοιπη διαδικασία του αλγορίθμου. Στη συνέχεια, σε κάθε βήμα, επιλέγεται προς εισαγωγή στο S το σχόλιο με το μεγαλύτερο σκορ.

Η λογική των Αλγορίθμων 5 και 6 φαίνεται στην Εικόνα 6.4. Ο *MAXMIN* θα εξετάσει,

Algorithm 6 Produce diverse set of comments with MONO-OBJECTIVE

Input: Set of candidate comments T, size of diverse set k

Output: Set of diverse comments S

$S = \emptyset$

for each $x_i \in T$ **do**

Calculate $d(x_i) = w(x_i) + \frac{\lambda}{|T|-1} \sum_{v \epsilon T} d(x_i, v)$

end for

while $|S| < k$ **do**

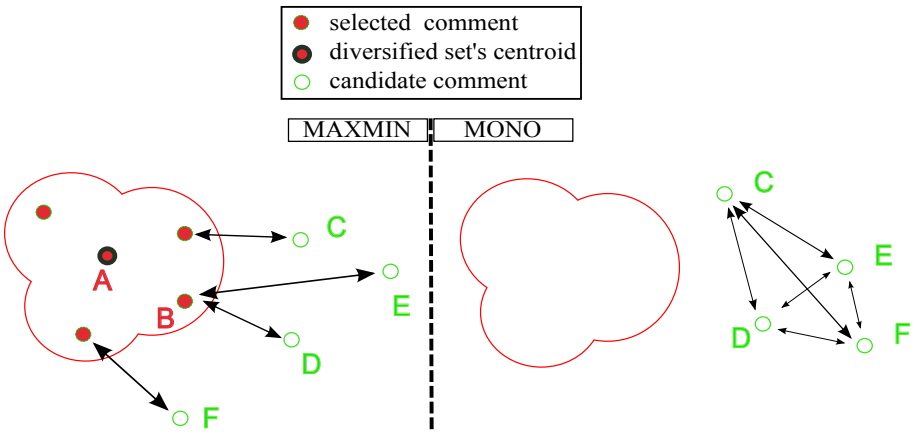
Find the candidate comment $u = \text{argmax} d(x_i)$

Set $S = S \cup \{u\}$

Set $T = T \setminus \{u\}$

end while

σε κάθε βήμα, όλες τις ανά δύο αποστάσεις μεταξύ των υποψηφίων και των επιλεγμένων σχολίων. Στη συνέχεια, θα επιλέξει το E ως το σχόλιο με την μέγιστη ελάχιστη απόσταση από κάποιο επιλεγμένο σχόλιο, προς εισαγωγή στο S. Σημειώνουμε ότι, σε κάθε επόμενο βήμα, όλες οι αποστάσεις πρέπει να επανυπολογιστούν, αφού ένα νέο σχόλιο εισέρχεται στο S. Ο *MONO-OBJECTIVE* θα υπολογίσει, κατά την αρχικοποίηση, ένα σκορ για κάθε υποψήφιο σχόλιο και, έπειτα, θα αρχίσει να εισάγει σχόλια στο S, βασιζόμενος στα αρχικώς υπολογισμένα σκορ. Στην περίπτωσή μας, το σχόλιο C θα είναι το πρώτο προς εισαγωγή, καθώς είναι προφανές ότι έχει τη μέγιστη μέση απόσταση από όλα τα άλλα υποψήφια σχόλια.



Σχήμα 6.4: Αλγόριθμοι διαφοροποίησης 5 και 6

Τα παραπάνω παραδείγματα καταδεικνύουν ότι διαφορετικές μέθοδοι μπορεί να επιφέρουν διαφορετικά τελικά αποτελέσματα, δηλαδή διαφορετικά σύνολα διαφοροποιημένων αποτελεσμάτων. Μία άλλη σημαντική παρατήρηση είναι ότι οι αλγόριθμοι *MAXSUM2* και *MAXMIN* επανυπολογίζουν αποστάσεις μεταξύ υποψήφιων και ήδη επιλεγμένων σχολίων σε κάθε βήμα. Ο *MAXSUM1* επίσης επανυπολογίζει αποστάσεις, αλλά μόνο μεταξύ υποψήφιων στοιχείων και ο *MONO-OBJECTIVE* διατηρεί και χρησιμοποιεί τις αρχικά υπολογισμένες αποστάσεις. Για αυτό το λόγο, περιμένουμε να υπάρχουν διαφορές στην αποτελεσματικότητα, όταν θα

συγκριθούν μεταξύ τους οι αλγόριθμοι.

Στη συνέχεια, περιγράφουμε τις συναρτήσεις αποστάσεων που εφαρμόζονται από τον κάθε αλγόριθμο.

6.2.4 Συναρτήσεις αποστάσεων

Στην Ενότητα 6.2.2 περιγράφηκαν οι υλοποιήσεις των κριτηρίων διαφοροποίησης, μέσω διανυσμάτων χαρακτηριστικών που αντιπροσωπεύουν διάφορες εκφάνσεις των σχολίων. Οι αλγόριθμοι διαφοροποίησης χρησιμοποιούν αυτά τα διανύσματα για να υπολογίζουν σε κάθε βήμα ένα αθροιστικό σκορ διαφοροποίησης για κάθε σχόλιο. Αυτό το σκορ, στη συνέχεια, σταθμίζεται με το σκορ ομοιότητας του σχολίου με το άρθρο για να παραγάγει ένα τελικό σκορ, από το οποίο καθορίζεται η επιλογή του επόμενου σχολίου-αποτελέσματος.

Προκειμένου να παράγουμε σκορ διαφοροποίησης, πρέπει να ορίζουμε μία συνάρτηση η οποία θα μετράει την απόσταση μεταξύ δύο στοιχείων (σχολίων). Υιοθετούμε την ευρέως χρησιμοποιούμενη συνάρτηση ομοιότητας συνημιτόνου (cosine similarity function) και ορίζουμε το σκορ διαφοροποίησης μεταξύ δύο στοιχείων, u, v, ως προς μία διάσταση-κριτήριο διαφοροποίησης i, ως εξής:

$$d_i(u, v) = 1 - \cos_i(u, v)$$

όπου $\cos(u, v)$ είναι κανονικοποιημένο στο διάστημα $[0, 1]$. Σημειώνουμε ότι η κανονικοποίηση γίνεται στο επίπεδο του κάθε κριτηρίου ξεχωριστά. Δηλαδή, υπολογίζουμε τη μέγιστη τιμή που μπορεί να πάρει κάποιο σχόλιο ενός συγκεκριμένου κριτηρίου και διαιρούμε τα αντίστοιχα σκορ των υπολοίπων σχολίων με αυτό, για κάθε κριτήριο ξεχωριστά.

Παρόλα αυτά, η ετερογένεια μεταξύ των σχολίων δεν είναι ο μόνος στόχος: τα σχόλια θα πρέπει να είναι, επιπλέον, σε κάποιο βαθμό σχετικά με το άρθρο. Έτσι, το τελικό σκορ για κάθε υποψήφιο σχόλιο είναι το σταθμισμένο άθροισμα του συνολικού σκορ διαφοροποίησής του και του σκορ ομοιότητάς του με το άρθρο. Ορίζουμε το σκορ ομοιότητας ενός σχολίου u με ένα άρθρο A, εφαρμόζοντας πάλι τη συνάρτηση συνημιτόνου στα διανύσματα όρων του άρθρου και του σχολίου:

$$r(u, A) = \cos(u, A)$$

Σημειώνουμε ότι και αυτό το σκορ κανονικοποιείται στο διάστημα $[0, 1]$.

Ανάλογα με τον αλγόριθμο διαφοροποίησης που εφαρμόζεται ορίζουμε τέσσερις τύπους που παράγουν το τελικό σκορ για κάθε υποψήφιο σχόλιο u, προκειμένου να επιλεγεί για εισαγωγή στο διαφοροποιημένο υποσύνολο S, όσον αφορά ένα άρθρο A το οποίο έχει αρχικό σύνολο υποψήφιων σχολίων T:

$$score_{MAXSUM1}(u, v, A) = (1 - w) \cdot \frac{r(u, A) + r(v, A)}{2} + w \cdot \sum_{i=1}^{4} \lambda_i \cdot d_i(u, v)$$

όπου (u, v) είναι ένα ζεύγος σχολίων, αφού αυτός ο αλγόριθμος εξετάζει ζεύγη σχολίων προς εισαγωγή, i είναι η διάσταση-κριτήριο διαφοροποίησης, $w \epsilon [0, 1]$ είναι το βάρος του συνολικού

σκορ διαφοροποίησης, έναντι του σκορ ομοιότητας με το άρθρο και $\lambda_i \epsilon [0,1]$ είναι το βάρος κάθε επιμέρους σκορ διαφοροποίησης, με $\sum_{i=1}^{4} \lambda_i = 1$.

$$score_{MAXSUM2}(u, A) = (1 - w) \cdot r(u, A) + w \cdot \sum_{i=1}^{4} \lambda_i \cdot d_i(u, C_i)$$

όπου C_i είναι το κεντροειδές του τρέχοντος διαφοροποιημένου συνόλου όσον αφορά τη διάσταση διαφοροποίησης i.

$$score_{MAXMIN}(u, A) = (1 - w) \cdot r(u, A) + w \cdot \sum_{i=1}^{4} \lambda_i \cdot d_i(u, \min v_{iu})$$

όου $\min v_{iu}$ είναι το σχόλιο από το τρέχον διαφοροποιημένο σύνολο με τη μικρότερη απόσταση από το υποψήφιο σχόλιο u.

$$score_{MONO}(u, A) = (1 - w) \cdot r(u, A) + w \cdot \sum_{i=1}^{4} \lambda_i \cdot \frac{1}{|T| - 1} \sum_{v \epsilon T} d(u, v)$$

6.3 Περιγραφή συστήματος

Σε αυτήν την ενότητα δίνουμε κάποιες τεχνικές λεπτομέρειες του συστήματος διαφοροποίησης που υλοποιήσαμε και περιγράφουμε συνοπτικά τη λειτουργικότητα της γραφικής διεπιφάνειας του εργαλείου.

Διαιρούμε τη διαδικασία της διαφοροποίησης σε δύο στάδια: Προεπεξεργασία και Εκτέλεση διαφοροποίησης. Η προεπεξεργασία περιλαμβάνει το κατέβασμα ειδησεογραφικών άρθρων και των αντιστοίχων σχολίων, την επεξεργασία/ανάλυσή τους ώστε να εξαχθούν τα διανύσματα χαρακτηριστικών των κριτηρίων διαφοροποίησης, αλλά και τα διανύσματα όρων για μέτρηση της ομοιότητας σχολίων-άρθρων και την αποθήκευσή τους στη βάση του συστήματος. Η φάση εκτέλεσης της διαφοροποίησης περιλαμβάνει το τρέξιμο των αλγορίθμων διαφοροποίησης πάνω στα παραγμένα διανύσματα χαρακτηριστικών. Όλα τα δεδομένα, πριν και μετά την επεξεργασία, αποθηκεύονται σε μία σχεσιακή βάση (MySQL), το σχήμα της οποίας φαίνεται στον Πίνακα 6.1.

Η εφαρμογή υλοποιείται σε γλώσσα Java. Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από τη διαδικτυακή σελίδα της εφημερίδας NY Times, χρησιμοποιώντας τα αντίστοιχα APIs. Για την εξαγωγή συναισθήματος από τα σχόλια χρησιμοποιήθηκε το εργαλείο SentiStrength[2] ([103]), ενώ για την εξαγωγή ονοματικών οντοτήτων το εργαλείο Stanford Named Entity Recognizer[3] ([102]).

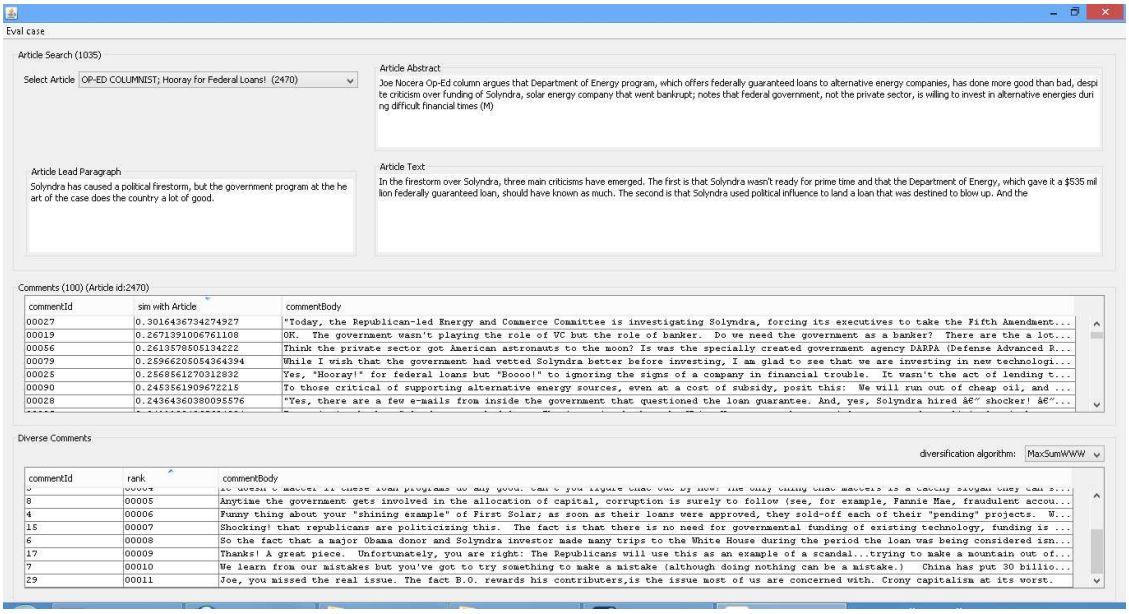
Η Εικόνα 6.5 παρουσιάζει μία οθόνη της υλοποιημένης εφαρμογής. Στο πάνω μέρος της γραφικής διεπιφάνειας «Article Search» ο χρήστης μπορεί να επιλέξει ένα άρθρο και να δει όλη τη διαθέσιμη για αυτό πληροφορία (τίτλο, περίληψη, βασική παράγραφο). Αφού επιλεγεί ένα άρθρο, τα σχόλιά του εμφανίζονται στο κάτω μέρος της οθόνης, ταξινομημένα ανάλογα

[2] http://sentistrength.wlv.ac.uk/

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

Table name	Description
article_data	Stores the text and metadata of the articles
comment_data	Stores the text and metadata of the comments
article_comments_dterms	Stores the distinct terms per article and its respective comments and the respective term frequencies for the article
comments_dterms	Stores term frequencies for the comments
article_cosine_vector	Stores the normalized term frequency vectors of the articles
comments_cosine_vector	Stores all diversification criteria in the form of feature vectors per comment
user_com_quality	Stores the user aggregated feature vector for the Aggregate Comment Writing Quality criterion

Πίνακας 6.1: Σχήμα βάσης



Σχήμα 6.5: Γραφικό περιβάλλον εφαρμογής

με τις επιλογές του χρήστη. Στην περιοχή «Comments» παρουσιάζεται το σύνολο των σχολίων, ταξινομημένα ανά ημερομηνία. Στην περιοχή «Diverse Comments», παρουσιάζεται ένα υποσύνολο διαφοροποιημένων σχολίων, διαφορετικό, ανάλογα με την επιλογή του αλγορίθμου και των κριτηρίων που έγινε από το χρήστη.

6.4 Πειραματική αξιολόγηση

6.4.1 Συγκρινόμενες μέθοδοι

Σε αυτήν την ενότητα παρουσιάζουμε μία αναλυτική αξιολόγηση των αλγορίθμων και των κριτηρίων που περιγράφηκαν προηγουμένως. Ως βασική μέθοδο προς σύγκριση με τις μεθόδους

μας θεωρούμε την αφελή (αν και καθιερωμένη στο σενάριο της διαφοροποίησης αποτελεσμάτων αναζήτησης) μέθοδο *Content Diversity - CONTENTDIV*, η οποία διαφοροποιεί τα σχόλια χρηστών βασιζόμενη μόνο στο κριτήριο της κειμενικής ομοιότητας. Οι υπόλοιπες μέθοδοι είναι κάποιες αντιπροσωπευτικές παραλλαγές της μεθόδου μας, όσον αφορά στα κριτήρια διαφοροποίησης που συνδυάζουν. Οι συγκρινόμενες μέθοδοι περιγράφονται παρακάτω:

- **Κειμενική διαφοροποίηση - CONTENTDIV**. Η βασική μέθοδος σύγκρισης που εφαρμόζει διαφοροποίηση μόνο πάνω στο κειμενικό περιεχόμενο.
- **Διαφοροποίηση συναισθήματος - SENTIDIV**. Η παραλλαγή που διαφοροποιεί μόνο πάνω στο αναγνωριζόμενο συναίσθημα των σχολίων.
- **Διαφοροποίηση ονοματικών οντοτήτων - NEDIV**. Η παραλλαγή που διαφοροποιεί μόνο πάνω στις αναγνωριζόμενες ονοματικές οντότητες.
- **Διαφοροποίηση συναισθήματος ονοματικών οντοτήτων - NESENTIDIV**. Η παραλλαγή που διαφοροποιεί πάνω στο συναίσθημα που αναγνωρίζεται *γύρω* από τις αναγνωριζόμενες ονοματικές οντότητες του σχολίου.
- **Υβριδική διαφοροποίηση - SEMIHYBRID**. Η παραλλαγή που διαφοροποιεί συνδυάζοντας τα κριτήρια της κειμενικής ομοιότητας, του συναισθήματος και των ονοματικών οντοτήτων, όπως παρουσιάστηκε στην αρχική μας δουλεια, στο [106].
- **Εκτεταμένη υβριδική διαφοροποίηση - HYBRID**. Η παραλλαγή που διαφοροποιεί συνδυάζοντας όλα προτεινόμενα κριτήρια.

Κάθε μία από τις παραπάνω μεθόδους συνδυάστηκε με καθέναν από τους τέσσερις αλγορίθμους διαφοροποίησης που παρουσιάστηκαν στην Ενότητα 6.2.3: *MAXSUM1, MAXSUM2, MAXMIN* και *MONO-OBJECTIVE*. Σημειώνουμε εδώ ότι, στις μεθόδους που συνδύαζαν κριτήρια, τα σκορ των επιμέρους κριτηρίων σταθμίζονταν ισοδύναμα ώστε να παραχθεί το τελικό σκορ διαφοροποίησης.

Επιπλέον, για όλες τις μεθόδους θέτουμε ένα σταθερό βάρος $w = 0.7$ για το σκορ διαφοροποίησης και αντίστοιχα βάρος $(1 - w) = 0.3$ για το σκορ ομοιότητας με το άρθρο. Με αυτόν τον τρόπο θέλουμε να εξασφαλίσουμε μία επαρκή κειμενική ομοιότητα των διαφοροποιημένων σχολίων με το άρθρο, η οποία θα βοηθούσε να εξαλειφθούν ακραία/άσχετα σχόλια. Παρόλα αυτά, θέτουμε τη βαρύτητα της διαφοροποίησης υπερδιπλάσια από τη βαρύτητα της ομοιότητας με το άρθρο, έτσι ώστε το σκορ ομοιότητας να μην επηρεάζει καταλυτικά την επιλογή ετερογενών σχολίων.

6.4.2 Σύνολο δεδομένων αξιολόγησης

Για την αξιολόγηση παραγάγαμε ένα σύνολο ειδησεογραφικών άρθρων και των αντιστοίχων σχολίων χρηστών από την εφημερίδα New York Times. Η διαδικτυακή έκδοση της εφημερίδας

προσφέρει ένα καλά οργανωμένο API για την ανάκτηση άρθρων[4] και σχολίων (The Community API[5]). Κάθε άρθρο/σχόλιο συνοδεύεται από τα μεταδεδομένα του, όπως ημερομηνία, θεματική κατηγοριοποίηση, χρήστης, κτλ.

Για να συγκεντρώσουμε μία επαρκή ποσότητα άρθρων, κατεβάσαμε άρθρα μέσω του API, χρησιμοποιώντας τη λέξη-κλειδί «financial». Αυτή η διαδικασία μας επέστρεψε 2800 άρθρα, για τα οποία εξετάσαμε αν υπάρχουν σχόλια. Τελικά ανασύραμε 1935 άρθρα με ένα σύνολο 293303 σχολίων, το οποίο δίνει ένα μέσο λόγο 152 σχολίων ανά άρθρο. Σημειώνουμε ότι, αφού (α) η λέξη-κλειδί που χρησιμοποιήσαμε είναι αρκετά γενική και (β) αναζητείται και στο κείμενο των άρθρων, το επεστραμμένο σύνολο άρθρων δεν περιορίζεται μόνο σε οικονομικά άρθρα, όντας αρκετά γενικό ώστε να περιέχει και άλλες θεματικές περιοχές, όπως πολιτική, επιχειρήσεις, κτλ.

6.4.3 Μεθοδολογία αξιολόγησης και μετρικές

Η αξιολόγηση πραγματοποιήθηκε ως εξής: Επιλέξαμε τυχαία 10 άρθρα και το σύνολο των σχολίων για το καθένα. Για κάθε μία από τις έξι παραλλαγές που παρουσιάστηκαν στην Ενότητα 6.4.1, επιστρέφουμε ένα σύνολο από 10 διαφοροποιημένα σχόλια. Κάθε μία από τις παραλλαγές εφαρμόζεται σε κάθε έναν από τους τέσσερις αλγορίθμους, κάτι που καταλήγει σε 24 συνδυασμούς προς αξιολόγηση.

Η αξιολόγηση στο σενάριό μας βασίζεται στην έννοια των μονάδων πληροφορίας, η οποία παρουσιάζεται στο [95]. Σε εκείνο το σενάριο (διαφοροποίηση αποτελεσμάτων αναζήτησης), οι μονάδες μπορεί να είναι διαφορετικές απαντήσεις ενός ερωτήματος - ερώτησης, ή διαφορετικές εκφάνσεις θεμάτων/εννοιών που αντιστοιχούν στους όρους του ερωτήματος. Στο δικό μας σενάριο, της διαφοροποίησης σχολίων, προσαρμόζουμε τον ορισμό των μονάδων πληροφορίας:

Ορισμός 6.3 (Μονάδα πληροφορίας). *Μονάδα πληροφορίας είναι κάθε έννοια ή θεματική κατηγορία ή υποκατηγορία που σχετίζεται με το κυρίως άρθρο ή κάθε σχετική άποψη/συναίσθημα ή επέκταση των παραπάνω εννοιών και κατηγοριών, τα οποία μπορεί να εντοπιστούν, είτε στο κυρίως άρθρο, είτε στα σχόλια του.*

Φυσικά, ο παραπάνω ορισμός δεν είναι αρκετά αυστηρός και, για αυτό, η εφαρμογή του στα ίδια άρθρα/σχόλια μπορεί να δώσει διαφορετικά αποτελέσματα, ανάλογα με την αυστηρότητα και την προσωπική αντίληψη του κάθε ανθρώπου επισημειωτή. Παρόλα αυτά, είναι ένας βολικός ορισμός για το σενάριο αξιολόγησής μας, αφού επιτρέπει την αντιστοίχιση της ετερογένειας σε θέματα και έννοιες που εντοπίζονται στις κειμενικές περιγραφές άρθρων και σχολίων και διευκολύνει των ορισμό κατάλληλων μετρικών διαφοροποίησης:

Μετρική 1 (Nugget Coverage - NC@n). Με την πρώτη μετρική προσπαθούμε να ποσοτικοποιήσουμε την έκταση της κάλυψης των μονάδων πληροφορίας από τα σχόλια που περιέχονται σε ένα σύνολο-αποτέλεσμα, για κάθε δοκιμαζόμενη μέθοδο. Ουσιαστικά, είναι μία μετρική βασισμένη στην Precision at N μετρική και μετράει πόσες μονάδες πληροφορίας, συνολικά, εντοπίζονται στα σχόλια του διαφοροποιημένου συνόλου. Ορίζεται ως εξής:

[4] http://developer.nytimes.com/docs/read/article_search_api

[5] http://developer.nytimes.com/docs/community_api

$$NC@n = \frac{\sum_{k=1}^{n} I_k}{n \cdot |I|}$$

όπου n είναι ο αριθμός των σχολίων του συνόλου, I_k ο αριθμός των διακριτών μονάδων πληροφορίας που περιέχονται στο σχόλιο k και $|I|$ ο συνολικός αριθμός διακριτών μονάδων πληροφορίας. Αφού ο μέγιστος αριθμός διακριτών μονάδων πληροφορίας σε ένα σχόλιο είναι $|I|$, οπότε $|I_k| \leq |I|$, η μετρική είναι κανονικοποιημένη στο διάστημα $[0, 1]$.

Μετρική 2: (Distinct Nugget Coverage). Αυτή η μετρική είναι συμπληρωματική με την πρώτη και μετράει το λόγο των **διακριτών** μονάδων πληροφορίας που βρίσκονται στα σχόλια ενός συνόλου, προς το συνολικό αριθμό διακριτών μονάδων:

$$DN@n = \frac{\sum_{i=1}^{|I|} DFI_i}{|I|}$$

όπου το DFI_i ορίζεται ως εξής:

$$DFI_i = \begin{cases} 1 & : FI_i > 0 \\ 0 & : FI_i = 0 \end{cases}$$

όπου FI_i είναι η συχνότητα της μονάδας i στο σύνολο των σχολίων του διαφοροποιημένου συνόλου n σχολίων.

Μετρική 3: (nugget Uniformity NU@n). Με αυτήν την μετρική μετράμε την ετερογένεια των μονάδων μέσα στο σύνολο σχολίων, απαιτώντας οι μονάδες να είναι όσο το δυνατόν πιο ομοιόμορφα κατανεμημένες. Την ορίζουμε ως τη *διακύμανση* των συχνοτήτων των μονάδων στο διαφοροποιημένο σύνολο σχολίων, ως προς την μέση τιμή των συχνοτήτων. Ορίζοντας τη μέση τιμή των συχνοτήτων των μονάδων ως:

$$\bar{I} = \frac{\sum_{i=1}^{|I|} FI_i}{|I|}$$

τότε η μετρική ορίζεται ως:

$$NU@n = \frac{\sum_{i=1}^{|I|} (FI_i - \bar{I})^2}{|I|}$$

Για κάθε άρθρο, αφού και οι 24 συνδυασμοί μεθόδων-αλγορίθμων τρέξουν, κρατάμε το διακριτό σύνολο σχολίων που προκύπτει συνδυάζοντας τα 10 πρώτα σχόλια από κάθε μέθοδο. Κρατάμε μόνο τα αναγνωριστικά των σχολίων και αφαιρούμε κάθε πληροφορία προέλευσης του σχολίου, δηλαδή από ποιες από τις 24 μεθόδους προέκυψε. Στη συνέχεια, εκτελούμε τα ακόλουθα δύο βήματα:

α) Εξαγωγή μονάδων πληροφορίας. Αφού διαβάσουμε κάθε άρθρο και τα αντίστοιχα σχόλιά του, παράγουμε χειροκίνητα ένα σύνολο με όλες τις διακριτές μονάδες πληροφορίας που σχετίζονται με τα θέματα του άρθρου. Με αυτόν τον τρόπο δημιουργούμε μία δεξαμενή από μονάδες πληροφορίας. Σημειώνουμε ότι επιλέξαμε να εξάγουμε μονάδες πληροφορίας, εκτός από το άρθρο, και από τα σχόλιά του, αφού, όπως δηλώνουμε στην Εισαγωγή, θεωρούμε τα σχόλια πολύτιμα μεταδεδομένα του άρθρου, που το συμπληρώνουν και το εμπλουτίζουν.

Article's main topic	**Indicative nuggets**
Tax evasion	Tax evasion, Ethics, Law and Legislation, Politics
Obama's anti-foreclosure plan	Housing bubble, Politics, People's irresponsibility, Mortgage crisis
Scandal with politician and bankers	Bailout, Bank's name, Legislation, Corruption, Discrimination
Financial reform related to elections	Goldman Sachs, Subprime mortgage crisis, Economic ideologies, Wall Street
Federal loans on energy programs	Alternative energy, Politics, Competitiveness, Financial crisis, Political parties
US consumption rate	Consumption comparisons, Free-market economy, Solutions, Self criticism
Democrats nominations	Clintons unite Republicans, Obama represents change, Critisism on candidates
Prescriptions decrease: consequences/reasons	Critisism on corporations, Economical drug solutions, Patients examples
Obama measures on financial crisis	Critisize irresponsible americans, Blame free market, Measures are moderate
Relation between successful people/elite colleges	Community college inferior to others, How students exploit education matters

Πίνακας 6.2: Αξιολογημένα άρθρα και ενδεικτικές μονάδες πληροφορίας

Σημειώνουμε, επίσης, ότι αυτή η διαδικασία περιελάμβανε περισσότερες από μία επαναλήψεις ανά ομάδα άρθρου-σχολίων: ορισμένες φορές, αφού ανακαλύπταμε μία νέα μονάδα πληροφορίας, επιστρέφαμε πίσω και επανεξετάζαμε το άρθρο και τα προηγούμενα σχόλια, έτσι ώστε να βεβαιωθούμε ότι σε κάθε σχόλιο ανατίθεντο όλες οι σχετικές με αυτό μονάδες.

Για την παραπάνω διαδικασία δοκιμάστηκαν, επιπλέον, και αυτόματες μέθοδοι εξαγωγής μονάδων πληροφορίας. Συγκεκριμένα, χρησιμοποιήσαμε το OpenCalais Web Service[6], ένα εργαλείο που εξάγει αυτόματα σημασιολογικά μεταδεδομένα από κείμενο: ονοματικές οντότητες, γεγονότα και συμβάντα. Επίσης, χρησιμοποιήσαμε το AlchemyAPI[7], το οποίο επίσης αναλύει κείμενο και επιστρέφει αφηρημένες έννοιες. Τα αποτελέσματα, όμως, που προέκυψαν ήταν πολύ χαμηλής ποιότητας σε σχέση με τις μονάδες πληροφορίας που αναγνωρίστηκαν χειροκίνητα, οπότε απορρίψαμε την συγκεκριμένη, αυτόματη, εκδοχή αξιολόγησης.

β) Επισημείωση σχολίων. Στη συνέχεια, αναθέτουμε σε δύο εξωτερικούς κριτές/επισημειωτές την επισημείωση των σχολίων με μονάδες πληροφορίας: Παρουσιάζουμε στους επισημειωτές: (α) το κείμενο του άρθρου, (β) το διακριτό σύνολο σχολίων που προκύπτουν από τα πρώτα 10 επεστραμμένα σχόλια, από όλες τις μεθόδους, ξανά, απαλλαγμένα από πληροφορία προέλευσης, ώστε να μην επηρεαστούν οι κριτές, για παράδειγμα, παρατηρώντας κάποιο μοτίβο προέλευσης σχολίων από συγκεκριμένη μέθοδο και (γ) το σύνολο των μονάδων πληροφορίας. Ζητείται από τους κριτές να επισημειώσουν κάθε σχόλιο με τις μονάδες πληροφορίας που πιστεύουν ότι περιέχει. Αφού γίνει αυτή η διαδικασία για όλα τα σχόλια των αξιολογούμενων άρθρων, εφαρμόζουμε τις παραπάνω μετρικές αξιολόγησης. Σημειώνουμε ότι δεν επιβαρύνουμε τους επισημειωτές με τη διαδικασία της αναγνώρισης μονάδων πληροφορίας: μπορούν να επιλέξουν μόνο μονάδες από τη δεξαμενή του βήματος (α).

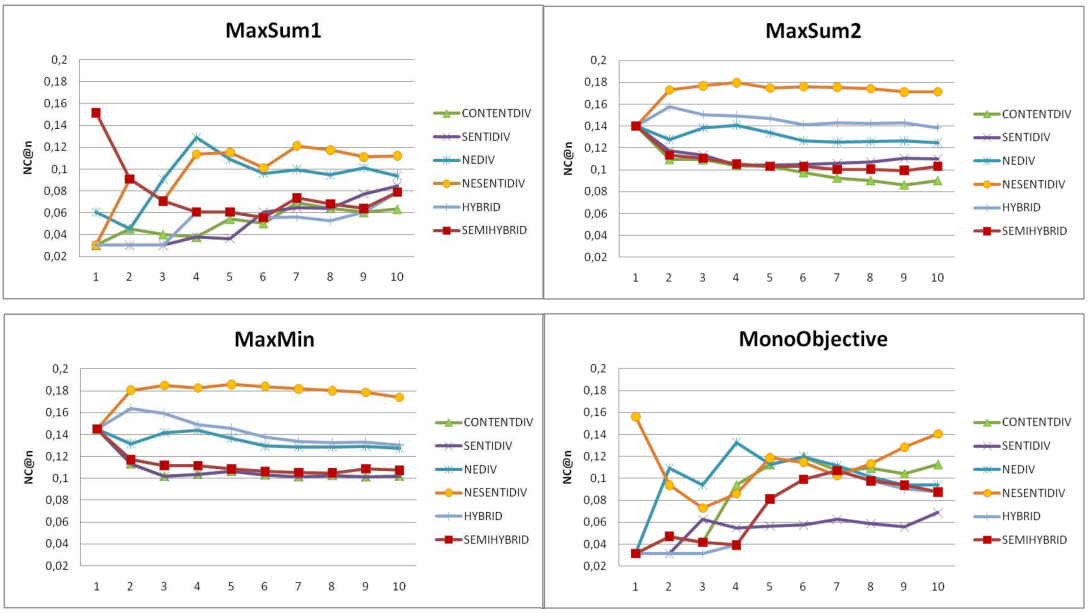
Στον Πίνακα 6.2 παρουσιάζουμε τα δέκα αξιολογηθέντα άρθρα, μαζί με κάποιες ενδεικτικές μονάδες πληροφορίας για το καθένα.

[6]www.opencalais.com

[7]http://www.alchemyapi.com/api/concept/

6.4.4 Αποτελέσματα αξιολόγησης

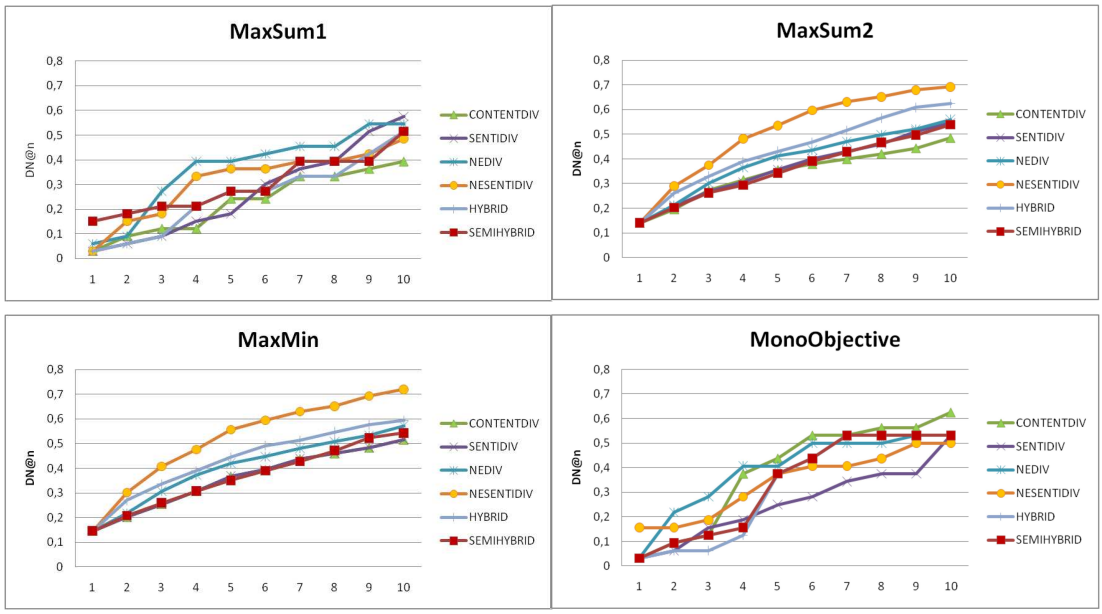
Στις Εικόνες 6.6 και 6.7 παρουσιάζουμε γραφικές παραστάσεις για Nugget Coverage και Distinct Nugget Coverage, για τις θέσεις κατάταξης σχολίων από 1 έως 10. Υπενθυμίζουμε ότι και οι δύο μετρικές είναι κανονικοποιημένες στο διάστημα $[0, 1]$. Οι γραφικές παρουσιάζονται ανά αλγόριθμο, για ευκολία παρουσίασης. Για την Nugget Coverage, η πρώτη παρατήρηση είναι ότι οι αλγόριθμοι *MAXSUM2* και *MAXMIN* είναι διακριτά πιο αποτελεσματικοί από τους άλλους δύο και, επιπλέον, ακολουθούν μία πιο συνεπή/ομαλή συμπεριφορά, όσον αφορά στη θέση κατάταξης, για όλες τις δοκιμαζόμενες παραλλαγές. Η δεύτερη παρατήρηση είναι ότι η βασική μέθοδος σύγκρισης (*CONTENTDIV*) αποδίδει σχεδόν πάντα χειρότερα από τις προτεινόμενες παραλλαγές, ακόμα και σε επίπεδο αλγορίθμου ξεχωριστά. Τέλος, η παραλλαγή που συνδυάζει ονοματικές οντότητες και συναίσθημα (*NESENTIDIV*), είναι διακριτά αποτελεσματικότερη από όλες τις άλλες παραλλαγές και τη βασική μέθοδο, και ακολουθείται από την παραλλαγή που συνδυάζει όλα τα κριτήρια (*HYBRID*) και την παραλλαγή των ονοματικών οντοτήτων (*NEDIV*).



Σχήμα 6.6: Nugget Coverage ανά αλγόριθμο

Για τη μετρική Distinct Nugget Coverage ισχύουν σχεδόν οι ίδιες παρατηρήσεις με την εξαίρεση της μεθόδου *CONTENTDIV* η οποία είναι αναπάντεχα αποτελεσματική στον αλγόριθμο *MONO-OBJECTIVE*, αλλά, και πάλι, χειρότερη από την παραλλαγή *NESENTIDIV* στους αλγορίθμους *MAXSUM2* και *MAXMIN*.

Για να τονίσουμε καλύτερα τις διαφορές μεταξύ των μεθόδων, θεωρούμε, στους Πίνακες 6.3, 6.4, 6.5 και 6.6 τις τιμές Nugget Coverage και Distinct Nugget Coverage μόνο για τις 5 και 10 πρώτες θέσεις κατάταξης, ξεχωριστά. Επίσης, σημειώνουμε τον καλύτερο αλγόριθμο ανά παραλλαγή κριτηρίων (τελευταία γραμμή), την καλύτερη παραλλαγή κριτηρίων ανά αλγόριθμο, (τελευταία στήλη) και τον συνολικά καλύτερο συνδυασμό παραλλαγής/αλγορίθμου.



Σχήμα 6.7: Distinct Nugget Coverage ανά αλγόριθμο

Ο συνδυασμός *MAXMIN/NESENTIDIV* ξεπερνάει όλους τους άλλους συνδυασμούς σε αποτελεσματικότητα. Επιπλέον, βελτιώνει την αποτελεσματικότητα της βασικής μεθόδου κατά 65%, 54%, 27% και 15% για τα NC@5, NC@10, DN@5 και DN@10 αντίστοιχα. Οι επί τοις εκατό διαφορές μεταξύ του *MAXMIN/NESENTIDIV* και της βασικής μεθόδου είναι αντίστοιχα: 7.3%, 6.1%, 11.8% ανδ 9.5%.

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.055	0.036	0.109*	0.115*	0.061*	0.061	NESENTIDIV
MAXSUM2	0.103	0.104	0.134	0.175*	0.147*	0.103	NESENTIDIV
MAXMIN	0.106	0.123	0.137	**0.186***	0.146	0.108	NESENTIDIV
MONO	0.113	0.056	0.113	0.119*	0.081	0.081	NESENTIDIV
Best Algorithm per criterion	MONO	MAXMIN	MAXMIN	MAXMIN	MAXSUM2	MAXMIN	MAXMIN/ NESENTIDIV

Πίνακας 6.3: Nugget Coverage στη θέση 5

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.064	0.085	0.094	0.112*	0.079*	0.079	NESENTIDIV
MAXSUM2	0.090	0.110	0.125	0.17*1	0.139*	0.103	NESENTIDIV
MAXMIN	0.102	0.115	0.128	**0.174***	0.131	0.107	NESENTIDIV
MONO	0.113	0.069	0.094	0.141*	0.088	0.088	NESENTIDIV
Best Algorithm per criterion	MONO	MAXMIN	MAXMIN	MAXMIN	MAXSUM2	MAXMIN	MAXMIN/ NESENTIDIV

Πίνακας 6.4: Nugget Coverage at position στη θέση 10

Επίσης, αναφέρουμε τιμές στατιστικής σημασίας των διαφορών μεταξύ των παραλλαγών και της βασικής μεθόδου, χρησιμοποιώντας το T-Test[8] με διάστημα εμπιστοσύνης 95%. Σε κάθε γραμμή των Πινάκων 6.3, 6.4, 6.5 και 6.6, σημειώνονται με αστερίσκο οι παραλλα-

[8]http://www.socialresearchmethods.net/kb/stat_t.php

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.242	0.182	0.394	0.364*	0.273*	0.273	NEDIV
MAXSUM2	0.355	0.357	0.411	0.536*	0.431	0.342	NESENTIDIV
MAXMIN	0.368	0.399	0.421	**0.556***	0.445	0.351	NESENTIDIV
MONO	0.438	0.250	0.406	0.375	0.375	0.375	CONTENTDIV
Best Algorithm per criterion	MONO	MAXMIN	MAXMIN	MAXMIN	MAXMIN	MONO	MAXMIN/ NESENTIDIV

Πίνακας 6.5: Distinct Nugget Coverage στη θέση 5

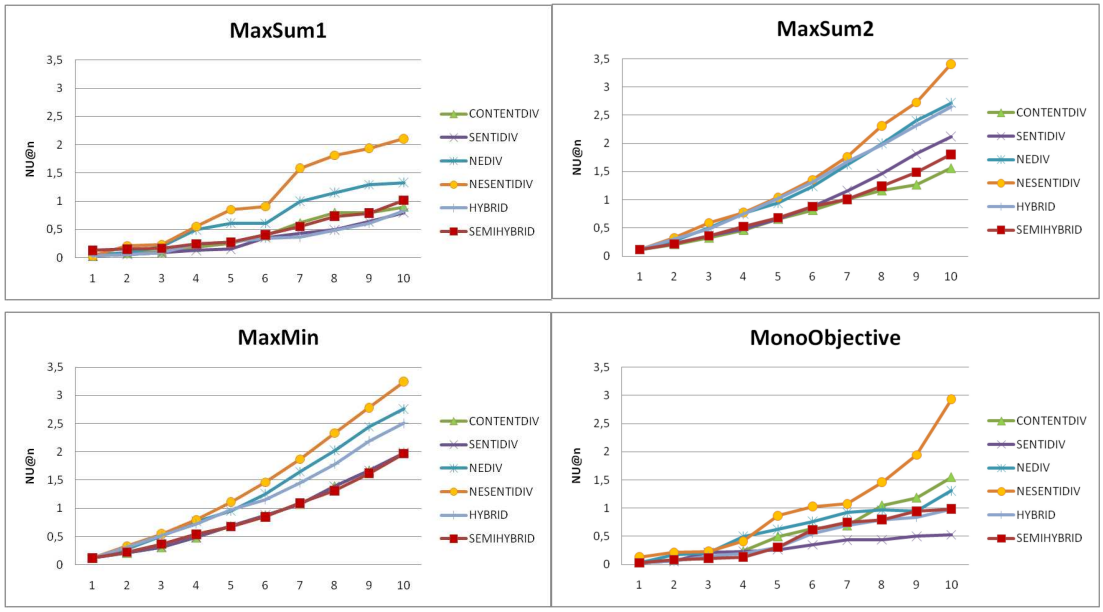
Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.394	0.576	0.546	0.485*	0.515*	0.515	SENTIDIV
MAXSUM2	0.485	0.545	0.560	0.692*	0.625*	0.538	NESENTIDIV
MAXMIN	0.516	0.557	0.572	**0.720***	0.594	0.543	NESENTIDIV
MONO	0.625	0.531	0.531	0.500*	0.531	0.531	CONTENTDIV
Best Algorithm per criterion	MAXMIN	MAXMIN	MAXMIN	MAXMIN	MAXSUM2	MAXMIN	MAXMIN/ NESENTIDIV

Πίνακας 6.6: Distinct Nugget Coverage στη θέση 10

γές εκείνες των οποίων η διαφορά στην αποτελεσματικότητα από τη βασική μέθοδο *CONTENTDIV* είναι στατιστικά σημαντική. Σημειώνουμε ότι ο καλύτερος συνδυασμός παραλλαγής/αλγορίθμου, *MAXMIN/NESENTIDIV* είναι στατιστικά σημαντικά καλύτερος από οποιονδήποτε συνδυασμό βασικής μεθόδου/αλγορίθμου.

Εξετάζοντας τις παραπάνω γραφικές και πίνακες είναι αρκετά προφανές ότι οι ονοματικές οντότητες είναι ένα σημαντικό κριτήριο διαφοροποίησης σχολίων χρηστών. Η αποτελεσματικότητα του κριτηρίου ενισχύεται ακόμα περισσότερο, όταν συνδυάζεται με αναγνώριση συναισθήματος γύρω από τις ονοματικές οντότητες. Αυτό δικαιολογείται πιθανότατα από το γεγονός ότι πολλά θέματα που περιγράφονται σε ειδησεογραφικά άρθρα είναι αρκετά πιθανόν να σχετίζονται με πρόσωπα ή οργανισμούς, οπότε οι ονοματικές οντότητες βοηθούν στην αποτελεσματικότερη αιχμαλώτιση αυτών των θεμάτων. Επιπλέον, η ετερογένεια στο συναίσθημα γύρω από αυτές τις ονοματικές οντότητες πιθανόν συνεπάγεται και ετερογένεια των αντιστοίχων θεμάτων. Δεύτερον, καταδεικνύεται ότι πιο εκλεπτυσμένα κριτήρια από την απλή κειμενική ομοιότητα είναι πιο αποτελεσματικά. Τέλος, ο αλγόριθμος *MAXMIN* είναι αποτελεσματικότερος όλων των άλλων, με ελαφρά χειρότερο τον *MAXSUM2*. Και οι δύο αποδίδουν διακριτά καλύτερα από τους *MAXSUM1* και *MONO-OBJECTIVE*. Η κύρια διαφορά μεταξύ των δύο καλύτερων και των δύο χειρότερων αλγορίθμων είναι ότι οι πρώτοι υπολογίζουν, σε κάθε βήμα, αποστάσεις μεταξύ υποψήφιων και ήδη επιλεγμένων σχολίων, ενώ οι δεύτεροι, αποστάσεις **μόνο** μεταξύ υποψηφίων σχολίων.

Η Εικόνα 6.8 και οι Πίνακες 6.7,6.8 παρουσιάζουν τα αποτελέσματα για την τρίτη μετρική, την Nugget Uniformity, η οποία ποσοτικοποιεί τις διαφορές στις συχνότητες των μονάδων πληροφορίας σε κάθε διαφοροποιημένο σύνολο-αποτέλεσμα. Υπενθυμίζουμε ότι η συγκεκριμένη μετρική δεν είναι κανονικοποιημένη και χαμηλότερες τιμές υποδηλώνουν μεγαλύτερη αποτελεσματικότητα. Σε αυτήν την μετρική, αν και η συνολικά καλύτερη απόδοση για τις μετρικές NU@5, NU@10 επιτυγχάνεται από την παραλλαγή *SENTIDIV*, δεν υπάρχει κάποια παραλλαγή η οποία ξεκάθαρα ξεπερνάει όλες τις υπόλοιπες, σε όλες τις περιπτώσεις. Επιπλέον, είναι φανερό ότι οι μετρικές που αποδίδουν καλά στις δύο πρώτες μετρικές (που αφορούν



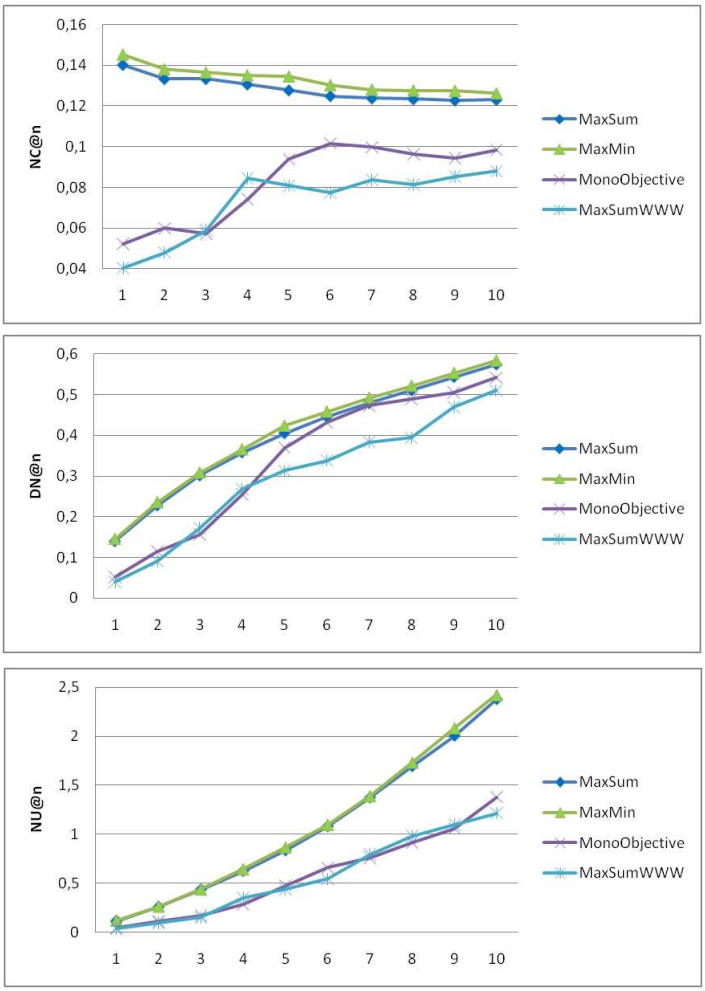
Σχήμα 6.8: Nugget Uniformity ανά αλγόριθμο

στην κάλυψη μονάδων πληροφορίας), αποδίδουν σχετικά άσχημα σε αυτήν την μετρική, και το αντίστροφο. Αυτό μπορεί να δικαιολογηθεί από το γεγονός ότι, όσο αυξάνεται ο αριθμός των εντοπιζόμενων μονάδων πληροφορίας σε ένα διαφοροποιημένο σύνολο σχολίων, αναμένεται πιο δημοφιλείς μονάδες (δηλαδή μονάδες που αντιστοιχούν σε πιο δημοφιλείς έννοιες, θέματα και απόψεις) να συνεισφέρουν περισσότερο στην αύξηση, ενώ για μικρό αριθμό συνολικών μονάδων, οι διαφορές στις επιμέρους συχνότητες των διαφορετικών μονάδων αναμένονται επίσης μικρές. Επίσης, αναμένεται να υπάρχουν κάποιες ακραίες μονάδες πληροφορίας, δηλαδή μονάδες που εμφανίζονται σε ελάχιστα σχόλια, αντιπροσωπεύοντας λιγότερο δημοφιλείς εκφάνσεις του άρθρου. Οπότε, όταν ο συνολικός αριθμός μονάδων αυξάνεται, οι συγκεκριμένες μονάδες αναμένεται να παραμένουν λίγες, επηρεάζοντας αρνητικά τις τιμές της Nugget Uniformity. Φυσικά, η Nugget Coverage και ειδικά η Distinct Nugget Coverage είναι πιο σημαντικές μετρικές για το πρόβλημα της διαφοροποίησης, οπότε η ιδανική μέθοδος διαφοροποίησης θα έπρεπε να επιλέγεται κυρίως με βάση αυτές. Από την άλλη πλευρά, υπάρχουν παραλλαγές που συνιστούν μία μέση λύση, όπως οι παραλλαγές *NEDIV*, *HYBRID* με τους αλγορίθμους *MAXMIN* και *MAXSUM2*.

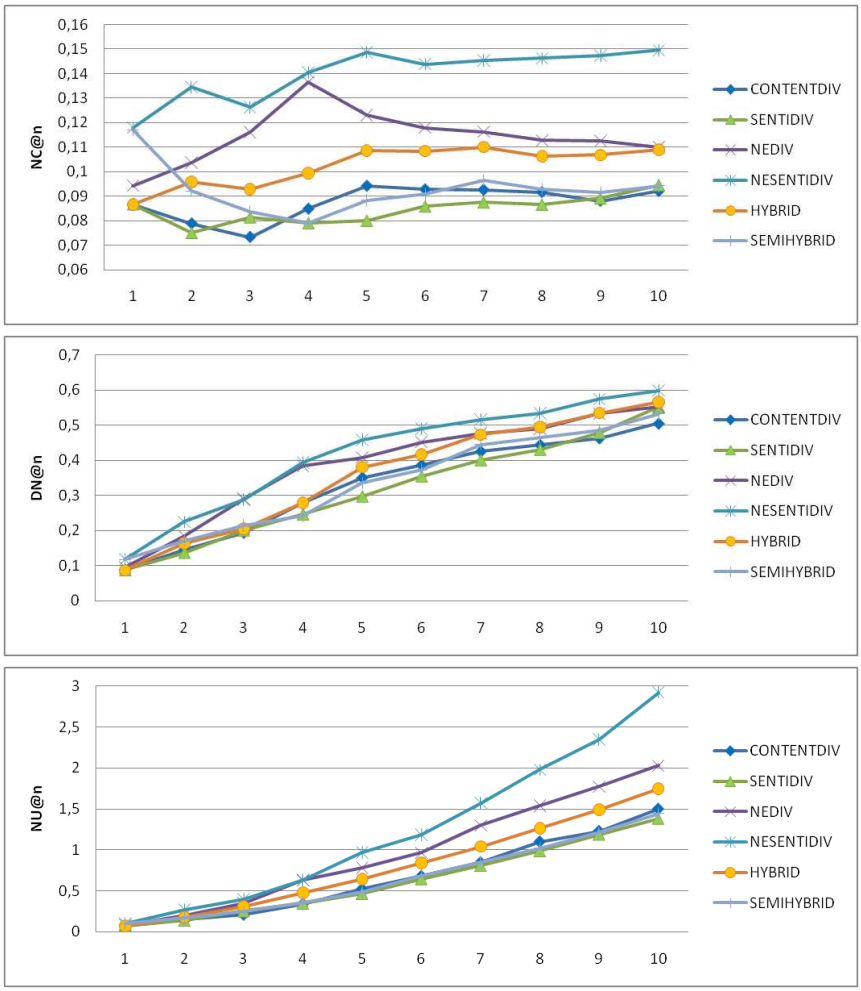
Σημειώνουμε ότι, αντιθέτως με τις δύο πρώτες μετρικές (που προσιδιάζουν μετρικές ακρίβειας), δε θεωρούμε χρήσιμο να πραγματοποιήσουμε τεστ στατιστικής σημαντικότητας στη Nugget Uniformity, αφού, αφενός μεν η μετρική δεν είναι κανονικοποιημένη, αφετέρου δε τα αποτελέσματα δεν θα είχαν κάποια διαισθητική σημασία.

Στην Εικόνα 6.9, παρουσιάζεται η μέση τιμή της αποτελεσματικότητας κάθε αλγορίθμου πάνω σε όλες της παραλλαγές. Η γραφικές επιβεβαιώνουν τις προηγούμενες παρατηρήσεις όσον αφορά στην υπεροχή των αλγορίθμων *MAXMIN* και *MAXSUM2* στις δύο μετρικές κάλυψης μονάδων και την υστέρησή τους, αντίστοιχα, στην μετρική ομοιομορφίας.

Ομοίως, στην Εικόνα 6.10, παρουσιάζεται η μέση αποτελεσματικότητα κάθε παραλλαγής,



Σχήμα 6.9: Μέση αποτελεσματικότητα κάθε αλγορίθμου διαφοροποίησης πάνω σε όλες τις παραλλαγές κριτηρίων



Σχήμα 6.10: Μέση αποτελεσματικότητα κάθε παραλλαγής κριτηρίων πάνω σε όλους τους αλγόριθμους διαφοροποίησης

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.259	**0.149**	0.612	0.850	0.272	0.272	SENTIDIV
MAXSUM2	0.659	0.662	0.940	1.041	1.018	0.678	CONTENTDIV
MAXMIN	0.688	0.793	0.954	1.112	0.974	0.677	SEMIHYBRID
MONO	0.496	0.265	0.621	0.866	0.304	0.304	SENTIDIV
Best Algorithm per criterion	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1/ SENTIDIV

Πίνακας 6.7: Nugget Uniformity στη θέση 5

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.898	0.795	1.330	2.107	0.834	1.016	SENTIDIV
MAXSUM2	1.565	2.122	2.718	3.407	2.649	1.805	CONTENTDIV
MAXMIN	1.983	2.079	2.760	3.245	2.509	1.969	SEMIHYBRID
MONO	1.547	**0.5273**	1.309	2.929	0.984	0.984	SENTIDIV
Best Algorithm per criterion	MAXSUM1	MONO	MONO	MAXSUM1	MAXSUM1	MONO	MONO/ SENTIDIV

Πίνακας 6.8: Nugget Uniformity στη θέση 10

πάνω σε όλους τους αλγορίθμους. Και εδώ επιβεβαιώνονται προηγούμενες παρατηρήσεις όσον αφορά την στατιστικά σημαντική υπεροχή της παραλλαγής *NESENTIDIV* στις μετρικές κάλυψης μονάδων και τις βολικότητας των παραλλαγών *NEDIV* και *HYBRID*, οι οποίες έχουν μία μέση απόδοση και στις τρεις μετρικές.

6.5 Συμπεράσματα

Σε αυτήν την ενότητα, παρουσιάσαμε μία πρωτότυπη προσέγγιση για διαφοροποίηση σχολίων χρηστών σε άρθρα και, γενικότερα, σε περιεχόμενο κοινωνικών δικτύων. Εισάγαγαμε κριτήρια διαφοροποίησης, εξειδικευμένα για σχόλια χρηστών και τα εφαρμόσαμε σε συνδυασμό με τέσσερις ευριστικούς αλγόριθμους (τρεις καθιερωμένους και μία δική μας παραλλαγή), ορίζοντας τις κατάλληλες συνθήκες αρχικοποίησης και συναρτήσεις ομοιότητας και συνάθροισης αποστάσεων. Υλοποιήσαμε τα παραπάνω σε μία πρωτότυπη, επιτραπέζια εφαρμογή η οποία τρέχει σε δεδομένα ενός ανοιχτού συνόλου δεδομένων. Η πειραματική αξιολόγηση, η οποία βασίζεται σε μετρικές αξιολόγησης διαφοροποίησης που ορίσαμε για το συγκεκριμένο πρόβλημα, κατέδειξε την αποτελεσματικότητα των μεθόδων μας, απέναντι στην τετριμμένη λύση της διαφοροποίησης μόνο με βάση το κειμενικό περιεχόμενο. Το πλαίσιο που αναπτύξαμε είναι αρκετά γενικό για να εφαρμοστεί σε διάφορα σενάρια διαφοροποίησης, όπως διαφοροποίηση συζητήσεων σε φόρουμ, διαφοροποίηση tweets, διαφοροποίηση αναρτήσεων σε μπλογκ, κτλ.

Η μελλοντική δουλειά έγκειται στη βελτίωση του κριτηρίου «Συν-σχολιασμός Χρηστών», εφαρμόζοντας αλγορίθμους θεματικής συσταδοποίησης προκειμένου να μειωθεί η αραιότητα των διανυσμάτων χαρακτηριστικών του κριτηρίου και να επανελεχθεί η αποτελεσματικότητά του. Επιπλέον, θέλουμε να εξετάσουμε αν υβριδικοί συνδυασμοί αλγορίθμων μπορούν να δώσουν καλύτερα αποτελέσματα. Τέλος, σκοπεύουμε να εξετάσουμε την αποτελεσματικότητα του προτεινόμενου πλαισίου σε άλλα σενάρια και να εξατάσουμε πιθανές βελτιώσεις/εξειδικεύσεις του πλαισίου για τα συγκεκριμένα σενάρια.

Κεφάλαιο 7

Λοιπές εργασίες

Σε αυτήν την ενότητα, παρουσιάζουμε εργασίες για τις οποίες πραγματοποιήθηκε μία πρώτη προεργασία κατά τη διάρκεια της διδακτορικής διατριβής, αλλά δεν αποτελούν δομικό κομμάτι της διατριβής. Η πρώτη δουλειά αφορά διαφοροποίηση αποτελεσμάτων αναζήτησης με λέξεις κλειδιά σε σημασιολογικά, δομημένα δεδομένα (RDF) [105] και η δεύτερη αφορά μοντελοποίηση των αλλαγών που πραγματοποιούνται σε ονομασίες συγκεκριμένων βιολογικών οντοτήτων, με απώτερο σκοπό την αποτελεσματική αναζήτησή τους [114, 115]. Και οι δύο δουλειές, παρόλο που βρίσκονται σε πρώιμο στάδιο, πραγματοποιούν μία αρχική μοντελοποίηση των αντιστοίχων προβλημάτων και αποτελούν ένα πρώτο, σημαντικό βήμα για την επίλυσή τους.

7.1 Διαφοροποίηση αναζήτησης σε σημασιολογικά δεδομένα

Σε αυτήν την ενότητα, παρουσιάζουμε μία πρώτη προσέγγιση του προβλήματος της διαφοροποίησης αποτελεσμάτων αναζήτησης σε σημασιολογικά δεδομένα. Το πρόβλημα έγκειται στην ανάκτηση ενός συνόλου από k σημασιολογικά αποτελέσματα αναζήτησης με λέξεις κλειδιά, τα οποία θα παρουσιάζουν τη μεγαλύτερη δυνατή ετερογένεια, τόσο από άποψη περιεχομένου, όσο και από άποψη δομής και συσχετίσεων, μιας και αναφερόμαστε σε δομημένα δεδομένα (RDF) που μπορεί να ακολουθούν συγκεκριμένο σχήμα και που συσχετίζονται μεταξύ τους με τη βοήθεια ιδιοτήτων. Αν και η διαφοροποίηση αποτελεσμάτων αναζήτησης σε αδόμητα δεδομένα (π.χ. ιστοσελίδες) είναι ευθύ πρόβλημα, αφού στόχος είναι η συλλογή αποτελεσμάτων ετερογενών ως προς το περιεχόμενο, δεν ισχύει το ίδιο για δομημένα δεδομένα. Η λύση που προτείνουμε είναι ένα πλαίσιο που θα συνδυάζει διάφορες πτυχές των σημασιολογικών δεδομένων (περιεχόμενο, σχήμα, δομή, συσχετίσεις οντοτήτων) και θα παράγει ετερογενή σύνολα από γράφους - αποτελέσματα, οι οποίοι θα αντιστοιχούν μεν στις λέξεις κλειδιά του αντίστοιχου ερωτήματος, αλλά θα διαφοροποιούν τις οντότητες - κόμβους, τις ιδιότητες - ακμές που τους συνδέουν και τις κλάσεις της οντολογίας που τους χαρακτηρίζουν. Η συγκεκριμένη δουλειά βρίσκεται υπό σχεδιασμό/ανάπτυξη αυτή τη στιγμή, παρόλα αυτά φαίνεται να είναι η πρώτη εργασία που αγγίζει, έστω και σε προκαταρκτικό στάδιο, το πρόβλημα της διαφοροποίησης RDF δεδομένων.

7.1.1 Περιγραφή προβλήματος και κίνητρο

Καθώς ένας μεγάλος αριθμός από εταιρείες, οργανισμούς και ιδρύματα (π.χ., *Europeana*, *DBpedia*, *data.gov*, *GeoNames*) έχουν αρχίσει να υιοθετούν σιγά σιγά το παράδειγμα των διασυνδεδεμένων δεδομένων και να δημοσιεύουν/διαχειρίζονται τα δεδομένα τους σε RDF πρότυπα, η υλοποίηση προχωρημένων, αποδοτικών και αποτελεσματικών τεχνικών αναζήτησης RDF δεδομένων καθίσταται σημαντική πρόκληση. Αυτή η ανάγκη ενισχύεται από τη σαφή προτίμηση των χρηστών να εκτελούν αναζήτηση με λέξεις κλειδιά και να αποφεύγουν πιο δομημένους τρόπους/γλώσσες ερωτήσεων (π.χ. SPARQL), οδηγώντας την ερευνητική κοινότητα προς αυτήν την κατεύθυνση [70, 71, 72].

Οι περισσότερες δουλειές που καταπιάνονται με το παραπάνω πρόβλημα επιστρέφουν τα *πιο σχετικά* RDF αποτελέσματα στη μορφή *γράφων ή δέντρων*. Η σχετικότητα ενός αποτελέσματος με το ερώτημα ορίζεται τυπικά με βάση (α) την *κειμενική ομοιότητα* μεταξύ του κειμένου των οντοτήτων του αποτελέσματος και των όρων του ερωτήματος και (β) τη *συνεκτικότητα αποτελέσματος* το οποίο σημαίνει ότι προτιμούνται γράφοι ή δέντρα όσο το δυνατόν μικρότερου μεγέθους. Το μειονέκτημα είναι ότι αυτή η στρατηγική οδηγεί, πολλές φορές, σε σύνολα από αποτελέσματα με μεγάλη επικάλυψη. Επιπλέον, σημαντική πληροφορία ενδέχεται να χάνεται, αφού δύο προς αναζήτηση οντότητες μπορεί να συνδέονται με ένα μεγάλο, πλην όμως χρήσιμο από πλευράς πληροφορίας μονοπάτι. Τέτοιου είδους αποτελέσματα πιθανόν απορρίπτονται, λόγω της απαίτησης για μικρά μεγέθη αποτελεσμάτων. Τέλος, οι περισσότερες προσεγγίσεις δεν εξετάζουν τη δομή και τη σημασιολογία που παρέχονται από το RDF μοντέλο. Για παράδειγμα, μία προσέγγιση αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα θα έπρεπε να αντιμετωπίζει τις ακμές του RDF, δηλαδή τις ιδιότητες που συνδέουν τις οντότητες μεταξύ τους, ως πολίτες πρώτης κατηγορίας, αφού οι συγκεκριμένες ιδιότητες περιέχουν, έμμεσα, σημασιολογία για τις οντότητες που συνδέουν.

Σχήμα 7.1: Συσχετίσεις μεταξύ των οντοτήτων *Scarlett Johansson*, *Woody Allen*

Ως παράδειγμα, έστω ένας χρήστης που ψάχνει για «*Scarlett Johansson*, *Woody Allen*», πάνω στο σύνολο δεδομένων της DBpedia. Μία αποτελεσματική και εξαντλητική προσέγγιση θα έπρεπε, τουλάχιστον σε αρχικό στάδιο, να εξετάσει όλους τους πιθανούς τρόπους με τους οποίους συνδέονται οντότητες που αντιστοιχούν στις λέξεις κλειδιά του ερωτήματος. Αφού υπάρχουν διάφοροι ρόλοι και συσχετίσεις που αντιστοιχούν στις εξεταζόμενες οντότητες (π.χ. ο Woody Allen μπορεί να είναι είτε σκηνοθέτης είτε ηθοποιός), η παραπάνω προσέγγιση θα οδηγούσε σε ένα υπερβολικά μεγάλο και σύνθετο σύνολο αποτελεσμάτων, το οποίο πιθανότατα θα περιείχε αρκετά επικαλυπτόμενα/παρόμοια αποτελέσματα. Η πληθώρα διαφορετικών συνδυασμών των ενδιάμεσων αποτελεσμάτων απαιτεί ένα μηχανισμό ο οποίος θα μειώνει την επικάλυψη πληροφορίας. Αυτό μπορεί να επιτευχθεί εισάγοντας μία διαδικασία διαφοροποίησης αποτελεσμάτων κατά την ανάκτηση και ταξινόμηση των αποτελεσμάτων. Ιδανικά, το σύστημα θα έπρεπε να επιστρέφει αποτελέσματα που καλύπτουν διαφορετικές όψεις των διασυνδέσεων μεταξύ των προς αναζήτηση οντοτήτων. Στο παράδειγμά μας, στο ερώτημα «*Scarlett Johansson*, *Woody Allen*» θα έπρεπε να επιστρέφονται, σε διαφορετικά αποτελέσματα, πληροφορίες για ταινίες όπου συμπρωταγωνίστησαν, για κοινά βραβεία που έχουν πάρει, για άλλους ηθοποιούς με τους οποίους συνδέονται, κτλ.

Αν και το πρόβλημα της διαφοροποίησης αποτελεσμάτων έχει μελετηθεί εκτεταμένα για αναζήτηση αδόμητων δεδομένων (εγγράφων, ιστοσελίδων), η δομημένη φύση των RDF δεδομένων απαιτεί διαφορετικά κριτήρια και μεθόδους. Οι περισσότερες σύγχρονες προσεγγίσεις για αναζήτηση με λέξεις κλειδιά σε γράφους [70, 71, 74] περιορίζουν τα αποτελέσματά τους σε μορφή δέντρων (ειδικότερα, παραλλαγές των Steiner δέντρων). Μόνο μερικές δουλειές επιτρέπουν αποτελέσματα σε μορφή γράφων [72, 73]. Μεταξύ αυτών, η δουλειά στο [72] είναι η πιο σχετική με τη δική μας. Παρόλα αυτά, δεν εξετάζει το πρόβλημα της διαφοροποίησης και ούτε λαμβάνει υπόψη το σχήμα των δεδομένων. Μία διαφορετική οπτική ακολουθείται στο [75], όπου ένα ερώτημα λέξεων κλειδιών μεταφράζεται πρώτα σε ένα σύνολο πιθανών δομημένων ερωτημάτων και, στη συνέχεια, τα πιο ετερογενή από αυτά τα ερωτήματα αποτιμώνται.

7.1.2 Διαφοροποιώντας αποτελέσματα αναζήτησης σε RDF δεδομένα

Θεωρούμε έναν RDF γράφο $G(V, E)$, όπου V είναι το σύνολο των κόμβων του και E των ακμών του. Προαιρετικά, ο γράφος G μπορεί να συσχετιστεί με ένα RDF σχήμα, το οποίο θα ορίζει ιεραρχίες από κλάσεις και ιδιότητες. Έστω $q = \{\{t_1, t_2, \ldots, t_m\}, k, \rho\}$ ένα ερώτημα αποτελούμενο από m όρους (λέξεις κλειδιά), μία παράμετρος k που προσδιορίζει το μέγιστο αριθμό αποτελεσμάτων και μία παράμετρος ρ που καθορίζει το μέγιστο μήκος μονοπατιού μεταξύ κόμβων που αντιστοιχούν σε λέξεις κλειδιά (όπως θα εξηγηθούν αργότερα). Έστω επίσης μία συνάρτηση $M\colon t \rightarrow V_t$ που αντιστοιχίζει μία λέξη (φράση) σε ένα σύνολο από κόμβους $V_t \subseteq V$.

Ορισμός 7.4 (Άμεσο μονοπάτι λέξεων κλειδιών). *Έστω δύο κόμβοι $u, v \in V$ που αντιστοιχούν σε δύο όρους (φράσεις) t, s ενός ερωτήματος q, δηλαδή $u \in V_t$ και $v \in V_s$. Έστω P ένα μονοπάτι ανάμεσα στους u και v. Το P ονομάζεται άμεσο μονοπάτι λέξεων κλειδιών αν δεν περιέχει κάποιον άλλο κόμβο που αντιστοιχίζεται σε κάποιον όρο του ερωτήματος q.*

Ορισμός 7.5 (Αποτέλεσμα). *Έστω ένας RDF γράφος G και ένα ερώτημα q. Ένας υπογράφος G_q του G αποτελεί αποτέλεσμα του q πάνω στον G, εάν: (α) για κάθε λέξη (φράση) κλειδί στο q, υπάρχει ακριβώς ένας κόμβος v στον G_q τέτοιος ώστε $v \in V_t$ (επονομαζόμενος κόμβος λέξης κλειδί) και (β) για κάθε ζεύγος κόμβων λέξης κλειδί u, v στον G_q, υπάρχει ένα μονοπάτι ανάμεσά τους με μέγιστο μήκος ρ και (γ) για κάθε ζεύγος κόμβων λέξης κλειδί u, v στον G_q, υπάρχει το πολύ ένα άμεσο μονοπάτι λέξεων κλειδιών ανάμεσά τους και (δ) κάθε κόμβος που δεν είναι κόμβος λέξης κλειδί βρίσκεται ενδιάμεσα σε ένα μονοπάτι που συνδέει κόμβους λέξης κλειδί.*

Οι παραπάνω ορισμοί οδηγούν σε αποτελέσματα που περιέχουν ανά δύο συνδέσεις (μονοπάτια) μεταξύ όλων των λέξεων κλειδιών ενός ερωτήματος. Δηλαδή, με βάση τα παραπάνω, απαιτούμε τα αποτελέσματα να έχουν μορφή γράφων και όχι ελάχιστων δέντρων. Αυτή η απαίτηση βασίζεται στη διαίσθηση του ότι θέλουμε να δώσουμε έμφαση στην πληρότητα των σχέσεων μεταξύ των όρων του ερωτήματος, παρά στην απαίτηση του αποτελέσματος ελάχιστου μεγέθους.

Έστω τώρα μία συνάρτηση $r\colon (G_q, q) \rightarrow [0,1]$ που ποσοτικοποιεί τη σχετικότητα ενός αποτελέσματος G_q με το ερώτημα q και μία συνάρτηση $d\colon (G_q, G'_q) \rightarrow [0,1]$ που μετράει την ανομοιότητα (απόσταση) μεταξύ δύο αποτελεσμάτων G_q και G'_q. Έστω, επίσης, η $f_{r,d}$ μία μονότονη αντικειμενική συνάρτηση που συνδυάζει τα δύο κριτήρια και αναθέτει σκορ σε ένα σύνολο αποτελεσμάτων R του ερωτήματος q, μετρώντας πόσο σχετικά είναι τα αποτελέσματα με το ερώτημα και πόσο ετερογενή μεταξύ τους. Στην περίπτωση που $|R| > k$, στόχος της διαφοροποίησης είναι να επιλέξει ένα υποσύνολο k αποτελεσμάτων ώστε η αντικειμενική συνάρτηση να μεγιστοποιείται. Τυπικά, το παραπάνω μπορεί να οριστεί ως εξής:

Ορισμός 7.6 (Διαφοροποιημένο σύνολο αποτελεσμάτων). *Έστω ένας RDF γράφος G, ένα ερώτημα q και οι συναρτήσεις r, d και $f_{r,d}$ όπως περιγράφηκαν παραπάνω. Έστω R το σύνολο αποτελεσμάτων. Το διαφοροποιημένο σύνολο αποτελεσμάτων R_k είναι ένα υποσύνολο του R με μέγεθος k, τέτοιο ώστε:* $R_k = \underset{R' \subseteq R, |R'|=k}{\operatorname{argmax}} f_{r,d}(R')$.

Ακολουθώντας την παραπάνω προσέγγιση πρέπει να οριστούν κατάλληλα οι συναρτήσεις r, d και $f_{r,d}$. Στο [96] ορίζονται διάφορες αντικειμενικές συναρτήσεις στόχοι και μελετώνται τα χαρακτηριστικά τους. Οι ίδιες συναρτήσεις μπορούν να εφαρμοστούν και στη δική μας περίπτωση, αφού η λειτουργία τους είναι ανεξάρτητη από τα υποκείμενα δεδομένα. Για αυτό, στη συνέχεια, επικεντρώνουμε σε μία γενική περιγραφή των συναρτήσεων ομοιότητας και ανομοιότητας r, d, στο σενάριό μας.

7.1.3 Κριτήρια διαφοροποίησης

Η βασική πρόκληση διαφοροποίησης είναι η ενσωμάτωση χαρακτηριστικών σημασιολογίας και δομής στις συναρτήσεις ομοιότητας και ανομοιότητας. Στη συνέχεια, ορίζουμε κάποια κριτήρια για αυτό το σκοπό.

Η ομοιότητα ενός αποτελέσματος με ένα ερώτημα λαμβάνει υπόψη δύο παράγοντες. Ο πρώτος αφορά στην κειμενική ομοιότητα ανάμεσα στους κόμβους (οντότητες) του RDF γρά-

φου και στους όρους του ερωτήματος. Ο δεύτερος παράγοντας αφορά στη συνεκτικότητα και ακρίβεια των αποτελεσμάτων, η οποία μπορεί να εξασφαλιστεί από τον ορισμό 7.5. Αυτή είναι μία ενδο-αποτελεσματική μετρική η οποία απαιτεί την εσωτερική ομοιογένεια ενός αποτελέσματος. Για παράδειγμα, αυτή η μετρική θα έδινε μεγαλύτερο σκορ σε ένα μονοπάτι του οποίου οι ακμές αντιστοιχούν στην ίδια RDF ιδιότητα. Επιπλέον, το RDF σχήμα μπορεί να ληφθεί υπόψη, για παράδειγμα υπολογίζοντας σκορ *ελαχίστου κοινού προγόνου* (least common ancestor).

Η ανομοιότητα μεταξύ αποτελεσμάτων μπορεί να οριστεί μέσω της σύγκρισης αντίστοιχων μονοπατιών ανάμεσα σε διαφορετικά γραφοαποτελέσματα. Η σύγκριση αυτή θα λάμβανε υπόψη δομικές ιδιότητες, όπως μεγέθη μονοπατιών ή κοινές ακμές και σημασιολογική πληροφορία, όπως κλάσεις και ιδιότητες που αντιστοιχούν σε κόμβους και ακμές των μονοπατιών. Στόχος εδώ είναι, για κάθε αποτέλεσμα, τα μονοπάτια να είναι όμοια με τα άλλα μονοπάτια του ίδιου αποτελέσματος, αλλά όσο γίνεται πιο ανόμοια με τα αντίστοιχα μονοπάτια άλλων αποτελεσμάτων.

7.2 Οργάνωση και αναζήτηση βιολογικών οντοτήτων

7.2.1 Εισαγωγή

Ένα από τα προβλήματα που καλούνται να επιλύσουν βιολογικές ομάδες είναι η συσχέτιση-ταίριασμα δύο συγκεκριμένων οντοτήτων: genes (γονίδια) και miRNAs. Για τέτοιες οντότητες προκύπτουν θέματα αλλαγής ονοματολογίας με την πάροδο του χρόνου. Σε πρώτη φάση, στη συγκεκριμένη δουλειά, καταπιαστήκαμε με αλλαγές ονοματολογίας σε miRNAs, ενώ παράλληλη προσέγγιση θα μπορούσε να ακολουθηθεί και για τις ονοματολογίες των γονιδίων.

Ύστερα από μελέτη αρχείων εκδόσεων βιολογικών ονομάτων, στα οποία κρατείται και στοιχειώδης πληροφορία ταυτοποίησης αλλαγών ονοματολογιών, αναγνωρίστηκαν οι εξής περιπτώσεις διαφορών στα ονόματα:

- Αλλαγή ονόματος με την πάροδο του χρόνου.
- Διαφορετικά (ταυτόχρονα) χρησιμοποιούμενα ονόματα με λίγο αλλαγμένο κάποιο μέρος του ονόματος (π.χ. κάποιο νούμερο στην αρχή ή το τέλος)
- Εντελώς διαφορετικά ονόματα.

Υπάρχουν βιολογικές ιστοσελίδες οι οποίες παρέχουν τις πληροφορίες (π.χ. ονόματα και κάποια χαρακτηριστικά) για όλα τα γονίδια[1] [2] [3] και για όλα τα miRNAs [4] που γνωρίζουμε. Όμως, καθώς η έρευνα προχωράει, νέες οντότητες ανακαλύπτονται, λάθη του παρελθόντος διορθώνονται κτλ. Για το λόγο αυτό ανανεώνεται και η πληροφορία στις συγκεκριμένες ιστοσελίδες. Οι παλιές εκδόσεις διατηρούνται και αυτές σε κάποια μορφή, ενώ, ταυτόχρονα με

[1] http://www.ensembl.org/
[2] http://www.ncbi.nlm.nih.gov/, http://www.genenames.org/
[3] http://www.ihop-net.org/UniPub/iHOP/
[4] http://www.mirbase.org/

κάθε καινούρια έκδοση, δίνεται και κάποιο *diff* αρχείο (αρχείο αλλαγών) το οποίο περιλαμβάνει αλλαγές που έχουν συμβεί σε σχέση με την προηγούμενη έκδοση.

Εν προκειμένω, τα miRNAs, χωρίζονται σε 2 κατηγορίες: τα hairpins και τα matures. Πρακτικά, τα hairpins παράγουν τα matures (σχέση m σε n). Στις περισσότερες περιπτώσεις, το όνομα του hairpin είναι ίδιο ή παρόμοιο με του mature. Ένα σημαντικό πρόβλημα είναι ότι οι βιολόγοι, στα αρχεία που παρουσιάζονται παρακάτω, κωδικοποιούν την πληροφορία με βάση τα hairpins. Αντιθέτως, κωδικοποιούν τις δημοσιεύσεις με βάση τα matures. Το δεύτερο, και πιο σημαντικό, πρόβλημα είναι ότι οι οντότητες των miRNAs αλλάζουν με το χρόνο (ονοματολογία, ακολουθία, κ.α.), αλλά και συσχετίζονται μεταξύ τους με διαφόρων ειδών συσχετίσεις (ίδια οικογένεια, ίδιο είδος, κ.α.).

Πληροφορίες που σχετίζονται με τις αλλαγές των hairpins κωδικοποιούνται από το mirBase[5], μία δημοφιλή βάση βιολογικών οντοτήτων, στα ακόλουθα αρχεία:

- miRNA.dat: Το συγκεκριμένο αρχείο συγκεντρώνει όλες τις πληροφορίες που αναφέρονται σε κάθε hairpin miRNA της τρέχουσας έκδοσης (αναγνωριστικό, όνομα, σχετιζόμενα matures, σημαντικές σχετιζόμενες δημοσιεύσεις, ακολουθία κ.α.).

- miRNA.diff: Το συγκεκριμένο αρχείο συγκεντρώνει γενικές πληροφορίες για τις αλλαγές που έχουν συμβεί σε σχέση με τις προηγούμενες εκδόσεις. Από ό,τι έχουμε διαπιστώσει οι αλλαγές αυτές περιορίζονται στις εξής περιπτώσεις:

  - NEW: Πρόκειται για κάποιο καινούριο miRNA.
  - DELETE: Πρόκειται για κάποιο miRNA που υπήρχε σε προηγούμενη έκδοση της βάσης, όμως διαγράφηκε στην τρέχουσα για κάποιον λόγο. Τα miRNA που γίνονται DELETE εγγράφονται στο αρχείο dead. Υπάρχουν 2 υποπεριπτώσεις: το miRNA να διαγράφεται εντελώς ή να διαγράφεται και να δείχνει σε κάποιο άλλο με το οποίο με κάποιο τρόπο σχετίζεται.
  - SEQUENCE: Πρόκειται για κάποιο miRNA του οποίου η ακολουθία έχει τροποποιηθεί σε σχέση με προηγούμενες εκδόσεις.
  - NAME: Πρόκειται για κάποιο miRNA το οποίο έχει καινούριο όνομα/αναγνωριστικό (αυτό που αναφέρεται). Για παράδειγμα, στην έκδοση 13 υπήρχε το bta-mir-26a, το οποίο μετονομάστηκε στην έκδοση 14 se bta-mir-26a-2.
  - MATURE: Αλλαγή στο mature που παράγεται από το συγκεκριμένο hairpin.
  - ΣΥΝΔΥΑΣΜΟΙ: των sequence, name και mature σε ενιαία αλλαγή.

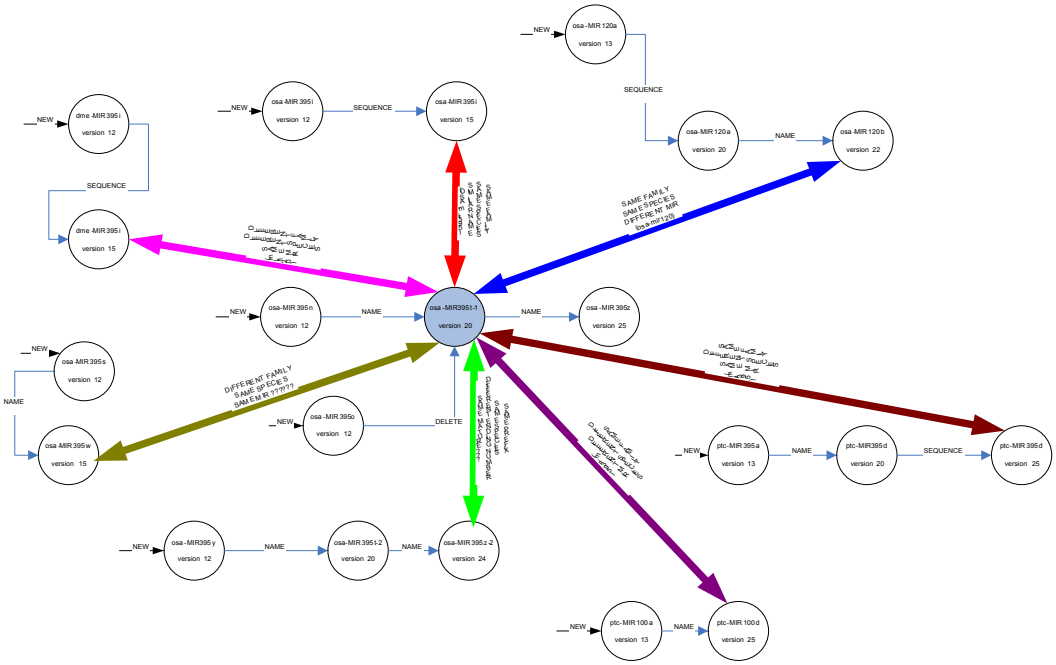
- miRNA.dead: Το συγκεκριμένο αρχείο συγκεντρώνει όλα τα miRNA τα οποία έχουν πια διαγραφεί. Το αρχείο αυτό είναι αυξητικό, με την έννοια ότι σε κάθε νέα έκδοση το αρχείο αυτό περιέχει ό,τι περιείχε και το αντίστοιχο αρχείο στην προηγούμενη έκδοση, συν κάποιες επιπλέον εγγραφές.

[5]http://www.mirbase.org/index.shtml

- miFam.dat: Το συγκεκριμένο αρχείο οργανώνει τα hairpin miRNAs σε οικογένειες, δηλαδή σε ομάδες των οποίων κοινό χαρακτηριστικό είναι τα παρόμοια matures.

Επιπλέον, από μελέτη των παραπάνω αρχείων και συζητήσεις με τους βιολόγους, καταλήξαμε ότι διάφορα miRNAs μπορεί να συσχετίζονται μεταξύ τους με τις παρακάτω συσχετίσεις:

- Ίδιο παραγόμενο mature.
- Ίδιο family (οικογένεια), δηλαδή παρόμοιο παραγόμενο mature.
- Ίδιο είδος.
- Ίδιος αύξων αριθμός στο όνομα του miRNA.
- Συν-εκφραζόμενα miRNAs.



Σχήμα 7.2: Μοντελοποίηση βιολογικών αλλαγών/συσχετίσεων

Μία πρώτη προσπάθεια μοντελοποίησης των πιθανών αλλαγών και συσχετίσεων που μπορεί να αφορούν ένα miRNA φαίνεται στο παρακάτω παράδειγμα:

Σχήμα βάσης

Όλα τα δεδομένα από τα διαθέσιμα αρχεία έχουν οργανωθεί και φορτωθεί στους ακόλουθους πίνακες.

Το ζητούμενο από όλα τα παραπάνω είναι η οργάνωση ενός μηχανισμού αναζήτησης και ταξινόμησης αποτελεσμάτων, έτσι ώστε, όταν ένας χρήστης ψάχνει στη βάση του PubMed[6]

[6] http://www.ncbi.nlm.nih.gov/pubmed/

biosearch.hairpins
hid : int(11)
hname : varchar(30)
version : int(11)
vchange : int(11)
comment : varchar(200)

Πίνακας 7.1: hairpins: Πίνακας που οργανώνει όλες τις αλλαγές των hairpins κρατώντας όνομα, αναγνωριστικό, αλλαγή, και έκδοση του mirbase στην οποία έγινε η αλλαγή.

biosearch.famtable
fid : int(11)
fname : varchar(30)
fhid : int(11)
fhname : varchar(30)
fversion : int(11)

Πίνακας 7.2: famtable: Πίνακας που οργανώνει τις ομαδοποιήσεις των hairpins σε οικογένειες, ανά έκδοση του mirbase.

biosearch.mattable
hid : int(11)
mid : int(11)
mname : varchar(50)

Πίνακας 7.3: mattable: Πίνακας που συσχετίζει hairpins με matures.

biosearch.mature
mid : int(11)
mname : varchar(30)
version : int(11)
mchange : int(11)

Πίνακας 7.4: mature: Πίνακας που οργανώνει όλες τις αλλαγές των matures κρατώντας όνομα, αναγνωριστικό, αλλαγή, και έκδοση του mirbase στην οποία έγινε η αλλαγή.

biosearch.pubstable
hid : int(11)
prank : int(11)
pid : int(11)

Πίνακας 7.5: pubstable: Πίνακας που συσχετίζει hairpins με σημαντικές δημοσιεύσεις.

biosearch.weight
chid : int(11)
chweight : float
chname : varchar(100)

Πίνακας 7.6: weight: Πίνακας που κρατάει το βάρος κάθε αλλαγής και συσχέτισης.

biosearch.searchtable
mid : int(11)
mname : varchar(50)
msyn : varchar(50)
mweight : float
mschange : int(11)

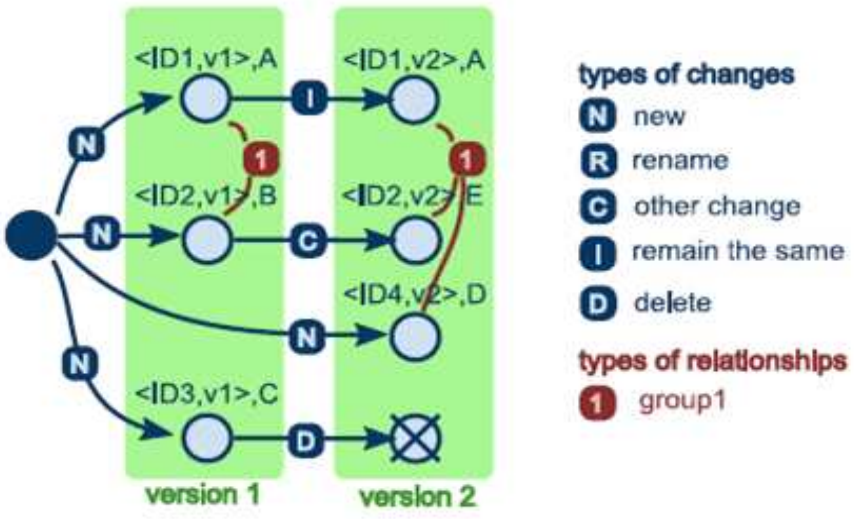
Πίνακας 7.7: searchtable: Πίνακας που τελικά θα κρατάει για κάθε hairpin, όλα τα σχετιζόμενα ονόματα, μαζί με συγκεκριμένο βάρος που θα έχει υπολογιστεί από το ranking μοντέλο μας.

για δημοσιεύσεις σχετικές με ένα mature miRNA, να του επιστρέφει αποτελέσματα με βάση όλες τις δυνατές συσχετίσεις που προκύπτουν από αυτά που περιγράφησαν παραπάνω, κατάλ-

ληλα ταξινομημένες και με τις κατάλληλες επεξηγήσεις όσον αφορά στην προέλευση της κάθε δημοσίευσης. Μεσοπρόθεσμα, στόχος μας είναι η προσθήκη training μηχανισμού, έτσι ώστε η βαρύτητα κάθε συσχέτισης να προκύπτει από τις αναζητήσεις και τις επιλογές που έκαναν στο παρελθόν οι χρήστες.

Μοντέλο δεδομένων

Στο παρακάτω σχήμα παρουσιάζεται ένα πιθανό μοντέλο οργάνωσης δεδομένων με βάση την προηγούμενη ανάλυση.



Σχήμα 7.3: Μοντελοποίηση βιολογικών οντοτήτων και συσχετίσεων σε μορφή εκδόσεων

Αλγόριθμος διάσχισης γράφου εκδόσεων βιολογικών οντοτήτων

Στη συνέχεια ορίζουμε μία πρώτη εκδοχή ενός αλγορίθμου διάσχισης και ενός αλγορίθμου ευρετηρίασης των παραπάνω ορισμένων μοντέλων δεδομένων. Ο αλγόριθμος διάσχισης βασίζεται στο Σχήμα 7.3.

Algorithm 7 Traversal Algorithm

1. Start from the name that was searched by the user (circles in the green area).

2. Go backward and forward gathering changes (blue arrows between circles) between different names of the SAME entity (hairpin).

a. Assign weights on the different names based on the gathered changes between two different names.

3. Go to the last version of the entity containing the searched name and find relations with other entities.

a. Assign weights relating the LAST names of each entity.

4. Combine weights from 2. and 3.

$S = \emptyset$

Algorithm 8 Index Creation

1. Create a first conventional inverted index, indexing bio entities.

2. Query the first index and combine returned scores and the database stored relations (weights):

a. Query for a name and all its synonyms

b. Create fake replicated docs: 1 for each synonym:

i. Replicated documents have no content (text), only the same title and a field with the name and a field with value: (synonym score*corresponding weight).

c. Create a second index containing both initial and fake documents

3. When asking, pose the initial query (a) to title and content fields and (b) to the synonyms fields. In this way, the documents regarding the initial name are returned from the first query and the documents regarding the synonyms from the second query.

Κεφάλαιο 8

Σύνοψη

Η παρούσα εργασία επικεντρώνεται σε προβλήματα αναταξινόμησης αποτελεσμάτων με χρήση μεθοδολογιών μηχανικής μάθησης, ευριστικών αλγορίθμων και σημασιολογικής μεταπληροφορίας, με σκοπό την εξατομίκευση, διαφοροποίηση και συνδυασμό των αποτελεσμάτων αναζήτησης. Συγκεκριμένα, προτείνονται και υλοποιούνται μεθοδολογίες για (α) τον εμπλουτισμό των αρχικών δεδομένων εκπαίδευσης συναρτήσεων ταξινόμησης αποτελεσμάτων, για ταχύτερη εκπαίδευση χωρίς επιπλέον επιβάρυνση των χρηστών, (β) τη συνεργατική εκπαίδευση πολλαπλών συναρτήσεων ταξινόμησης και την επιλεκτική χρησιμοποίησή τους, με σκοπό τη βελτίωση της ποιότητας εκπαίδευσης, (γ) προσαρμοστική αναζήτηση με τη βοήθεια σημασιολογικών δεδομένων (υβριδική αναζήτηση σημασιολογικά επισημειωμένων εγγράφων, εξατομικευμένη αναζήτηση σημασιολογικών δεδομένων, διαφοροποίηση αναζήτησης σημασιολογικών δεδομένων) και (δ) διαφοροποίηση σχολίων χρηστών σε κοινωνικά δίκτυα. Οι παραπάνω μεθοδολογίες είναι αρκετά ευέλικτες ώστε να μπορούν να υιοθετηθούν από διαφορετικές προσεγγίσεις συστημάτων αναζήτησης, εξατομίκευσης και διαφοροποίησης και μπορούν να εφαρμοστούν σε πληθώρα ακαδημαϊκών και εμπορικών συστημάτων (διαδικτυακές ή επιτραπέζιες μηχανές αναζήτησης, ψηφιακές βιβλιοθήκες κ.α.).

Οι επεκτάσεις της υπάρχουσας δουλειάς αφορούν τις εξής κατευθύνσεις:

1. Επέκταση/βελτίωση των ίδιων των μεθόδων, όπως προτείνεται στην περιγραφή της μελλοντικής δουλειάς των επιμέρους εργασιών.

2. Εφαρμογή των παραπάνω μεθοδολογιών σε διάφορα επιστημονικά πεδία με κατάλληλη προσαρμογή τους ώστε να ικανοποιούν τις συγκεκριμένες απαιτήσεις της εκάστοτε περιοχής. Ήδη έχει γίνει μία πρώτη προεργασία για εφαρμογή των παραπάνω σε βιολογικές βάσεις δεδομένων και συγκεκριμένα σε αναζήτηση βιολογικών δημοσιεύσεων με βάση σύνθετες ονοματολογίες βιολογικών οντοτήτων.

3. Επέκταση και βελτίωση της εφαρμογής των παραπάνω μεθοδολογιών σε δεδομένα σημασιολογικού ιστού ώστε να υπολογίζονται, πλέον, η δομή των δεδομένων και οι σχέσεις που τα συνδέουν, ως επιπλέον συνιστώσες για την εξατομίκευση/διαφοροποίηση των αποτελεσμάτων αναζήτησης. Η διατριβή εξέτασε σε σημαντικό βαθμό αυτήν την κατεύθυνση, αλλά, αδιαμφισβήτητα, υπάρχει περιθώριο επέκτασης της δουλειάς ή εισαγωγής

νέων/διαφορετικών μεθόδων.

Βιβλιογραφία

[1] R. Herbrich, T. Graepel and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers, MIT Press*, Pages: 115-132, 2000.

[2] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

[3] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248, 2005.

[4] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In Poster Abstract, Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), 2004.

[5] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, pages 675–684, 2004.

[6] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2005.

[7] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, 2007.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[9] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, 2006.

[10] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[11] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4:933–969, 2003.

[12] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193, 2006.

[13] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–579, 2007.

[14] S. Pandey, S. Roy, C. O. J. Cho, and S. Chakrabarti. Shuffling a stacked deck: the case for partially randomized ranking of search engine results. In *Proceedings of the 31st international conference on Very large data bases*, pages 781–792, 2005.

[15] T.-H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526, 2002.

[16] G. Jeh, and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, 2003.

[17] U. Rohini and V. Ambati. Improving Re-ranking of Search Results Using Collaborative Filtering. *Information Retrieval Technology, Third Asia Information Retrieval Symposium, AIRS 2006*, pages 205–216, 2006.

[18] P.-A. Chirita, C.-S. Firan, and W. Nejdl. Summarizing local context to personalize global web search. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 287–296, 2006.

[19] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 718–723, 2006.

[20] T. Qin, X.-D. Zhang, D.-S. Wang, T.-Y. Liu, W. Lai, and H. Li. Ranking with multiple hyperplanes. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 279–286, 2007.

[21] X. Li, N. Wang, and S.-Y. Li. A fast training algorithm for svm via clustering technique and gabriel graph. In *Proceedings of the International Conference on Intelligent Computing*, 2007.

[22] J. Diez, J. J. del Coz, O. Luaces, and A. Bahamonde. Clustering people according to their preference criteria. *Expert Systems with Applications: An International Journal*, 34:1274–1284, 2008.

[23] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.

[24] Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168, 2005.

[25] S. E. Robertson. Overview of the okapi projects. *Journal of Documentation*, 53(1):3–7, 1997.

[26] P.-N. Tan, M. Steinbach, and V. Kumar. Cluster Analysis: Basic Concepts and Algorithms. *Introduction to Data Mining. Pearson Addison Wesley, Boston*, 2006.

[27] T. Qin, T.-Y. Liu, J. Xu, and H. Li. LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. *Information Retrieval Journal*, 2010.

[28] http://www.dmoz.org.

[29] http://research.microsoft.com/en-us/um/beijing/ projects/letor/LETOR4.0/Data/ Features_in_LETOR4.pdf.

[30] http://lucene.apache.org/.

[31] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4dataset.aspx.

[32] http://trec.nist.gov/.

[33] S. Fox, K. Karnawat, M. Mydland, S. Dumais and T. White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23(2):147–168, 2005.

[34] Z. Dou, R. Song, J.-R. Wen, and X. Yuan. Evaluating the Effectiveness of Personalized Web Search. *IEEE TKDE*, 21:1178–1190, 2008.

[35] Z. Zheng, K. Chen, G. Sun, and H. Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the ACM SIGIR Conference*, 2007.

[36] W. Chu, and S.-S. Keerthi. Support Vector Ordinal Regression. *Neural Computation*, 19:792–815, 2007.

[37] J.-W. Kim, and K.-S. Candan. Skip-and-prune: cosine-based top-k query processing for efficient context-sensitive document retrieval. In *Proceedings of the ACM SIGMOD Conference*, 2009.

[38] J. Teevan, S.-T. Dumais, and D.-J. Liebling. To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent. In *Proceedings of the ACM SIGIR Conference*, 2008.

[39] M. Dudev, S. Elbassuoni, J. Luxemburger, M. Ramanath, and G. Weikum. Personalizing the Search for Knowledge. In *2nd International Workshop on Personalized Access, Profile Management, and Context Awareness: Databases* , 2008.

[40] C. Rocha, D. Schwabe, and M. P. Poggi. Hybrid approach for searching in the semantic web. In *Proc. of the 13th international conference on World Wide Web* , 374–383, 2004.

[41] X. Jiang and A.H. Tan. Learning and inferencing in user ontology for personalized Semantic Web search. In *Information Sciences: an International Journal*, Vol. 179, Issue 16, 2794–2808, 2009.

[42] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Learning Semantic Query Suggestions. In *Proc. of the 8th International Semantic Web Conference* , 424–440, 2009.

[43] L. Dali and B. Fortuna, T. Tran and D. Mladenic. Query-Independent learning to rank for RDF entity search. In *Proc. of the 9th international conference on The Semantic Web: research and applications* , 484–498, 2012.

[44] A. Sieg, B. Mobasher and R. Burke. Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search. In *IEEE Intelligent Informatics Bulletin* , Vol. 8, No. 1, November 2007.

[45] K. W.-T. Leung, D. L. Lee, W. Ng and H. Y. Fung. A Framework for Personalizing Web Search with Concept-Based User Profiles. In *ACM Transactions on Internet Technology* , Vol. 11, No. 4, Article 17, 2012.

[46] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (39)*, pages 1–38, 1977.

[47] A. Banerjee, I. Dhillon, j. Ghosh and S. Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning*, 38(6):1345–1382, 2005.

[48] T. Joachims. Training Linear SVMs in Linear Time. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[49] J. Bian, X. Li, F.-Li. Liu, Z. Zheng, and H. Zha. Ranking Specialization for Web Search: A Divide-and-Conquer Approach by Using Topical RankSVM. In *Proceedings of the 19th international conference on World wide web*, pages 131–140, 2010.

[50] Mitchell T.M. Machine Learning. In *WCB/McGraw-Hill*, 1997.

[51] Handschuh, S., and Staab, S. (eds.) Annotation for the Semantic Web. In *IOS Press*, 2003.

[52] Agosti, M., and Ferro, N. A Formal Model of Annotations of Digital Content. In *ACM Transactions on Information Systems (TOIS)*, volume 26, issue 3, pages 1–57, 2008.

[53] Reeve L., and Han H. Survey of semantic annotation platforms. In *Proceedings ofthe ACM Symposium on Applied Computing*, 2005.

[54] Uren V. S., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E., and Ciravegna F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. In *Journal of Web Semantics*, volume 4, 2006.

[55] Kiyavitskaya N., Zeni N., Cordy J.R., Mich L., and Mylopoulos J. Cerno: Lightweight tool support for semantic annotation of textual documents. In *Data Knowl. Eng. (DKE)*, volume 68, issue 12, 2009.

[56] Hogue A., and Karger D. Thresher: automating the unwrapping of semantic content from the World Wide Web. In *Proceedings of the WWW Conference*, 2005.

[57] Cimiano P., Handschuh S., and Staab S. Towards the self-annotating web. In *Proceedings of the WWW Conference*, 2004.

[58] Dill S., Eiron N., Gibson D., Gruhl D., Guha R., Jhingran A., Kanungo T., McCurley K. S., Rajagopalan S., Tomkins A., Tomlin J. A., and Zien J. Y. A Case for Automated Large-Scale Semantic Annotation. In *Journal of Web Semantics*, volume 1, issue 1, 2003.

[59] SMORE: Create OWL Markup for HTML Web Pages.
http://www.mindswap.org/2005/SMORE/.

[60] Handschuh, S., Staab, S., Ciravegna, F. S-CREAM: Semi-automatic CREAtion of Metadata. In *Proceedings of EKAW*, 2002.

[61] Vargas-Vera, M., Motta, E., Domingue, and J., Lanzoni (et.al). MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup. In *Proceedings of EKAW*, 2002.

[62] Cunningham H., Maynard D., Bontcheva K., and Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the ACL*, 2002.

[63] Kiryakov A., Popov B., Terziev I., Manov D., and Ognyanoff D. Semantic annotation, indexing, and retrieval. In *Journal of Web Semantics*, volume 2, issue 1, 2004.

[64] Chakravarthy A., Lanfranchi V., and Ciravegna F. Cross-media document annotation and enrichment. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop*, 2006.

[65] Eriksson H. An annotation tool for semantic documents. In *Proceedings of the ESWC*, 2007.

[66] Tallis M. SemanticWord processing for content authors. In *Proceedings of the Knowledge Markup and Semantic Annotation Workshop*, 2003.

[67] Mangold C. A survey and classification of semantic search approaches. In *Int. J. Metadata Semantics and Ontology*, volume 2, issue 1, 2007.

[68] Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D. Hybrid search: Effectively combining keywords and semantic searches. In *Proceedings of the ESWC*, 2008.

[69] Giunchiglia F., Kharkevich U., Zaihrayeu I. Concept search. In *Proceedings of the ESWC*, 2009.

[70] Tran T., Wang H., Rudolph S., Cimiano P. Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data. In ICDE 2009.

[71] Zhou Q., Wang C., Xiong M., Wang H., Yu Y. SPARK: Adapting Keyword Query to Semantic Search. In ISWC 2007.

[72] Elbassuoni S., Blanco R. Keyword Search over RDF Graphs. In CIKM 2011.

[73] Li G., Ooi B-C, Feng J., Feng J., et.al EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured data. In SIGMOD 2008.

[74] He H., Wang H., Yang J., Yu P. BLINKS: Ranked Keyword Searches on Graphs. In SIGMOD 2007.

[75] Demidova E., Fankhauser P., Zhou X., Nejdl W. DivQ: Diversification for Keyword Search over Structured Databases. In SIGIR 2010.

[76] Stefanidis K., Drosou M., Pitoura E. PerK: Personalized Keyword Search in Relational Databases through Preferences. In EDBT 2010.

[77] Kucuktunc, O., and Cambazoglu, B. B., and Weber, I., and Ferhatosmanoglu, H. A large-scale sentiment analysis for Yahoo! answers. In *Proceedings of the 5th ACM international conference on Web search and data mining (WSDM'12)*, pages 633–642, 2012.

[78] Tsagkias, E., Weerkamp, W., and de Rijke, M. News Comments: Exploring, Modeling, and Online Predicting. In *Proceedings of the 32nd European Conference on Information Retrieval (ECIR '10)*, pages 109–203, 2010.

[79] Tsagkias, E., Weerkamp, W., and de Rijke, M. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, pages 1765–1768.

[80] Diakopoulos, N., and Naaman, M. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW '11)*, pages 133–142, 2011.

[81] Park, S., Ko, M., Kim, J., Liu, Y., and Song, J. The Politics of Comments: Predicting Political Orientation of News Stories with Commenters" Sentiment Patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW '11)*, pages 113–122, 2011.

[82] Herring, S.C., Kouper, I., Paolillo, J.C., Scheidt, L.A., Tyworth, M., Welsch, P., Wright, E., and Ning Y. Conversations in the Blogosphere: An Analysis "From the Bottom Up". In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, (HICSS '05)*, pages 107b–107b, 2005.

[83] Potthast, M. Measuring the descriptiveness of web comments. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development (SIGIR '09)*, pages 724–725, 2009.

[84] Li, Q., Wang, J., Chen, Y. P., and Lin, Z. User comments for news recommendation in forum-based social media. In *Information Sciences: an International Journal*, volume 180, issue 24, pages 4929–4939, 2010.

[85] Shmueli, E., Kagian, A., Koren, Y., and Lempel, R. Care to Comment? Recommendations for Commenting on News Stories. In *Proceedings of the 18th international conference on World wide web (WWW '12*, to appear, 2012.

[86] Hu, M., Sun, A., and Lim, E. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*, pages 291–298, 2008.

[87] Mishne, G.A., and Glance, N. Leave a Reply: An Analysis of Weblog Comments. In *Proceedings of the WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW ': the 15th international conference on World Wide Web*, 2006.

[88] Wong, D., Faridani, S., Bitton, E., Hartmann, B., and Goldberg, K. The diversity donut: enabling participant control over the diversity of recommended responses. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems (CHI EA '11)*, pages 1471–1476, 2011.

[89] Munson, S. A., and Resnick, P. Presenting diverse political opinions: how and how much. In *Proceedings of the 28th international conference on Human factors in computing systems (CHI '10)*, pages 1457–1466, 2010.

[90] Erkut, E. The discrete p-dispersion problem. In *Operations Research Letters*, volume 46, issue 1, pages 48–60, 1990.

[91] Erkut, E., Ülküsal, Y., and Yeniçerioglu, O. A comparison of p-dispersion heuristics. In *Computers & Operations Research*, volume 21, issue 10, pages 1103–1113, 1994.

[92] Chandra, B., and Halldórsson, M. M. Approximation Algorithms for Dispersion Problems. In *Journal of Algorithms*, volume 38, issue 2, pages 438•-465, 2001.

[93] Carbonell, J., and Goldstein, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98)*, pages 335–336, 1998.

[94] Chen, H., and Karger, D. R. Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*, pages 429–436, 2006.

[95] Clarke, C. L.A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*, pages 659–666, 2008.

[96] Gollapudi, S., and Sharma, A. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web (WWW '09)*, pages 381–390, 2009.

[97] Agrawal R., Gollapudi S., Halverson A., and Ieong S. Diversifying search results. In *Proceedings of the Second International Conference on Web Search and Web Data Mining (WSDM 2009)*, pages 5-14, 2009.

[98] Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., and Yahia, S. A. Efficient Computation of Diverse Query Results. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE '08)*, pages 228–236, 2008.

[99] Drosou, M., and Pitoura, E. Search result diversification. In *ACM SIGMOD Record*, volume 39, issue 1, pages 41–47.

[100] Hassin R., Rubinstein S., and Tamir A. Approximation algorithms for maximum dispersion. In *Operations Research Letters*, volume 21, issue 3, pages 133–137, 1997.

[101] Ravi, SS., Rosenkrantzt D.J., and Tayi. G.K. Approximation Algorithms for Facility Dispersion. In *In Teofilo F. Gonzalez, editor, Handbook of Approximation Algorithms and Metaheuristics.*, Chapman & Hall/CRC, 2007.

[102] Finkel, J. R., Grenager, T., and Manning, C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 363–370, 2005.

[103] Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., and Kappas, A. Sentiment strength detection in short informal text. In *Journal of the American Society for Information Science and Technology*, volume 61, issue 12, pages 2544-•2558, 2010.

[104] Giannopoulos, G., Biliri, E., and Sellis, T. Personalizing keyword search on RDF data. in *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'13)*, September 2013.

[105] Bikakis, N., Giannopoulos, G., Liagouris, J., Skoutas, D., Dalamagas, T., and Sellis, T. RDivF: Diversifying Keyword Search on RDF Graphs. in *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'13)*, September 2013.

[106] Giannopoulos, G., Weber, I., Jaimes, A., and Sellis, T. Diversifying User Comments on News Articles. In *Proceedings of the 13th International Conference Web Information Systems Engineering (WISE '12)*, pages 100–113, 2012.

[107] Giannopoulos, G. Personalizing Search Results on User Intent. In *Proceedings of the VLDB PhD workshop, in conjunction with the 38th International Conference on Very Large Databases (VLDB'12)*, 2011.

[108] G. Giannopoulos, U. Brefeld, T. Dalamagas, and T. Sellis. Learning to rank user intent. In *Proceedings of the 20th ACM International Conference on Information and knowledge management*, 2011.

[109] G. Giannopoulos, T. Dalamagas, and T. Sellis. Search behavior-driven training for result re-ranking. In *Proceedings of the 15th international conference on Theory and practice of digital libraries*, 2011.

[110] Bikakis, N., Giannopoulos, G., Dalamagas, T., and Sellis, T. Integrating Keywords and Semantics on Document Annotation and Search. In *OTM Conferences*, pages 921–938, 2010.

[111] Giannopoulos, G., Bikakis, N., Dalamagas, T., and Sellis, T. GoNTogle: A Tool for Semantic Annotation and Search. In *Proceedings of the The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010*, pages 376-380, 2010.

[112] G. Giannopoulos, T. Dalamagas and T. Sellis. Collaborative Ranking Function Training for Web Search Personalization. In *PersDB Workshop*, 2009.

[113] G. Giannopoulos, T. Dalamagas, M. Eirinaki and T. Sellis. Boosting the ranking function learning process using clustering. *Proceedings of the 10th ACM International Workshop on Web Information and Data Management (WIDM 2008)*, pages 125–132, 2008.

[114] Manolis Maragkakis, Panagiotis Alexiou, Giorgio L Papadopoulos, Martin Reczko, Theodore Dalamagas, George Giannopoulos, George Goumas, Evangelos Koukis, Kornilios Kourtis, Victor A Simossis, Praveen Sethupathy, Thanasis Vergoulis, Nectarios Koziris, Timos Sellis, Panagiotis Tsanakas and Artemis G Hatzigeorgiou. Accurate microRNA target prediction correlates with protein repression levels. In *BMC Bioinformatics*, volume 10, pages 295, 2009.

[115] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. DIANA-microT web server: elucidating microRNA functions through target prediction. In *Nucleic Acids Research*, volume 37, pages 273–276, 2009.

Παράρτημα Α΄

Μεταφράσεις Ξένων Όρων

Μετάφραση	Αγγλικός όρος
αδόμητα δεδομένα	unstructured data
ανάδραση χρήστη	user feedback
ανάκτηση πληροφορίας	information retrieval
αναταξινόμηση	reranking
ανεστραμμένο ευρετήριο	inverted index
αντικείμενο	object
άπληστη ευριστική	greedy heuristic
απόλυτη/σχετική σχετικότητα	absolute/relative relevance
απόσπασμα	snippet
βάρος	weight
βασική αλήθεια	ground truth
βασισμένος στην κλίση	online gradient-based
γλώσσα οντολογιών ιστού	Web Ontology Language - OWL
δήλωση	statement
δεδομένα αναζήτησης	clickthrough data
διάδοση ενεργοποίησης	spreading activation
διάνυσμα βαρών	weight vector
διάνυσμα χαρακτηριστικών	feature vector
διάνυσμα όρων	term vector
διαφοροποίηση	diversification
διχοτόμηση	partition
εξαγωγή πληροφορίας	Information Extraction
εξατομίκευση αναζήτησης	searc personalization
ελάχιστος κοινός πρόγονος	least common ancestor
ελεύθερη επισημείωση	tagging
έννοια	concept
επιγραφή	label
ετερογένεια	diversity

ευριστική	heuristic
ευριστική αυστηρής ανάθεσης	hard assignment heuristic
ιδιότητα	property
ιστορικό αναζήτησης	search history
καθολικό αναγνωριστικό πόρου	Internationalized Resource Identifier - IRI
καινοτομία	novelty
κανονικοποιημένο αθροιστικό κέρδος	normalized discount cumulative gain - NDCG@n
κατευθυνόμενος, ονοματισμένος γράφος	directed labelled graph
κατηγόρημα	predicate
κατηγοριοποίηση	classification
κεντροειδές	centroid
κειμενική ομοιότητα	textual similarity
κλάσση	class
κρίση σχετικότητας αποτελέσματος	relevance judgment
λεξιλόγιο	vocabulary
μέγιστη οριακή σχετικότητα	maximal marginal relevance
μεγιστοποίηση της προσδοκίας	expectation maximization - EM
μέση ακρίβεια	mean average precision - MAP
μηχανές διανυσμάτων στήριξης	support vector machines - SVM
μηχανική μάθηση	machine learning
μονάδα πληροφορίας	information nugget
μπλογκ	blog
ν-γράμμα	n-gram
οντολογία	ontology
ονοματική οντότητα	named entity
παγκόσμιος ιστός	World Wide Web
παραγωγή προτάσεων	recommendation
πληροφοριακή ανάγκη	information need
πολωμένος	biased
πόρος	resource
προσέγγιση βασισμένη στη μνήμη	memory based approach
προσέγγιση βασισμένη σε μοντέλο	model based approach
προσομοιωμένη ανόπτηση	simulated annealing
πρόταση ερωτημάτων	query suggestion
προφίλ	profile
σκοπός αναζήτησης	search intent
σκορ	score
σημασιολογικός ιστός	semantic web
στατιστικά σημαντικός	statistically significant
συμπεριφορά αναζήτησης	search behavior
συμφραζόμενα	context

συνάρτηση αναταξινόμησης αποτελεσμάτων	ranking function
συνάρτηση-κριτήριο	criterion function
συνάρτηση απόστασης	distance function
συνάρτηση ομοιότητας	similarity function
συνάρτηση συνημιτόνου	cosine function
συνεδρία αναζήτησης	search session
συνεργατικό φιλτράρισμα	collaborative filtering
συσταδοποίηση	clustering
σχήμα	schema
ταίριασμα μοτίβων γράφων	graph pattern matching
τετραγωνικός προγραμματισμός	quadratic programming
υποκείμενο	subject
φιλτράρισμα	filtering
φόρουμ	forum
χαρακτηριστικό	feature
ψήφος πλειοψηφίας	majority vote

Παράρτημα Β΄

Βιογραφικό Σημείωμα

Στοιχεία Επικοινωνίας

Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Ηρώων Πολυτεχνείου 9. Ζωγράφου
157 80 Αθήνα, Ελλάδα

Τηλέφωνο: (+30) 210 772 1402
Fax: (+30) 210 772 1442
Ηλεκτρονικό ταχυδρομείο (e-mail): giann@dblab.ece.ntua.gr
Προσωπική Σελίδα: http://www.dblab.ece.ntua.gr/~giann

Σπουδές

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2006–σήμερα)
  Υποψήφιος Διδάκτωρ της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών.
  Περιοχή έρευνας: Διαχείριση Δεδομένων και Υπηρεσιών στον Παγκόσμιο Ιστό
  Επιβλέπων: καθ. Τιμολέων Σελλής

- **Yahoo! Research**, Βαρκελώνη, Ισπανία (10/2011 – 12/2011)
  Σύντομη ερευνητική επίσκεψη.

- **Yahoo! Research**, Βαρκελώνη, Ισπανία (2/2011 – 4/2011)
  Σύντομη ερευνητική επίσκεψη.

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2001–2006)
  Δίπλωμα Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών.
  Βαθμός: 8.82/10
  Διπλωματική εργασία: GoNToggle: Έξυπνη μηχανή αναζήτησης με χρήση οντολογιών

Επιβλέπων: καθ. Τιμολέων Σελλής

Ερευνητικά Ενδιαφέροντα

- Ανάκτηση πληροφορίας
- Εξατομικευμένη/διαφοροποιημένη αναζήτηση
- Σημασιολογικός ιστός
- Εξόρυξη δεδομένων ιστού
- Συσταδοποίηση
- Επεξεργασία φυσικής γλώσσας

Διακρίσεις

- **Ε.Σ.Π.Α, Υπουργείο Παιδείας - Ευρωπαϊκή Ένωση** (2010-2013)
  Υποτροφία υποστήριξης διδακτορικής έρευνας (Ηράκλειτος ΙΙ)
- **Ίδρυμα Ευγενίδου** (2007-2008)
  Υποτροφία για Μεταπτυχιακές Σπουδές στο Ε.Μ.Π.

Ακαδημαϊκή Εμπειρία

- **Εθνικό Μετσόβιο Πολυτεχνείο**, Ελλάδα (2007-2010)
  Βοηθός Διδασκαλίας
  - Προχωρημένα Θέματα Βάσεων Δεδομένων (Χειμερινό 2009)
  - Προχωρημένα Θέματα Βάσεων Δεδομένων (Χειμερινό 2008)
  - Προχωρημένα Θέματα Βάσεων Δεδομένων (Χειμερινό 2007)

  Επιβλέπων Διπλωματικών Εργασιών
  - *Θ. Μαρούλης*, **Τεχνικές ολοκλήρωσης σημασιολογικών γεωχωρικών δεδομένων** (2013)
  - *Ε. Μπιλίρη*, **Προσαρμογή και αξιολόγηση μεθόδων εξατομίκευσης αναζήτησης με λέξεις κλειδιά σε σημασιολογικά δεδομένα** (2013)
  - *Π. Πάρχας*, **Μελέτη μεθόδων για την έμμεση αύξηση των δεδομένων εκπαίδευσης συναρτήσεων ταξινόμησης σε αποτελέσματα αναζήτησης** (2011)

  Συνεπιβλέπων Διπλωματικών Εργασιών

- *Π. Γεωργίου*, **Ανάκτηση Πληροφορίας Στον Ιστό με Χρήση Ταξινόμησης Όψεων και Συσταδοποίησης** (2009)
- *Α. Κόλλιας*, **Εργαλείο Συλλογής και Οργάνωσης Γνώσης με Μηχανισμούς Μετα-Αναζήτησης στον Ιστό** (2009)
- *Α. Νικολαΐδης*, **Τεχνικές ταξινόμησης αποτελεσμάτων μηχανών αναζήτησης με βάση την ιστορία του χρήστη** (2009)

Έργα

- **ΙΠΣΥ, Ε. Κ. Αθηνά** , Ελλάδα. Έργο GeoKnow (EU - FP7).
  - Έρευνα, ανάπτυξη και διαχείριση έργου πάνω στη διαχείριση γεωχωρικών σημασιολογικών δεδομένων. (1/2013 -)

- **ΙΠΣΥ, Ε. Κ. Αθηνά** , Ελλάδα. Έργο Arcomem (EU - FP7).
  - Έρευνα και ανάπτυξη πάνω στη διαφοροποίηση αναζήτησης δεδομένων από κοινωνικά δίκτυα. (1/2012 - 12/2012)

- **ΙΕΛ, Ε. Κ. Αθηνά**, Ελλάδα. Έργο 'Εκφραση'.
  - Ανάπτυξη εφαρμογής εξαγωγής πολυλεκτικών εκφράσεων από σώματα κειμένων και βάσης δεδομένων για την αποθήκευσή τους. (2/2007 - 10/2008)

- **ΙΠΣΥ, Ε. Κ. Αθηνά - Αλέξανδρος Φλέμινγκ, Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών**, Ελλάδα
  - Ανάπτυξη διαδικτυακής εφαρμογής και βάσης δεδομένων αποθήκευσης, επεξεργασίας και αναζήτησης βιολογικών δεδομένων (microRNA). (5/2008 - 6/2008)

- **Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων, Ε.Μ.Π. - Ελληνική Εταιρία Ύπατος Παγκρέατος και Χοληφόρων**, Ελλάδα
  - Ανάπτυξη διαδικτυακής εφαρμογής και βάσης δεδομένων σταδιοποίησης, επεξεργασίας και αποθήκευσης ιατρικών δεδομένων (υπατικών όγκων). (5/2007 - 11/2007)

Κρίσεις Άρθρων

- Επιτροπή Κρίσης
  - ECIR'14, ECIR'13, CIKM'11(Posters), SIGIR'10
- Εξωτερικός Κριτής

- EDBT, ICDE, TPDL, IJAIT, IJDWM (2013)
- ECIR, TPDL, IS, DKE, WIDM, SETN (2012)
- EDBT, SIGMOD, SIGMOD, CIKM, KAIS, WISE, WEBDB, DOLAP, DESIRE (2011)
- SIGIR, KAIS, DEXA, WISE, ECDL, SETN (2010)
- CIKM, JIS, AIAI, SITIS, ECI, PSI (2009)
- JIR, WIDM, INEWS (2008)

Δημοσιεύσεις

1. Giorgos Giannopoulos, Evmorfia Biliri and Timos Sellis,
Personalizing keyword search on RDF data, in Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'13), Valletta, Malta, September 2013.

2. Nikos Bikakis, Giorgos Giannopoulos, John Liagouris, Dimitrios Skoutas, Theodore Dalamagas and Timos Sellis,
RDivF: Diversifying Keyword Search on RDF Graphs, in Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'13), Valletta, Malta, September 2013.

3. Giorgos Giannopoulos, Ingmar Weber, Alejandro Jaimes and Timos Sellis,
Diversifying User Comments on News Articles, in Proceedings of the 13th International Conference on Web Information System Engineering (WISE'12), Cyprus, November 2012.

4. Giorgos Giannopoulos,
Personalizing Search Results on User Intent, in Proceedings of the VLDB PhD workshop, in conjunction with the 38th International Conference on Very Large Databases (VLDB'12), Turkey, 2012.

5. Giorgos Giannopoulos, Ulf Brefeld, Theodore Dalamagas and Timos Sellis,
Learning to rank user intent, in Proceedings of the 20th ACM International Conference on information and Knowledge Management (CIKM'11), UK, 2011.

6. Giorgos Giannopoulos, Theodore Dalamagas and Timos Sellis,
Search Behavior-Driven Training for Result Re-ranking, in Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'11), Germany, 2011.

7. Nikos Bikakis, Giorgos Giannopoulos, Theodore Dalamagas and Timos Sellis,
Integrating Keywords and Semantics on Document Annotation and Search, in Proceedings of the 9th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE'10), Greece, 2010

8. Giorgos Giannopoulos, Nikos Bikakis, Theodore Dalamagas and Timos Sellis,
GoNTogle: a Tool for Semantic Annotation and Search, in Proceedings of the 7th Extended Semantic Web Conference (ESWC'10) (Demo), Greece, 2010.

9. Manolis Maragkakis, Panagiotis Alexiou, Giorgio L Papadopoulos, Martin Reczko, Theodore Dalamagas, Giorgos Giannopoulos, George Goumas, Evangelos Koukis, Kornilios Kourtis, Victor A Simossis, Praveen Sethupathy, Thanasis Vergoulis, Nectarios Koziris, Timos Sellis, Panagiotis Tsanakas and Artemis G Hatzigeorgiou,
Accurate microRNA target prediction correlates with protein repression levels, BMC Bioinformatics, published on September 18, 2009.

10. Giorgos Giannopoulos, Theodore Dalamagas and Timos Sellis,
Collaborative Ranking Function Training for Web Search Personalization, in Proceedings of the 3rd International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases (PersDB 09) in conjunction with the 35th International Conference on Very Large Data Bases, France, 2009.

11. M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou,
DIANA-microT web server: elucidating microRNA functions through target prediction, Nucleic Acids Res., Advance Access published on April 30, 2009.

12. Giorgos Giannopoulos, Theodore Dalamagas, Magdalini Eirinaki and Timos Sellis,
Boosting the Ranking Function Learning Process using Clustering, in Proceedings of the 10th ACM International Workshop on Web Information and Data Management (WIDM'08) in conjunction with the 17th ACM International Conference on Information and Knowledge Management (CIKM'08), California, USA, 2008.

Εργασιακή Εμπειρία

- **ΙΠΣΥΠ, Ε. Κ. Αθηνά**, Ελλάδα (11/2008 - σήμερα)
Βοηθός έρευνας - Υποτροφία επιμόρφωσης, εξειδίκευσης και επιστημονικής εκπαίδευσης. Τμήμα Κατανεμημένων Πληροφοριακών Συστημάτων και Διαδικτύου.
Ανάπτυξη εφαρμογών σημασιολογικού χαρακτηρισμού και αναζήτησης.
Ανάπτυξη εφαρμογών διαχείρισης βιολογικών δεδομένων και δημοσιεύσεων.
Ανάπτυξη αλγορίθμων και μεθοδολογιών για εξατομίκευση της αναζήτησης.

- **ΙΕΛ, Ε. Κ. Αθηνά**, Ελλάδα (2/2007 - 10/2008)
Μηχανικός λογισμικού - Υποτροφία επιμόρφωσης, εξειδίκευσης και επιστημονικής εκπαίδευσης. Τμήμα Ηλεκτρονικής Λεξικογραφίας και Γλωσσικών Πόρων, έργο Έκφραση.
Ανάπτυξη εφαρμογής εξαγωγής πολυλεκτικών εκφράσεων από σώματα κειμένων με βάση γραμματικούς κανόνες και στατιστικές μεθόδους (Perl).

Ανάπτυξη βάσης δεδομένων και εφαρμογής διαχείρισης και επισημείωσης πολυλεκτικών εκφράσεων (MySQL, Java, NetBeans).
Ανάπτυξη εφαρμογών ολοκλήρωσης των παραπάνω με εργαλεία και τεχνολογίες του Ι.Ε.Λ., όπως το Μορφολογικό Λεξικό, τον Γραμματικό Επισημειωτή και την Οντολογία της Έκφρασης (C++, Protege).

Τεχνικές Ικανότητες

- **Προγραμματισμός**: Perl, C, C#, C++, Java, Hadoop, Hbase, Pig, SQL, HTML, PHP, XML, XSLT, Javascript, Ajax.
- **Άλλα**: MySQL, Semantic Web technologies (RDF(S)/OWL/SPARQL, Protege, RDF Stores)

Ξένες Γλώσσες

- Αγγλικά (Cambridge Proficiency)
- Γερμανικά (Mittelstufenprufung)