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

Μελέτη μεθόδων για την έμμεση αύξηση των δεδομένων εκπαίδευσης συναρτήσεων ταξινόμησης σε αποτελέσματα αναζήτησης

Διπλωματικη Εργασια

του

ΠΑΝΑΓΙΩΤΗ Γ. ΠΑΡΧΑ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Βασεων Γνωσεων και Δεδομενων
Αθήνα, Μάρτης 2011

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Μελέτη μεθόδων για την έμμεση αύξηση των δεδομένων εκπαίδευσης συναρτήσεων ταξινόμησης σε αποτελέσματα αναζήτησης

Διπλωματικη Εργασια

του

ΠΑΝΑΓΙΩΤΗ Γ. ΠΑΡΧΑ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3η Μαρτίου 2011

.....	.....	.....
Τιμολέων Σελλής	Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτης 2011

(Υπογραφή)

.....

Παναγιωτης Παρχας
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Σελλή για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων. Επίσης ευχαριστώ ιδιαίτερα τον Γιώργο Γιαννόπουλο για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε.

Περίληψη

Καθώς ο όγκος πληροφορίας που διακινείται μέσω του διαδικτύου αυξάνεται με εκθετικό ρυθμό, η ανάγκη αποδοτικής αναζήτησης γίνεται όλο και πιο επιβεβλημένη. Η αναζήτηση πλέον διαφοροποιείται ανάλογα με τον χρήστη και τις συνήθειές του. Καθώς ο χρήστης αλλάζει συνεχώς το μοτίβο (pattern) των επιλογών του, δεν μπορεί να κατασκευαστεί ντετερμινιστικός αλγόριθμος που να επιστρέφει την βέλτιστη σειρά αποτελεσμάτων για τον εκάστοτε χρήστη. Μια προσέγγιση για την επίτευξη της προσωποποιημένης αναζήτησης είναι η χρήση νευρωνικών δικτύων. Τα νευρωνικά δίκτυα αποτελούν προσεγγιστικά μοντέλα τα οποία ουσιαστικά ελαχιστοποιούν μια συνάρτηση βάση κάποιων περιορισμών. Για την υλοποίηση τους χρειαζόμαστε κάποιο σύνολο εκμάθησης. Θεωρητικά, όσο μεγαλύτερο και πιο αντιπροσωπευτικό είναι το σύνολο εκμάθησης τόσο καλύτερο είναι το μοντέλο που κατασκευάζεται. Το σύνολο όμως των αποτελεσμάτων που αξιολογούν οι χρήστες σε κάθε αναζήτησή τους είναι μικρό (συνήθως μόνο τα 10 πρώτα αποτελέσματα). Η ιδέα αυτής της διπλωματικής είναι να επεκτείνουμε τις πραγματικές αξιολογήσεις των χρηστών σε αποτελέσματα που είναι συναφή με βάση κάποια συγκεκριμένα χαρακτηριστικά. Υλοποιούμε δηλαδή συσταδοποίηση (clustering) των αποτελεσμάτων, αποφασίζουμε ποιες συστάδες θα κρατήσουμε, ορίζουμε μια ενιαία αξιολόγηση για κάθε συστάδα και την επεκτείνουμε σε όλα τα μέλη του.

Επίσης δοκιμάστηκε μεγάλο σύνολο αλγορίθμων για το clustering ώστε να επιτευχθεί το καλύτερο δυνατό αποτέλεσμα. Αυτά τα νέα σύνολα δόθηκαν σαν είσοδος στο νευρωνικό (SVM) και κατασκευάστηκε το μοντέλο που στη συνέχεια ελέγχθηκε με τη χρήση ενός διαφορετικού συνόλου ελέγχου(dataset) που προσομοιώνει την νέα αναζήτηση του χρήστη. Όλη η διαδικασία αναλύεται με λεπτομέρειες στις σελίδες που ακολουθούν.

Λέξεις Κλειδιά

LETOR, SVM, CLUSTERS, personalization, feature vectors, IR, ομαδοποίηση, νευρωνικό δίκτυο

Abstract

The vast amount of information traveling through the Internet makes the need of efficient web search crucial. The web search gets personalized and the results of the same query differ among users. There have been stated different approaches for the problem of personalization in web search. Our approach is based on the training of neural networks so as to boost the result of a new search, considering the user's options of previous searches. We use the click-through data of the user in order to rank the results and provide the most relevant ones, higher in ordering. The problem we face is that there are limited click-through data per query as the user normally checks only the first n results of the thousands provided by the search engine. In our approach we try to expand the judgment of these first results in the unjudged ones so as to have a bigger training set for our SVM. This expansion is basically done through clustering of the results into relevant groups considering different criteria. In our experiments, we use the LETOR benchmark. In the following pages we elaborate on our method and we explain the results of our experiments.

Keywords

clustering, SVM, personalization, LETOR benchmark, feature vectors, IR

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Έχει ειπωθεί, πολύ εύστοχα, ότι *το Ίντερνετ είναι η μεγαλύτερη βιβλιοθήκη του κόσμου. Το μόνο πρόβλημα, είναι ότι όλα τα βιβλία είναι απλωμένα στο πάτωμα.* Η αναζήτηση είναι συνυφασμένη με το ίδιο το Ίντερνετ. Πράγματι η τρομερή ανάπτυξη του Ίντερνετ στις μέρες μας οφείλεται σε μεγάλο βαθμό στην εκπληκτική βελτίωση των μηχανών αναζήτησης. Τόσο από πλευράς ταχύτητας, όσο και από πλευράς ποιότητας, οι μηχανές αναζήτησης μας λύνουν τα χέρια. Αυτό αφενός διευκολύνεται από την σταδιακή εξοικείωση των χρηστών στον τρόπο αναζήτησης (που είναι βασισμένος στις λέξεις κλειδιά), αφετέρου αντιμετωπίζει το σοβαρό ανάχωμα της ραγδαίας αύξησης του όγκου της πληροφορίας που διακινείται στο διαδίκτυο, πληροφορία σε μεγάλο βαθμό άχρηστη για τους περισσότερους χρήστες. Οι διάσημες μηχανές αναζήτησης καλούνται να γίνουν όλο και πιο αποδοτικές μιας και ο ρόλος τους είναι αναντικατάστατος. Ποίες άραγε σελίδες του ιστού θέλει να επισκεφτεί ο χρήστης πληκτρολογώντας τις συγκεκριμένες λέξεις κλειδιά; Υπάρχουν χιλιάδες σελίδες που περιέχουν αυτές τις λέξεις κλειδιά, παρόλα αυτά ο χρήστης ενδιαφέρεται μόνο για ένα μικρό υποσύνολο. Και αυτό θα επέστρεφε ιδανικά μια μηχανή αναζήτησης.

Από την άλλη πλευρά εμφανίζεται το πρόβλημα της προσωποποίησης της αναζήτησης. Η ίδια λέξη κλειδί ενδέχεται να έχει διαφορετικές ερμηνείες και ο εκάστοτε χρήστης μπορεί να ενδιαφέρεται μόνο για κάποια από αυτές. Για παράδειγμα, η λέξη κλειδί 'eclipse". Για έναν αστρονόμο μια αναζήτηση με αυτή τη λέξη κλειδί, θα υπονοούσε μάλλον το φαινόμενο της έκλειψης ενός ουράνιου σώματος, οπότε τα αποτελέσματα που θα επιθυμούσε θα ήταν πίνακες με ημερομηνίες εκλείψεων, φωτογραφίες και επιστημονικά άρθρα. Από την άλλη πλευρά ένας προγραμματιστής, μάλλον θα έψαχνε για το προγραμματιστικό περιβάλλον *eclipse* οπότε θα επιθυμούσε αποτελέσματα όπως συνδέσμους για να το προμηθευτεί, forums συζητήσεων για επίλυση προβλημάτων κ.α. Είναι προφανές ότι μια βέλτιστη μηχανή, θα επέστρεφε σε κάθε έναν το πραγματικό αντικείμενο της έρευνάς του. Αυτό είναι το πρόβλημα της προσωποποίησης το οποίο συνδέεται άμεσα με τις τεχνικές που προτείνονται σε αυτή τη διπλωματική.

Όπως είναι προφανές, για να ξεπεράσουμε τα παραπάνω εμπόδια, είναι απαραίτητη η συμβολή του χρήστη. Αυτός πρέπει να μας υποδηλώσει πιο είναι το αντικείμενο της έρευνας του, πληροφορία την οποία θα λαμβάνουμε υπόψη μας στις μελλοντικές αναζητήσεις του. Θα μπορούσαμε απλά να ζητήσουμε από τον χρήστη να αξιολογήσει τα αποτελέσματα άμεσα.

Γνωρίζοντας τις προτιμήσεις του, θα μπορούσαμε σε επόμενη αναζήτηση να του παρέχουμε καλύτερα αποτελέσματα βελτιστοποιώντας ή ακόμα προσωποποιώντας την συνάρτηση αξιολόγησης των αποτελεσμάτων. Παρόλα αυτά, έχει αποδειχθεί στην πράξη ότι ο χρήστης σπάνια θα αξιολογήσει το αποτέλεσμα μιας αναζήτησης του. Διάφορες μελέτες [4], αλλά και η εμπειρία από τη χρήση μηχανών αναζήτησης δείχνουν ότι οι χρήστες συνήθως δυσανασχετούν και αγνοούν συστήματα που τους ζητούν 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. Μεγαλύτερο ενδιαφέρον έχει ο *έμμεσος* τρόπος αξιολόγησης μέσω της καταγραφής των *clicktrough data*, δεδομένων δηλαδή όπως το αν ο χρήστης επισκέφτηκε μια συγκεκριμένη σελίδα ή όχι, αν επέστρεψε άμεσα με τη χρήση του back κτλ. Οι μηχανές αναζήτησης εγκαθιστούν έναν proxy server από όπου πρώτα περνούν τα clicks των χρηστών πριν επανακατευθυνθούν στην επιθυμητή διεύθυνση. Έτσι καταγράφονται οι προτιμήσεις του συγκεκριμένου χρήστη (ή της ομάδας στην οποία ανήκει) και λαμβάνονται υπόψη σε μελλοντικές αναζητήσεις.

1.1 Αντικείμενο της διπλωματικής

Το βασικό ζήτημα που προκύπτει στην *εξατομικευμένη αναζήτηση* είναι το πως θα επεκταθούν τα *δεδομένα προτίμησης* (που έχουν καταγραφεί σε προηγούμενες αναζητήσεις του χρήστη), στην νέα και πιθανότατα εντελώς διαφορετική (από πλευράς περιεχομένου) αναζήτηση. Μια λύση που έχει προταθεί [4], είναι η χρήση νευρωνικών δικτύων. Μπορούμε να εκπαιδεύσουμε ένα νευρωνικό δίκτυο με τα δεδομένα προτίμησης του χρήστη και σε κάθε επόμενη αναζήτηση, τα αποτελέσματα α) να επαναξιολογούνται (reranking) με βάση το μοντέλο του νευρωνικού και β) αφού αξιολογηθούν, να χρησιμοποιούνται στην εκ νέου εκπαίδευση του μοντέλου. Αυτή η μέθοδος αντιμετωπίζει το βασικό πρόβλημα της έλλειψης δεδομένων εκπαίδευσης. Έχει αποδειχθεί στην πράξη ότι ο μέσος χρήστης εξετάζει (και άρα έμμεσα αξιολογεί) μόνο τα 10-20 πρώτα αποτελέσματα κάθε αναζήτησης, ανεξάρτητα από το πόσα βρέθηκαν με βάση τις λέξεις κλειδιά που χρησιμοποίησε. Έπειτα προσπαθεί να κάνει πιο συγκεκριμένη την αναζήτησή του, αλλάζοντας τις λέξεις κλειδιά, ή προσθέτοντας κάποιες επιπλέον. 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 καταναλώνει χρόνο και υπολογιστική ισχύ στην εκπαίδευση του μοντέλου και όχι στην επαναταξινόμηση των αποτελεσμάτων της αναζήτησης. Για να κατασκευαστεί ένα αξιόπιστο μοντέλο, πρέπει να μεσολαβήσουν πολλές αναζητήσεις ώστε να υπάρχουν αρκετά δεδομένα εκπαίδευσης. Αυτό προφανώς δεν είναι καθόλου αποδοτικό

και πρέπει να ξεπεραστεί

2. Στην περίπτωση που συμμετέχουν πολλοί χρήστες, μαζεύονται 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 αποτέλεσμα, οι εκπαιδευμένες συναρτήσεις ταξινόμησης προσπαθούν να ικανοποιήσουν όλες τις διαφορετικές συμπεριφορές αναζήτησης, καταλήγοντας τελικά, να είναι μεν γενικευμένες, αλλά λιγότερο ακριβείς.

Αντικείμενο της διπλωματικής είναι ο εμπλουτισμός του συνόλου εκπαίδευσης προσθέτοντας στα ήδη αξιολογημένα από τον χρήστη αποτελέσματα, αυτά που δεν έχει αξιολογήσει και πιθανώς δεν έχει δει καν. Για να το καταφέρουμε αυτό αρχικά συσταδοποιούμε τα αποτελέσματα με βάση την ομοιότητα τους στη λεκτική περιγραφή. Στη συνέχεια εντοπίζουμε αποτελέσματα-κλειδιά τα οποία φέρουν την έμμεση αξιολόγηση του χρήστη (ανήκουν δηλαδή στο σύνολο αυτών που έχει αξιολογήσει) και προσπαθούμε να επεκτείνουμε την αξιολόγηση και στα υπόλοιπα μέλη της συστάδας. Κατ΄ αυτόν τον τρόπο αυξάνουμε κατά πολύ το σύνολο εκπαίδευσης το οποίο πλέον αποτελείται από την ένωση των συνόλων αξιολόγησής και επέκτασης. Αυτή η τεχνική είναι πλήρως αυτοματοποιημένη και δεν απαιτεί την παρέμβαση του χρήστη. Το νέο σύνολο εκπαίδευσης χρησιμοποιείται στην κατασκευή της συνάρτησης ταξινόμησης. Σε αυτό το πλαίσιο παρουσιάζονται αναλυτικά τα εργαλεία που χρησιμοποιήθηκαν, τα πειράματα που διεξήχθησαν και τα αποτελέσματα που προέκυψαν. Αποδεικνύεται ότι το αποτέλεσμα της μεθόδου της επέκτασης προσεγγίζει αρκετά το ιδανικό αποτέλεσμα της εκπαίδευσης που θα είχαμε χρησιμοποιώντας το πλήρες σύνολο των δεδομένων στην εκπαίδευση του νευρωνικού, γεγονός μάλλον ανέφικτο καθώς θα απαιτούσε την αξιολόγηση όλων των αποτελεσμάτων προηγούμενων αναζητήσεων από τον χρήστη.

1.2 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται στα εξής κεντρικά στοιχεία:

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

2. Υλοποιούμε τις προτεινόμενες μεθόδους, προσαρμόζοντας τες σε μία ευρέως διαδεδομένη τεχνική μηχανικής μάθησης, τις Μηχανές Διανυσμάτων Στήριξης. Η υλοποίηση αυτή συνίσταται σε μία σειρά από αυτόνομες φάσεις που καθιστούν την προτεινόμενη λύση αρκετά γενική, ώστε να μπορεί να υιοθετηθεί από διαφορετικά συστήματα και προσεγγίσεις (για παράδειγμα, διαφορετικές τεχνικές μηχανικής μάθησης). Πέρα από την παραπάνω τεχνική μηχανικής μάθησης, σημαντικό ρόλο στις μεθόδους μας παίζουν αλγόριθμοι συσταδοποίησης (clustering), οι οποίοι επιτρέπουν την αναγνώριση και εκμετάλλευση ομοιογενών (ως προς διάφορα κριτήρια όπως θα αναλυθεί σε επόμενες ενότητες) ομάδων δεδομένων.

3. Εκτελούμε πειράματα με πολλούς διαφορετικούς συνδυασμούς παραμέτρων συσταδοποίησης συνεισφέροντας στον τομέα της απόδοσης των αλγορίθμων συσταδοποίησης αφού με αυτόν τον τρόπο ελέγχονται τα αποτελέσματα και αναδεικνύονται κάποια πλεονεκτήματα και μειονεκτήματά τους

4. Εκτελούμε πειράματα σε διαφορετικά σετ δεδομένων τα οποία δείχνουν την αποδοτικότητα των προτεινόμενων μεθόδων.

5. Εξάγουμε ενδιαφέροντα αποτελέσματα και δείχνουμε οτι η μέθοδος που προτείνουμε τείνει να πλησιάσει την ιδανική και ανέφικτη περίπτωση εκπαίδευσης με *όλα* τα αποτελέσματα προηγούμενων αναζητήσεων

1.3 Οργάνωση του τόμου

Ο τόμος οργανώνεται σε 6 κεφάλαια ως εξης: Στο κεφάλαιο 2 δίνεται η θεωρητική βάση της υλοποίησης στην οποία παρουσιάζονται βασικές έννοιες όπως η έννοια της συσταδοποίησης και κάποια βασικά για τα νευρωνικά δίκτυα, καθώς και τα εργαλεία που χρησιμοποιήσαμε κατά την υλοποίηση. Στη συνέχεια, στο κεφάλαιο 3, παρουσιάζεται αναλυτικά η μέθοδος που προτείνουμε, ενώ στο επόμενο κεφάλαιο δίνεται το πλαίσιο εργασίας και τα πειραματικά δεδομένα εισόδου που χρησιμοποιήσαμε για να ελέγξουμε την αποδοτικότητα της μεθόδου. Έπειτα, στο κεφάλαιο 5 παρουσιάζονται αναλυτικά τα αποτελέσματα των πειραμάτων και σχολιάζονται τα συμπεράσματα που προκύπτουν. Τέλος στο κεφάλαιο 6, δίνεται ο επίλογος, με τα ολικά συμπεράσματα και τις μελλοντικές επεκτάσεις.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά τα εργαλεία που χρησιμοποιήθηκαν για την διεξαγωγή των πειραμάτων και το θεωρητικό υπόβαθρο που συνοδεύει τις εκάστοτε υλοποιήσεις. Στα εργαλεία αυτά ανήκουν το εργαλείο συσταδοποίησης (clustering) Cluto και το πρόγραμμα εκπαίδευσης Μηχανών Διανυσμάτων Υποστήριξης (SVM).

2.1 Cluto

2.1.1 Τι είναι το Cluto;

Το Cluto είναι ένα πρόγραμμα συσταδοποίησης συνόλων δεδομένων. Παρέχει τρεις διαφορετικές ομάδες από αλγόριθμους συσταδοποίησης οι οποίοι ενεργούν είτε απευθείας στις ιδιότητες του αντικειμένου, είτε στο χώρο ομοιότητάς του αντικειμένου. Το βασικό χαρακτηριστικό του Cluto είναι ότι αντιμετωπίζει το πρόβλημα της συσταδοποίησης ως πρόβλημα βελτιστοποίησης, μεγιστοποίησης ή ελαχιστοποίησης δηλαδή μιας συγκεκριμένης συνάρτησης -κριτήριο η οποία ορίζεται είτε τοπικά είτε γενικά, στο χώρο επίλυσης του προβλήματος. Συνολικά το πακέτο περιέχει 7 διαφορετικές συναρτήσεις -κριτήρια που μπορούν να χρησιμοποιηθούν σε συνδυασμό με τους διαφορετικούς αλγορίθμους συσταδοποίησης. Εμείς, λαμβάνοντας υπόψη τα αποτελέσματα της έρευνάς [7] περιοριστήκαμε σε δυο κριτήρια, όπως φαίνεται αναλυτικά πιο κάτω. Μια σημαντική παράμετρος των αλγορίθμων συσταδοποίησης που χρησιμοποιούνται είναι η μέθοδος που χρησιμοποιείται για να βελτιστοποιηθεί αυτή η συνάρτηση-κριτήριο. Το Cluto χρησιμοποιεί έναν άπληστο στη φύση του αλγόριθμο που έχει μικρές υπολογιστικές απαιτήσεις και έχει αποδειχθεί ότι παράγει συστάδες υψηλής ποιότητας [7]. Τέλος οι αλγόριθμοι που χρησιμοποιούνται έχουν βελτιστοποιηθεί ώστε να μπορούν να χειρίζονται αποδοτικά μεγάλα σύνολα δεδομένων (μερικές δεκάδες χιλιάδες) τα οποία περιλαμβάνουν διανύσματα σε χώρους μεγάλων διαστάσεων (μερικές χιλιάδες). Επίσης λαμβάνεται υπόψη το γεγονός ότι οι περισσότεροι πίνακες είναι αραιοί και έχουν υλοποιηθεί αποδοτικοί τρόποι διαχείρισης της μνήμης, ώστε τελικά η μνήμη που χρειάζεται να είναι γραμμική ως προς την είσοδο.

2.1.2 Γιατί επιλέξαμε το Cluto;

Στο σύνολο των εργαλείων συσταδοποίησης που εξετάστηκε, επιλέξαμε αυτό το πρόγραμμα, κυρίως για τη μεγάλο εύρος δυνατοτήτων που μας παρέχει. Προσφέρει μεγάλη γκάμα αλγορίθμων και μεθόδων συσταδοποίησης ενώ παράλληλα το αποτέλεσμα που προσφέρει είναι υψηλής ποιότητας σε μικρό χρόνο. Παρόλα αυτά, αντιμετωπίσαμε κάποια προβλήματα κατά την υλοποίηση, όπως την αδυναμία ένταξης αντικειμένων σε κάποια συστάδα, αλλά ευτυχώς ήταν σπάνια και τα διαχειριστήκαμε εύκολα. Επίσης η κοινότητα στήριξης βοήθησε αρκετά, με τον εντοπισμό σφαλμάτων και την επίλυσή τους στην ασταθή ακόμα καινούρια έκδοση (2.1.2beta).

2.1.3 Παρουσίαση των παραμέτρων που χρησιμοποιήθηκαν

Παρουσιάζουμε εδώ συνοπτικά τους αλγορίθμους και τις συναρτήσεις -κριτήρια που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων συσταδοποίησης. Αναλύονται συνοπτικά κάποια βασικά χαρακτηριστικά τους.

Αλγόριθμοι

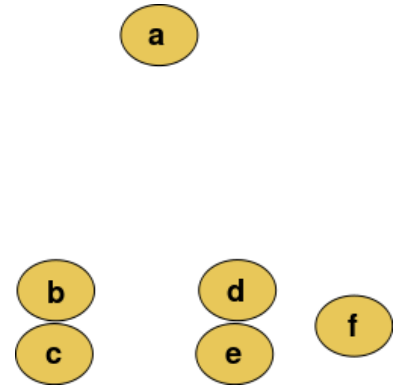
rb Σε αυτή τη μέθοδο συσταδοποίησης k συστάδων, χρησιμοποιήθηκαν *k-1* επαναλαμβανόμενες διχοτομήσεις (repeted bisections). Με αυτή την προσέγγιση ο πίνακας εισόδου χωρίζεται σε δύο συστάδες και στη συνέχεια επιλέγεται μια από αυτές η οποία ξανά διαχωρίζεται, μέχρι να κατασκευαστεί ο απαιτούμενος αριθμός συστάδων. Σε κάθε ένα από τα παραπάνω βήματα, ο διαχωρισμός γίνεται ώστε να βελτιστοποιηθεί μια συγκεκριμένη, δοθείσα συνάρτηση-κριτήριο. Με αυτή τη μέθοδο υπολογίζονται τα τοπικά ελάχιστα της συνάρτησης, αλλά γενικά,όχι τα ολικά.

rbr Αυτή η μέθοδος ακολουθεί γενικά τον ίδιο αλγόριθμο με την rb αλλά στο τέλος αναζητά το ολικό ελάχιστο. Ουσιαστικά χρησιμοποιεί το αποτέλεσμα της rb και προσπαθεί να βελτιστοποιήσει περαιτέρω την συνάρτηση-κριτήριο.

direct Σε αυτή τη μέθοδο, η λύση υπολογίζεται με την ταυτόχρονη εύρεση των k συστάδων. Γενικά, αυτή η μέθοδος είναι πιο αργή από τις παραπάνω. Από την άλλη πλευρά, για το πλήθος συστάδων που χειριστήκαμε, (λιγότερες από 20), η ποιότητα των αποτελεσμάτων είναι καλύτερη.

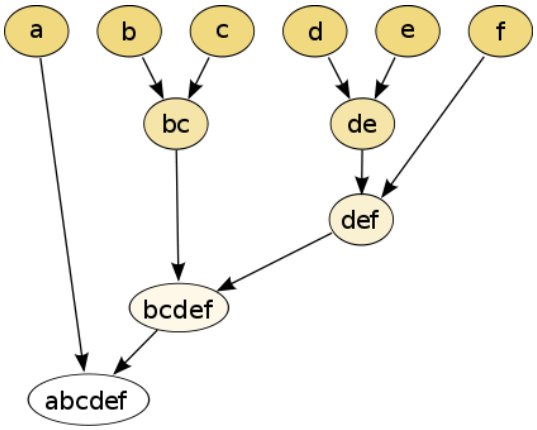
agglo Σε αυτή τη μέθοδο, η συσταδοποίηση γίνεται από κάτω προς τα πάνω, σχηματίζονται δηλαδή στην αρχή, τόσες συστάδες όσα και τα αντικείμενα, και στη συνέχεια συγχωνεύονται με βάση τη βελτιστοποίηση της συνάρτησης- κριτήριο μέχρι να φτάσουμε στο επιθυμητό πλήθος συστάδων. Παρακάτω παρουσιάζεται ένα παράδειγμα για αυτή τη μέθοδο, όπου η συνάρτηση-κριτήριο ταυτίζεται με την ευκλείδεια απόσταση:

Το Σχήμα 2.1 φαίνονται τα δεδομένα εισόδου, προς συσταδοποίηση. Η συνάρτηση που πρέπει να ελαχιστοποιηθεί είναι αυτή της Ευκλείδειας απόστασης μεταξύ των clusters.



Σχήμα 2.1: Παράδειγμα agglomerative clustering Δεδομένα προς συσταδοποίηση.

Το Σχήμα 2.2 παρουσιάζεται το αποτέλεσμα της συσταδοποίησης με βάση τον agglomerative αλγόριθμο. Ξεκινώντας από πάνω, στην πρώτη σειρά υπάρχουν 6 συστάδες (όσες και τα αντικείμενα του προβλήματος) και στη συνέχεια οι συστάδες ενοποιούνται με βάση την απόστασή τους. Οπότε σχηματίζονται τέσσερις, τρεις, δύο και εν τέλει μια μόνο συστάδα που τα περιέχει όλα.



Σχήμα 2.2: Παράδειγμα agglomerative clustering Αποτέλεσμα συσταδοποίησης.

graph Σε αυτή τη μέθοδο, η λύση υπολογίζεται αφού πρώτα κατασκευαστεί ο γράφος γειτνίασης των αντικειμένων (κάθε αντικείμενο μετατρέπεται σε διάνυσμα που συνδέεται με τα πιο κοντινά του διανύσματα) ο οποίος στη συνέχεια χωρίζεται σε συνεκτικούς υπογράφους χρησιμοποιώντας τον min-cut αλγόριθμο.

bagglo Αυτή η μέθοδος λειτουργεί παρόμοια με την μέθοδο agglo αφού όμως πρώτα έχει υπολογιστεί μια λύση με βάση τον rb αλγόριθμο. Συγκεκριμένα, αρχικά φτιάχνεται μια λύση για $\sqrt{n}$ συστάδες μέσω της μεθόδου των επαναλαμβανόμενων διχοτομήσεων όπου n το αρχικό πλήθος αντικειμένων. Έπειτα ο διανυσματικός χώρος του προβλήματος εμπλουτίζεται με την προσθήκη μιας επιπλέον διάστασης για κάθε συστάδα η οποία περιέχει την απόσταση του αντικειμένου από το κεντροειδές της συστάδας. Έπειτα εφαρμόζεται για τα *τροποποιημένα* αντικείμενα ο κλασσικός agglomerative αλγόριθμος.

Πρόκειται ουσιαστικά για *ελεγχόμενη (biased)* έκδοση του agglo αλγορίθμου, εξού και η ονομασία bagglo

Συνάρτηση ομοιότητας

Σε αντίθεση με τις μεθόδους συσταδοποίησης που αναλύθηκαν πιο πάνω, στην παρούσα διπλωματική χρησιμοποιήθηκε μόνο μια συνάρτηση ομοιότητας, η *ομοιότητα συνημιτόνου* (cosine similarity)

Η συνάρτηση αυτή βασίζεται στο εσωτερικό γινόμενο διανυσμάτων και ορίζεται ως εξής:

$$cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

όπου στην περίπτωσή μας τα διανύσματα A, B έχουν διαστάσεις είτε στον χώρο λεκτικής απεικόνισης, είτε στο χώρο των ιδιοτήτων (features) των αποτελεσμάτων.

Συνάρτηση-κριτήριο

Αφού πρώτα μελετήσαμε τα αποτελέσματα σχετικής έρευνας [7] για την καταλληλότητα των κριτηρίων ανάλογα με το είδος του πειράματος, καταλήξαμε να χρησιμοποιήσουμε τα παρακάτω κριτήρια:

1.
$$E1 : minimize \sum_{i=1}^{k} (n_i \frac{\sum_{v \in S_i, u \in S} (sim(u, v))}{\sqrt{\sum_{v,u \in S_i} (sim(u, v))}})$$

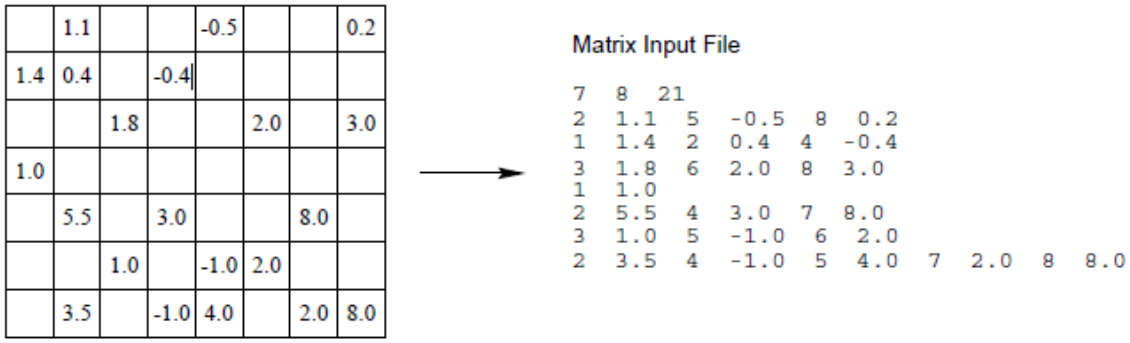
2.
$$H2 : maximize \frac{\sum_{i=1}^{k} (\sqrt{\sum_{u,v \in S_i} (sim(u, v))})}{E1}$$

όπου k, το πλήθος των clusters, S το σύνολο των αντικειμένων προς συσταδοποίηση, S_i το σύνολο των αντικειμένων που ανήκουν στην i συστάδα με πληθάριθμο n_i

Στα πειράματα που κάναμε δοκιμάσαμε τη συσταδοποίηση με όλους τους πιθανούς συνδυασμούς των παραπάνω παραμέτρων με σκοπό, την κατά το δυνατόν βέλτιστη συσταδοποίηση ανάλογα με την περίπτωση.

2.1.4 Χρήση του Cluto

Αφού έχουμε επιλέξει τις παραμέτρους, δίνουμε σαν είσοδο στο πρόγραμμα έναν πίνακα ο οποίος περιέχει την απεικόνιση των αντικειμένων μας σε κάποιο προκαθορισμένο χώρο. Ο χώρος αυτός μπορεί να είναι είτε ο χώρος της λεκτικής περιγραφής, είτε ο χώρος των ιδιοτήτων. Τα αντικείμενά μας είναι τα ζευγάρια ερωτημάτων-αποτελεσμάτων τα οποία εκφράζονται μέσω των διαστάσεών τους στον εκάστοτε χώρο. Κατασκευάζεται δηλαδή ένας πίνακας για κάθε query στον οποίο κάθε σειρά είναι ένα αποτέλεσμα και κάθε στήλη μια διάσταση στον χώρο διαστάσεων. Παρακάτω δίνεται ένα παράδειγμα:



	1.1			-0.5			0.2
1.4	0.4		-0.4				
		1.8			2.0		3.0
1.0							
	5.5		3.0			8.0	
		1.0		-1.0	2.0		
	3.5		-1.0	4.0		2.0	8.0

Σχήμα 2.3: Είσοδος στο cluto

Στο Σχήμα 2.3 φαίνεται η υλοποίηση αραιού πίνακα την οποία χρησιμοποιήσαμε και εμείς. Πιο συγκεκριμένα, το αρχείο εισόδου περιέχει μια γραμμή επιπλέον από τον πίνακα, την πρώτη. Σε αυτή, καταγράφονται 3 αριθμοί: το πλήθος των σειρών(αντικείμενα), το πλήθος των στηλών(διαστάσεις) και το πλήθος των μη μηδενικών στοιχείων. Έπειτα κάθε σειρά αντιστοιχίζεται 1:1 με τον πίνακα και περιέχει ζευγάρια αριθμών όπου ο πρώτος αριθμός υποδηλώνει τον αριθμό της στήλης και ο δεύτερος την τιμή της.

Η έξοδος είναι ένα αρχείο με τόσες σειρές όσες το πλήθος των αντικειμένων που σε κάθε σειρά υπάρχει ένας και μόνο αριθμός ο οποίος συμβολίζει σε ποια συστάδα ανήκει το κάθε αντικείμενο.

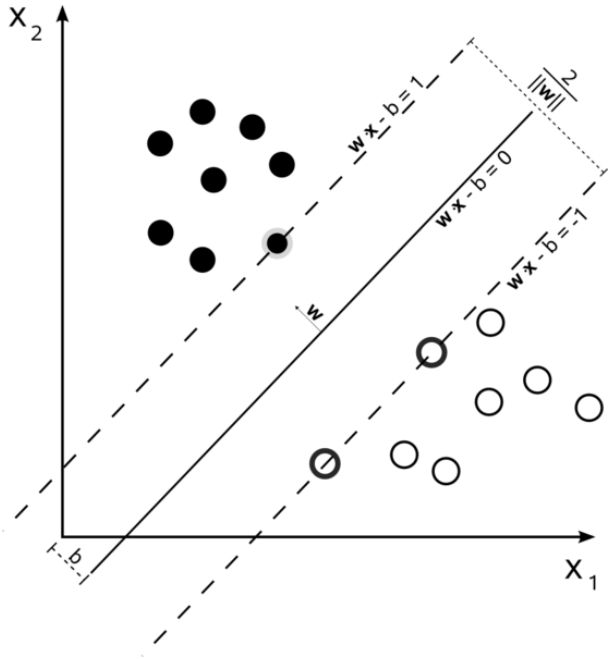
2.2 Μηχανές Διανυσμάτων Υποστήριξης

Ένα βασικό κομμάτι της ιδέας που υλοποιούμε σε αυτή τη διπλωματική, είναι η εκπαίδευση νευρωνικών δικτύων με σκοπό την κατηγοριοποίηση των αποτελεσμάτων προηγούμενων αναζητήσεων ώστε το μοντέλο που θα παραχθεί να μπορεί να κατηγοριοποιήσει τα αποτελέσματα των νέων αναζητήσεων. Για αυτή τη δουλειά, χρησιμοποιούμε *μηχανές διανυσμάτων υποστήριξης (Support Vector Machines-SVMs)* οι οποίες εισήχθησαν από τον V. Vapnik et al.[2] και είναι πολύ καλά θεωρητικά θεμελιωμένες. Στόχος μας είναι η κατηγοριοποίηση των αποτελεσμάτων αναζητήσεων σε τρεις βαθμίδες σχετικότητας (*0,1,2*) αναφορικά με το ερώτημα: **0-άσχετο αποτέλεσμα, 1-σχετικό αποτέλεσμα, 2-πολύ σχετικό αποτέλεσμα.** Κάθε αποτέλεσμα μπορεί να ανήκει μόνο σε μια από τις παραπάνω κατηγορίες. Χρησιμοποιώντας μηχανική μάθηση, ο στόχος μας είναι να εκπαιδεύσουμε κατηγοριοποιητές (classifiers) με βάση τα παραδείγματα για τα οποία έχουμε αποτελέσματα που έχουν αξιολογηθεί είτε από πραγματικούς χρήστες είτε από τη μέθοδο επέκτασης αξιολογήσεων μέσω συσταδοποίησης που έχει αναφερθεί στην εισαγωγή και εξηγείται με σαφήνεια παρακάτω. Πρόκειται ουσιαστικά για *πρόβλημα εκπαίδευσης με επιτήρηση* (supervised learning problem).

Τα SVMs βασίζονται στην αρχή της *μείωσης του δομικού κινδύνου (Structural Risk Minimization)* από τη θεωρία υπολογισμού. Η ιδέα αυτής της αρχής, είναι να βρεθεί μια

υπόθεση h για την οποία μπορεί να εγγυηθεί το μικρότερο πραγματικό σφάλμα. Το πραγματικό σφάλμα της h είναι η πιθανότητα ότι η υπόθεση h θα κάνει λάθος στην κατηγοριοποίηση ενός τυχαίου και άγνωστου δοκιμαστικού παραδείγματος. Μπορεί να βρεθεί ένα άνω όριο που συνδέει το πραγματικό σφάλμα της υπόθεσης h με το σφάλμα της h στο σύνολο εκπαίδευσης. Τα SVMs προσεγγίζουν αυτό το όριο.

Όπως φαίνεται και στο παρακάτω σχήμα 2.4 σκοπός των μηχανών διανυσμάτων υποστήριξης είναι η εύρεση του βέλτιστου υπερεπιπέδου που χωρίζει τα δεδομένα της εκπαίδευσης. Ουσιαστικά, πρόκειται για πρόβλημα βελτιστοποίησης μιας ορισμένης συνάρτησης με βάση περιορισμούς που ανακύπτουν κατά την εκπαίδευση (την είσοδο δηλαδή δεδομένων πάνω στα οποία θα στηριχτεί η κατασκευή του μοντέλου).



Σχήμα 2.4: Μηχανή διανυσμάτων υποστήριξης.

Πιο συγκεκριμένα, το πρόβλημα ανάγεται στην κατασκευή του υπερεπιπέδου που χαρακτηρίζεται από την εξίσωση:

$$\mathbf{w}^T\mathbf{x} + b = 0$$

τα δεδομένα εκπαίδευσης είναι ζευγάρια της μορφής

$$\mathcal{I} = (\mathbf{x}_i, d_i), i = 1..N$$

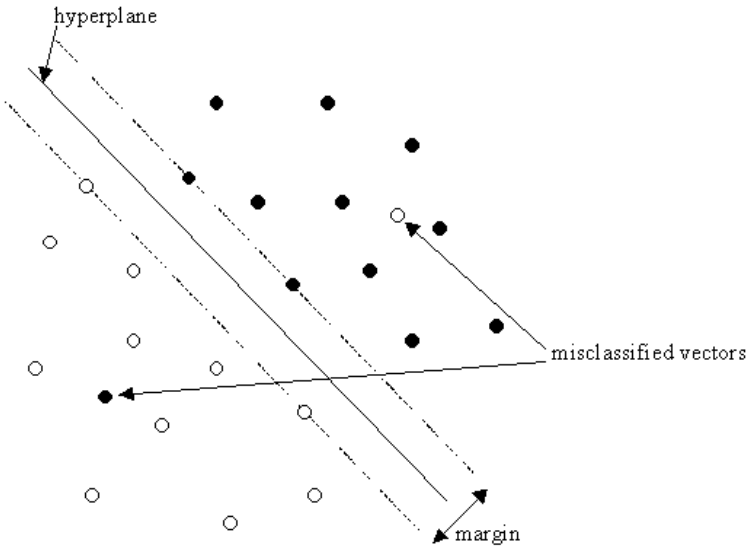
όπου x_i το διάνυσμα i και d_i η αντίστοιχη αξιολόγηση που μπορεί να είναι είτε $+1$ (θετικό ημιεπίπεδο) είτε -1 (αρνητικό ημιεπίπεδο) και πρέπει να ικανοποιούν την παρακάτω σχέση:

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

για τις τιμές εκείνες του διανύσματος βαρών $\mathbf{w}$ που ελαχιστοποιούν τη συνάρτηση κόστους

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Τα SVM ανταποκρίνονται πολύ καλά και σε περιπτώσεις όπου ο διαχωρισμός είναι αδύνατος καθώς υπάρχουν μεμονωμένα δεδομένα εκπαίδευσης που ενώ θα έπρεπε να ανήκουν στο ένα υπερεπίπεδο, πέφτουν στο άλλο. Ένα παράδειγμα αυτής της περίπτωσης φαίνεται στο παρακάτω σχήμα 2.5:



Σχήμα 2.5: Μηχανή διανυσμάτων υποστήριξης με σφάλμα εκπαίδευσης.

σε αυτήν την περίπτωση εισάγονται και οι μεταβλητές χαλάρωσης ξ_i που είναι μη αρνητικοί αριθμοί, και ο ορισμός του υπερεπιπέδου διαχωρισμού γίνεται:

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \text{ για ι= 1..N και } \xi_i \geq 0$$

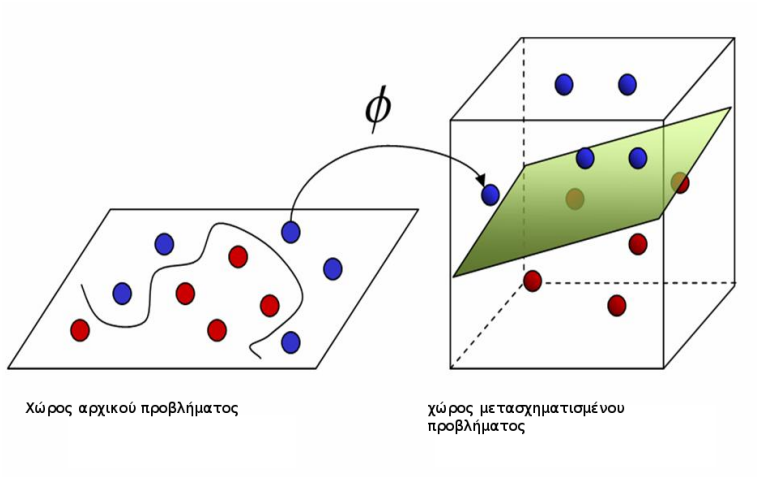
οπότε η επίλυση του προβλήματος ανάγεται στον υπολογισμό των

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}(\xi_i)$$

όπου η παράμετρος C ορίζεται από το χρήστη και καθορίζει το ισοζύγιο ανάμεσα στην πολυπλοκότητα του μοντέλου και τον αριθμό των λανθασμένων σημείων στο στάδιο της εκπαίδευσης. Η επιλογή της παραμέτρου μπορεί να γίνει :

- είτε *πειραματικά*, με δοκιμές πάνω σε ένα σύνολο δεδομένων επαλήθευσης
- είτε *αναλυτικά* χρησιμοποιώντας τα κατάλληλα μαθηματικά μοντέλα.

Κατά αντιστοιχία, για την επίλυση μη γραμμικά διαχωρίσιμων προβλημάτων χρησιμοποιούμε την τεχνική προβολής του δεδομένου προβλήματος σε έναν χώρο (περισσοτέρων διαστάσεων), στον οποίο η λύση είναι γραμμική (σχήμα 2.6). Με αυτόν τον τρόπο τα SVMs αποτελούν πολύ καλή προσέγγιση του προβλήματος της μηχανικής μάθησης.



Σχήμα 2.6: Μηχανή διανυσμάτων υποστήριξης για μη γραμμικά προβλήματα.

2.2.1 Τί είναι το SVM-Rank;

Για την κατασκευή του νευρωνικού μοντέλου που χρειαζόμαστε χρησιμοποιούμε την υλοποίηση του Thorsten Joachims, *SVM-Rank*. Πρόκειται για μια αποδοτική υλοποίηση κατασκευασμένη και προσαρμοσμένη στο σύνολο δεδομένων που επεξεργαζόμαστε. Η παραμετροποίηση που δέχεται το SVMrank έχει να κάνει με την επιλογή της παραμέτρου C η οποία σχετίζεται με την εξισορρόπηση ανάμεσα στην πολυπλοκότητα και την ακρίβεια του μοντέλου. Το benchmark που χρησιμοποιήσαμε απαιτεί τις ακόλουθες τιμές για αυτή 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}. Όσο μεγαλύτερη η τιμή της παραμέτρου, τόσο πιο πολύπλοκο απαιτούμε να είναι το μοντέλο, με κίνδυνο ορισμένες φόρες να μην συγκλίνει ποτέ στη λύση.

Κεφάλαιο 3

Περιγραφή Της Μεθόδου

Στο κεφάλαιο αυτό παρουσιάζουμε αναλυτικά τη μέθοδό μας με τις διάφορες προσεγγίσεις και παραλλαγές που υλοποιήσαμε. Αρχικά, δίνουμε μια γενική εικόνα της μεθόδου και στη συνέχεια εξετάζουμε αναλυτικότερα τα επιμέρους βήματα.

3.1 Διάγραμμα μεθόδου

Ακολουθεί το διάγραμμα ροής της μεθόδου μας. Πρόκειται για συνοπτική παρουσίαση της ροής των δεδομένων μέχρι την εξαγωγή των επιθυμητών στατιστικών (σχήμα 3.1).

Η μέθοδος συνοπτικά ακολουθεί τα εξής βήματα:

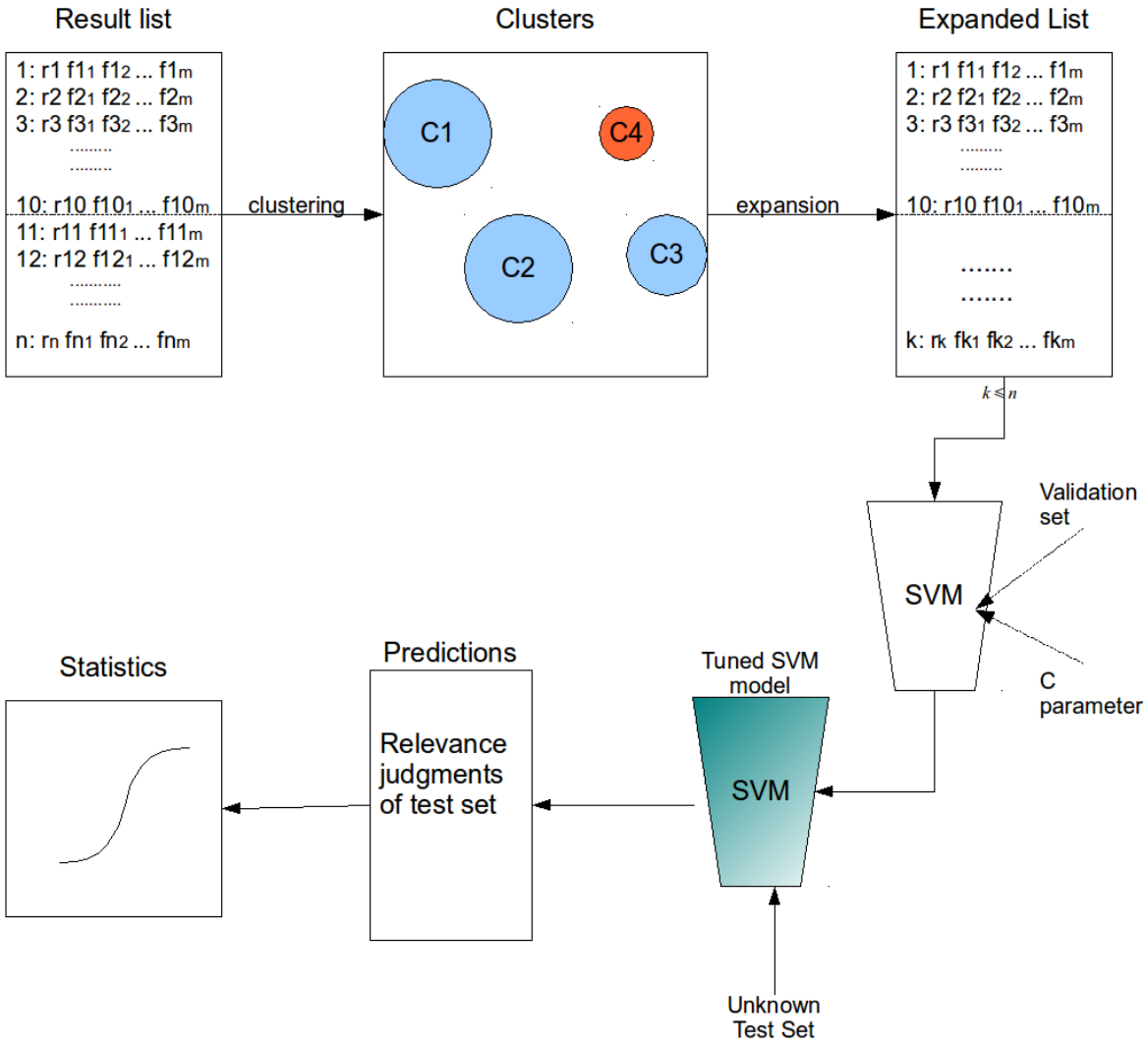
1. *Προεπεξεργασία* των αποτελεσμάτων κάθε αναζήτησης. Από κάθε αποτέλεσμα, αντλούμε τη χρήσιμη πληροφορία που θα επεξεργαστούμε στη συνέχεια.

2. Διατήρηση των πρώτων n αποτελεσμάτων κάθε αναζήτησης. Τα αποτελέσματα ταξινομούνται και διατηρούνται τα πρώτα n. Οι αξιολογήσεις αυτών των αποτελεσμάτων θα θεωρηθούν ως *έγκυρες αξιολογήσεις του χρήστη*. Έχουμε πειραματιστεί με διαφορετικές τιμές του n.

3. *Συσταδοποίηση* των αποτελεσμάτων. Σε αυτό το στάδιο ομαδοποιούμε τα αποτελέσματα με βάση κάποια συγκεκριμένα κριτήρια ομοιότητας. Δοκιμάζουμε διάφορους αλγόριθμους συσταδοποίησης. Το αποτέλεσμα αυτού του βήματος είναι συστάδες *παρόμοιων* αποτελεσμάτων, για κάθε αναζήτηση.

4. *Επέκταση* αξιολόγησης. Με βάση τις συστάδες που κατασκευάσαμε στο προηγούμενο βήμα, προσπαθούμε να επεκτείνουμε τις αρχικές *έγκυρες αξιολογήσεις χρήστη* στα υπόλοιπα αποτελέσματα. Σκοπός μας είναι τα αποτελέσματα που ανήκουν στην ίδια συστάδα να έχουν την ίδια γενικευμένη αξιολόγηση.

5. *Εκπαίδευση* μοντέλου μηχανικής μάθησης. Εκπαιδεύουμε πολλά μοντέλα με διαφορετικές παραμέτρους εκπαίδευσης πάνω στο ίδιο σύνολο εκπαίδευσης που έχει προκύψει



Σχήμα 3.1: Διάγραμμα μεθόδου.

από το προηγούμενο βήμα. Στη συνέχεια δίνουμε σαν είσοδο το *σύνολο επικύρωσης* και κρατάμε το μοντέλο που ανταποκρίνεται καλύτερα.

6. *Έλεγχος* μοντέλου. Εισάγουμε στη συνάρτηση ταξινόμησης το σύνολο ελέγχου και καταγράφουμε τα στατιστικά.

Στη συνέχεια ακολουθεί λεπτομερής ανάλυση των παραπάνω βημάτων σε συνδυασμό με λεπτομερή παρουσίαση των διαφορετικών υλοποιήσεων.

3.2 Προεπεξεργασία των αποτελεσμάτων

Το σύνολο δεδομένων μας αποτελείται από δυο μεγάλες κατηγορίες δεδομένων. Το αρχείο **εγγράφων**, δηλαδή τη συλλογή των κειμένων χωρίς μορφοποίηση των αποτελεσμάτων αναζητήσεων, και το αρχείο των **αξιολογήσεων**. Το αρχείο εγγράφων περιλαμβάνει

στοιχεία για κάθε αποτέλεσμα όπως κωδικό -κλειδί αποτελέσματος, τίτλο, σώμα κειμένου και συγγραφέα. Το αρχείο αξιολογήσεων είναι αυτό που κάνει τη σύνδεση ανάμεσα σε μια αναζήτηση (που χαρακτηρίζεται από τον μοναδικό κωδικό αριθμό της) και σε ένα κωδικό εγγράφου-αποτελέσματος. Σε αυτά τα ζευγάρια αναζήτησης -αποτελέσματος προστίθενται επιπλέον ιδιότητες. Αυτές είναι α) η αξιολόγηση του αποτελέσματος στο εύρος (0,1,2) από πραγματικούς χρήστες και β) η διανυσματική απεικόνιση του αποτελέσματος στο χώρο των χαρακτηριστικών ιδιοτήτων (βλ. ενότητα 4.1, σελίδα 35).

Κατά την προεπεξεργασία πραγματοποιούμε την ένωση join των δυο αυτών αρχείων δημιουργώντας ένα αρχείο ανά αναζήτηση το οποίο περιέχει τη λεκτική περιγραφή του εκάστοτε αποτελέσματος σε διαφορετική σειρά. Τα αποτελέσματα δηλαδή διαχωρίζονται με τον χαρακτήρα αλλαγής γραμμής. Στη συνέχεια μετατρέπουμε αυτή τη λεκτική περιγραφή σε διάνυσμα. Ο διανυσματικός χώρος έχει σαν διαστάσεις σύνολο των όρων που εμφανίζονται σε όλο το πλήθος των αποτελεσμάτων κάθε αναζήτησης. Η συνιστώσα κάθε διάστασης ταυτίζεται με τη συχνότητα εμφανίσεων του συγκεκριμένου όρου στο εκάστοτε αποτέλεσμα. Έτσι κατασκευάζουμε μια δομή συμβατή με τα όσα περιγράφηκαν στο κεφάλαιο 2. Αυτή η δομή πίνακα θα αποτελεί την είσοδο στο εργαλείο συσταδοποίησης Cluto (βλ. ενότητα 2.1.4 σελίδα 20).

3.3 Διατήρηση Αξιολογήσεων

Για κάθε αναζήτηση, ταξινομούμε τα αποτελέσματα με βάση το BM25 score τους, απομονώνουμε τα πρώτα n αποτελέσματα και διατηρούμε τις αξιολογήσεις τους. Με αυτόν τον τρόπο προσομοιάζουμε το ρεαλιστικό σενάριο της αναζήτησης κατά το οποίο ο χρήστης αξιολογεί μόνο τα πρώτα αποτελέσματα κάθε έρευνας που πραγματοποιεί. Τα υπόλοιπα τις περισσότερες φορές δεν τα βλέπει καν οπότε δεν μπορούμε να έχουμε κάποια αξιολόγηση για αυτά. Στις αξιολογήσεις αυτών των πρώτων n αποτελεσμάτων θα αναφερόμαστε με τον όρο *έγκυρες αξιολογήσεις χρήστη*. Αυτές είναι οι αξιολογήσεις που παίρνουμε ως βάση για την μετέπειτα επέκταση και εμπλουτισμό του συνόλου εκπαίδευσης. Κατά τη διάρκεια των πειραμάτων μας, χρησιμοποιήσαμε διαφορετικές τιμές του n μένοντας όμως στη ρεαλιστική υπόθεση ότι ο χρήστης σε κάθε αναζήτηση αξιολογεί τα πρώτα 5-20 αποτελέσματα. Τα αποτελέσματα αυτών των πειραμάτων μπορούν να επηρεάσουν τον τρόπο εμφάνισης των αποτελεσμάτων αναζήτησης, όπως π.χ πόσα αποτελέσματα να εμφανίζονται ανά σελίδα, δεδομένου του ότι ο μέσος χρήστης εποπτεύει και άρα έμμεσα αξιολογεί μόνο τις 1-2 πρώτες σελίδες κάθε αναζήτησης.

Οι αξιολογήσεις των υπόλοιπων αποτελεσμάτων δεν λαμβάνονται πρακτικά υπόψη στα παρακάτω. Χρησιμοποιούνται μόνο για τον έλεγχο των μεθόδων επέκτασης και την εξαγωγή στατιστικών σχετικά με την ακρίβεια της υλοποίησης. Κατά τα άλλα αγνοούνται.

3.4 Συσταδοποίηση

Αρχικά επιλέγουμε τον διανυσματικό χώρο στον οποίο εκφράζουμε τα αποτελέσματα προς συσταδοποίηση. Οι διανυσματικοί χώροι που μας ενδιαφέρουν είναι οι παρακάτω:

- Ο χώρος της *λεκτικής περιγραφής* των αποτελεσμάτων. Περιέχει την Περίληψη, το Σώμα και τον Συγγραφέα του εκάστοτε αποτελέσματος. Για να μεταφέρουμε το κείμενο στον διανυσματικό χώρο λεκτικής περιγραφής ακολουθούμε τη διαδικασία που περιγράφεται στην ενότητα 3.2

- Ο χώρος των *διανυσμάτων χαρακτηριστικών* των αποτελεσμάτων. Πρόκειται για τα διανύσματα χαρακτηριστικών που δίνονται από το benchmark. 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 [14]. Κάποια άλλα χαρακτηριστικά βασίζονται σε πληροφορία από υπερσυνδέσμους (για παράδειγμα τιμές pagerank) [15] ή σε συγκεκριμένες ιδιότητες των αποτελεσμάτων, όπως το domain της διεύθυνσης ή την κατάταξη του αποτελέσματος σε διάφορες μηχανές αναζήτησης [9]. 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 [16].

- Ο *υβριδικός* χώρος. Πρόκειται για συνένωση των δύο παραπάνω χώρων. Ουσιαστικά, ο χώρος της λεκτικής περιγραφής επαυξάνεται από τον χώρο των χαρακτηριστικών διανυσμάτων προσθέτοντας τις επιπλέον διαστάσεις, χωρίς κάποια κανονικοποίηση ή βαρύτητα.

Στους παραπάνω διανυσματικούς χώρους τοποθετούνται τα αποτελέσματα και επιχειρείται η συσταδοποίηση τους. Πειραματιζόμαστε με όλους τους αλγορίθμους που αναλύονται στην ενότητα 2.1.3. Χρησιμοποιούμε την ομοιότητα συνημιτόνου.

Στον χώρο της λεκτικής περιγραφής ισχύουν τα εξής:

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.

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

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

Κατά αντιστοιχία, στο χώρο των διανυσμάτων χαρακτηριστικών κάθε αποτέλεσμα v_i εκφράζεται από το χαρακτηριστικό του διάνυσμα ως εξής: $v_i = \{f_{i1}, f_{i2}, \ldots, f_{in}\}$ και η μετρική ομοιότητας γίνεται:

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

Τέλος ο υβριδικός χώρος προκύπτει από τη συνένωση των παραπάνω οπότε κάθε αποτέλεσμα v_i εκφράζεται ως $v_i = \{wtd_{i1}, wtd_{i2}, \ldots, wtd_{in}, f_{i1}, f_{i2}, \ldots, f_{im}\} = \{a_{i1}, a_{i2}, \ldots, a_{i(n+m)}\}$ και η συνάρτηση ομοιότητας δίνεται αντίστοιχα:

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

Πειραματιστήκαμε με διαφορετικό πλήθος από συστάδες κατά τη συσταδοποίηση. Το πλήθος των συστάδων ανήκει στο σύνολο $\{2, 5, 7, 8, 9, 10, 15, 20\}$

3.5 Επέκταση(expansion) αξιολόγησης

Σκοπός μας είναι να επεκτείνουμε την αξιολόγηση των αρχικών *έγκυρων αξιολογήσεων χρήστη* και να τις γενικεύσουμε ώστε τα μέλη μιας συστάδας να έχουν όλα την ίδια αξιολόγηση.

Στο παρακάτω σχήμα 3.2 βλέπουμε σε γενικές γραμμές τη μέθοδο επέκτασης της αξιολόγησης. Παρακάτω διατυπώνουμε αυστηρά τη μέθοδο και εξετάζουμε μεμονωμένες περιπτώσεις εφαρμογής της.

Η διαδικασία αυτή ονομάζεται επέκταση αξιολόγησης και μπορεί να παρασταθεί με τον παρακάτω αλγόριθμο:

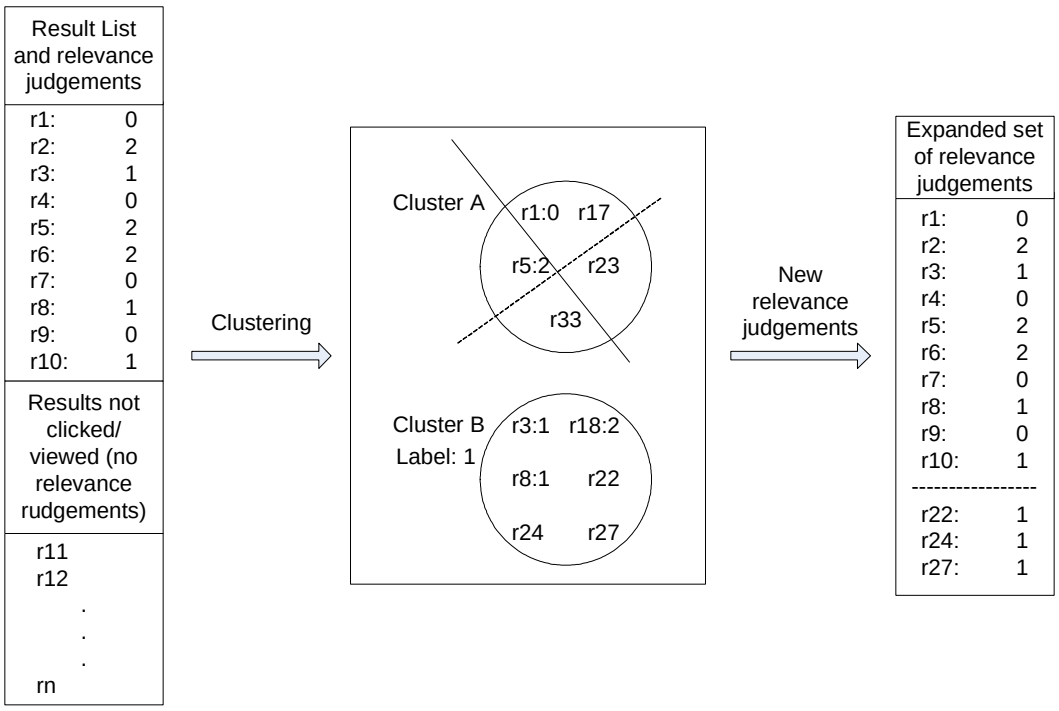
Έστω το σύνολο $V = \{v_1, v_2, \ldots v_l\}$ των έγκυρων (valid) αποτελεσμάτων χρήστη με τις αντίστοιχες αξιολογήσεις $W = \{w_1, w_2, \ldots w_l\}$. Σε κάθε συστάδα S με μέλη το σύνολο $A_S = \{a_1, a_2 \ldots a_k\}$ ορίζουμε τις αντίστοιχες αξιολογήσεις $R_S = \{r_1, r_2 \ldots r_k\}$ όπου:

$$r_i = \begin{cases} w_j & \text{αν } a_i = v_j \text{ για κάποιο } j \in [1 \ldots l] \\ -1 & \text{αλλιώς} \end{cases}$$

- Αν $A_S \cap V = \emptyset$ τότε η συστάδα S απορρίπτεται
- αλλιώς αν $A_S \cap V \neq \emptyset$ τότε $\forall r_i \in R_S$ με $r_i \geq 0$ θεωρώ π_n =πλήθος$\{i : r_i = n\}$ με $n \in \{0, 1, 2\}$
  - αν $|\max\{r_i\} - \min\{r_i\}| < 2$ τότε η γενικευμένη αξιολόγηση J_S δίνεται:

$$J_S = \begin{cases} 0 & \text{αν } \pi_0 > \pi_1, \pi_2 \\ 1 & \text{αν } \pi_1 > \pi_0, \pi_2 \\ 2 & \text{αλλιώς} \end{cases}$$

  - αλλιώς αν $(\pi_0 - \pi_2) > k_1$ τότε $J_S = 0$

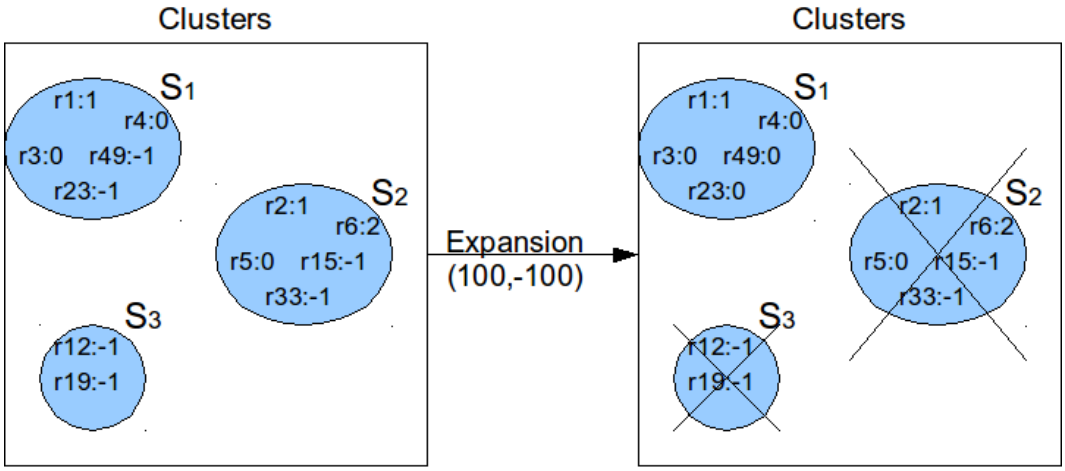


Σχήμα 3.2: Εκπαίδευση μοντέλου μηχανικής μάθησης.

- αλλιώς αν $(\pi_2 - \pi_0) > k_1$ τότε $J_S = 2$
- αλλιώς αν $(\pi_i + \pi_j < k_2 + \pi_m)$ όπου $i, j, m \in \{0, 1, 2\}$ και $i \neq j \neq m$ τότε $J_S = m$
- αλλιώς η συστάδα S απορρίπτεται

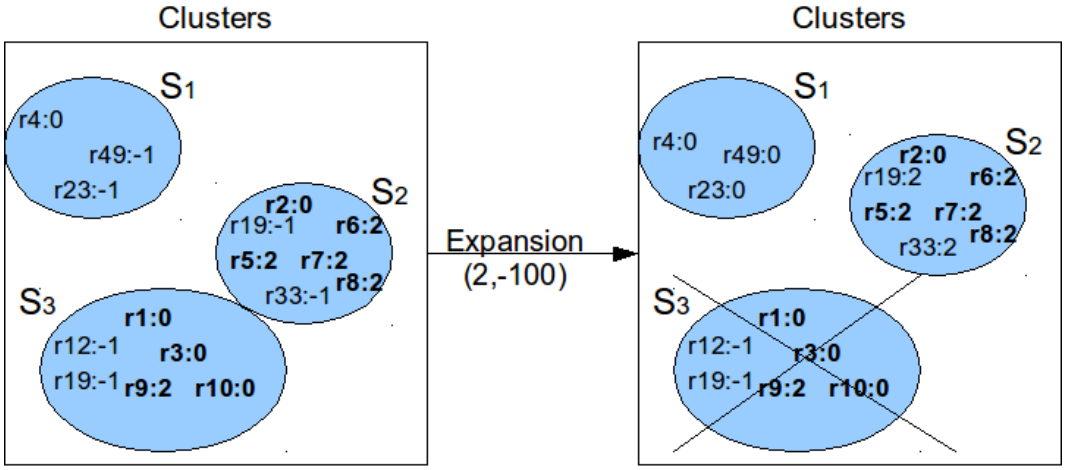
Αναλυτικά, η παραπάνω διαδικασία μπορεί να περιγραφεί ως εξής: Αρχικά, συστάδες που δεν περιέχουν κανένα έγκυρο αποτέλεσμα χρήστη απορρίπτονται και δεν θα χρησιμοποιηθούν περαιτέρω καθώς δεν έχουμε καμία αξιολόγηση για αυτές. Από τις υπόλοιπες συστάδες αν δεν υπάρχουν αποτελέσματα με ακραίες τιμές (0,2) στην ίδια συστάδα τότε η γενικευμένη αξιολόγηση συστάδας δίνεται πλειοψηφικά. Δηλαδή η αξιολόγηση που απαντάται στα περισσότερα αποτελέσματα θα είναι η γενικευμένη αξιολόγηση. Για συστάδες με ακραίες αξιολογήσεις μόνο, αν το πλήθος της μιας αξιολόγησης είναι *αρκετά μεγαλύτερο* (κατώφλι k_1) από το πλήθος της άλλης τότε γενικεύουμε την πολυπληθή αξιολόγηση. Τέλος σε συστάδες που συνυπάρχουν όλες οι αξιολογήσεις αν το πλήθος των αποτελεσμάτων κάποιας από αυτές είναι *αρκετά* (κατώφλι k_2) μεγαλύτερο από το *άθροισμα* των άλλων δύο, τότε αυτή η αξιολόγηση γενικεύεται. Σε διαφορετική περίπτωση, η συστάδα απορρίπτεται.

Στο σχήμα 3.3 φαίνεται ένα παράδειγμα επέκτασης με κατώφλια k1=100 και k2=-100. Αυτή την περίπτωση την ονομάζουμε *απλή επέκταση* καθώς συστάδες με ακραίες κρίσεις απορρίπτονται. Σε αυτό το παράδειγμα, μετά την επέκταση γενικεύεται η αξιολόγηση μόνο της συστάδας S_1 και γίνεται $J_1 = 0$ σημειώνεται εδώ ότι κατά την επέκταση δεν πειράζουμε τις αρχικές έγκυρες αξιολογήσεις χρήστη, αλλά τις διατηρούμε. Η επέκταση επηρεάζει τα αποτελέσματα με κρίση -1. Η συστάδα S_2 απορρίπτεται γιατί περιέχει ακραίες αξιολογήσεις το πλήθος των οποίων δεν επαρκεί για επέκταση με βάση τα κατώφλια k1, k2 που έχουμε



Σχήμα 3.3: Απλή επέκταση με k1=100 και k2=-100.

επιλέξει. Η συστάδα S_3 απορρίπτεται επίσης καθώς δεν περιέχει κανένα έγκυρο αποτέλεσμα χρήστη, οπότε δεν μπορεί να γίνει κάποια επέκταση.



Σχήμα 3.4: Μερική επέκταση με k1=2 και k2=0.

Αντίστοιχα, στο σχήμα 3.4 βλέπουμε ένα παράδειγμα *μερικής επέκτασης*. Εδώ το κατώφλι k1 είναι πλέον 2. Αυτό θα πει ότι για να γίνει επέκταση σε συστάδες με ακραίες αξιολογήσεις, πρέπει το πλήθος των αποτελεσμάτων που φέρουν την μια αξιολόγηση μείον το πλήθος των αποτελεσμάτων που φέρουν την άλλη να είναι κατάπόλυτη τιμή **αυστηρά μεγαλύτερο** του *2*. Δηλαδή, κάνοντας χρήση των παραπάνω συμβολισμών, πρέπει

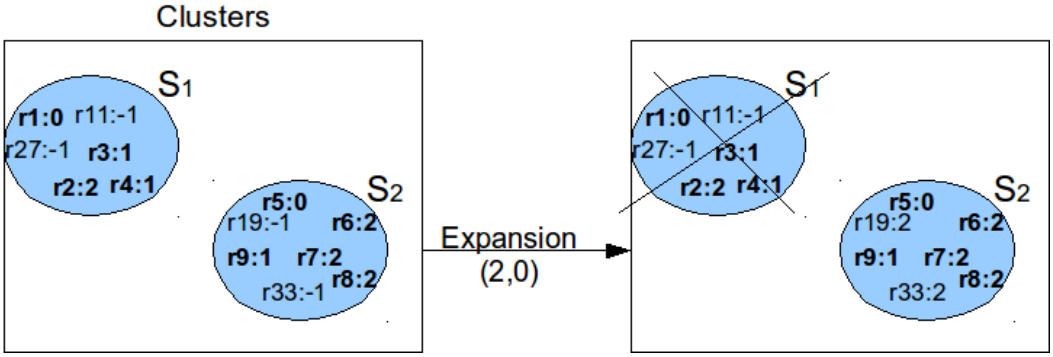
$|\pi_0 - \pi_2| > 2$

Παρατηρούμε ότι αυτή η συνθήκη πληρείται μόνο στην περίπτωση της συστάδας S_2 όπου $|\pi_0 - \pi_2| = |1 - 4| = 3 > 2$ οπότε $J_2 = 2$.

Αντίθετα, στη συστάδα S_3 το πλήθος των έγκυρων αξιολογήσεων δεν επαρκεί για να ικανοποιήσει την παραπάνω σχέση, οπότε τα μέλη της συστάδας απορρίπτονται. Στη συστάδα S_1 γενικεύεται η αξιολόγηση του μοναδικού έγκυρου αποτελέσματος χρήστη $r_4 = 0$ οπότε

$J_1 = 0$

Τέλος παραθέτουμε (σχήμα 3.5) ένα παράδειγμα *πλήρους επέκτασης*, επέκτασης δηλαδή αποτελέσματος σε συστάδα που περιέχει αποτελέσματα και των τριών αξιολογήσεων.



Σχήμα 3.5: Πλήρης επέκταση με k1=2 και k2=0.

Σε αυτή την περίπτωση, η συστάδα S_1 δεν πληροί την αντίστοιχη σχέση ($\pi_i + \pi_j < 0 + \pi_m$) για κανέναν συνδυασμό των (i, j, m). Αντίθετα, στη δεύτερη συστάδα, για ($i = 0$, $j = 1$, $m = 2$) ισχύει η παραπάνω σχέση. Οπότε γενικεύεται η κρίση $J_2 = 2$

Με την παραπάνω μέθοδο αυξάνουμε κατά πολύ το σύνολο εκπαίδευσης. Στα πειράματά μας δοκιμάσαμε διάφορες τιμές των κατωφλιών k_1, k_2 για να βρούμε την καλύτερη ισορροπία, καθώς όπως είναι λογικό αυτή η επέκταση δεν 'μαντεύει' πάντα τις σωστές αξιολογήσεις. Παρόλα αυτά, όπως θα δούμε και στο κεφάλαιο με τα αποτελέσματα, με τη χρήση αυτής της μεθόδου προσεγγίζουμε κατά πολύ την ιδανική περίπτωση κατά την οποία θα χρησιμοποιούσαμε όλες τις αρχικές αξιολογήσεις (στο μη ρεαλιστικό σενάριο κατά το οποίο ο χρήστης αξιολογεί *όλο* το σύνολο των αποτελεσμάτων μιας αναζήτησης).

3.6 Εκπαίδευση μοντέλου μηχανικής μάθησης και έλεγχος

Το εμπλουτισμένο πλέον σύνολο εκπαίδευσης, δίνεται σαν είσοδος στο SVM-rank. Εκπαιδεύεται κατά τα γνωστά (ενότητα 2.2) η συνάρτηση-μοντέλο και πειραματιζόμαστε με διάφορες τιμές της παραμέτρου C (ενότητα 2.2.1). Τα μοντέλα που δημιουργούνται ελέγχονται μέσω του συνόλου επικύρωσης (*validation set*) για να βρεθεί η καλύτερη τιμή της παραμέτρου C. Αυτό το μοντέλο είναι το τελικό μοντέλο της μεθόδου μας. Από αυτό το μοντέλο θα περάσουμε τα αποτελέσματα επόμενων αναζητήσεων για να τα ανακατατάξουμε και να τα παρουσιάσουμε επιδιώκοντας (ιδανικά) να τοποθετήσουμε τα πιο σχετικά αποτελέσματα πιο ψηλά στην λίστα και τα άσχετα αποτελέσματα τελευταία.

Στην υλοποίησή μας αυτό γίνεται μέσω του τελευταίου συνόλου δεδομένων, του συνόλου ελέγχου. Αυτό το σύνολο περιέχει αποτελέσματα νέων αναζητήσεων με τις αξιολογήσεις που τους έχει δώσει ο χρήστης. Εμείς αγνοούμε αρχικά αυτές τις αξιολογήσεις και περνάμε

τα αποτελέσματα από το νευρωνικό δίκτυο που κατασκευάσαμε. Αυτό τα ανακατατάσει και τα τοποθετεί με μια νέα σειρά. Τώρα πια αποκαλύπτουμε την πραγματική αξιολόγηση και βλέπουμε κατά πόσο τα *απόλυτα σχετικά* αποτελέσματα (με αξιολόγηση 2) βρίσκονται ψηλά στη λίστα.

Κεφάλαιο 4

Διεξαγωγή πειραμάτων

Το κεφάλαιο αυτό αφορά στην πειραματική διάταξη της υλοποίησης μας. Αρχικά παρουσιάζουμε το πειραματικό σύνολο που χρησιμοποιήσαμε(LETOR benchmark) με τα δυο ανεξάρτητα σύνολα δεδομένων του (OHSUMED, .GOV). Στη συνέχεια παρουσιάζουμε το περιβάλλον εργασίας (framework) που υλοποιήσαμε για την διεξαγωγή των πειραμάτων. Δίνουμε επίσης το διάγραμμα ροής των δεδομένων κατά την πειραματική διαδικασία.

4.1 LETOR Benchmark

Προκειμένου να δοκιμάσουμε τη μέθοδό μας χρειαζόμαστε ένα σύνολο δεδομένων πραγματικών αναζητήσεων από πραγματικούς χρήστες που συμπεριλαμβάνει τα αποτελέσματα και την αντίστοιχη αξιολόγηση. Τα τελευταία χρόνια (από το 2007) έχει επικρατήσει το δοκιμαστικό σύνολο *LETOR (LEarning To Rank)*. Το σύνολο αυτό προσφέρει τα εξής πλεονεκτήματα:

1. Η δημιουργία ενός δοκιμαστικού συνόλου εμπεριέχει μεγάλο κόστος καθώς πρέπει να γίνει αποκλειστικά από ανθρώπους και οι μεγάλες μηχανές αναζήτησης δεν προσφέρουν ελεύθερα τα ιστορικά αναζήτησης χρηστών στους ερευνητές. Μέσω του LETOR έχουμε ένα αξιόπιστο δοκιμαστικό σύνολο.

2. Το LETOR αποτελεί πλέον σημείο αναφοράς και σύγκρισης ανάμεσα σε διαφορετικές ερευνητικές εργασίες που παλιότερα στήριζαν τα αποτελέσματά τους σε διαφορετικά σύνολα δεδομένων.

3. Εκτός από την αξιολόγηση, κάθε αποτέλεσμα παρουσιάζεται και ως ένα *διάνυσμα χαρακτηριστικών Feature vector*. Αυτό διευκολύνει πολύ καθώς ποσοτικοποιούνται κάποιες ιδιότητες του αποτελέσματος όπως θα δούμε και πιο κάτω.

Το LETOR αποτελείται από δυο συλλογές, την συλλογή OHSUMED και την συλλογή .GOV

4.1.1 Η συλλογή OHSUMED

Αυτή η συλλογή αποτελεί ένα υποσύνολο του MEDLINE, μιας βάσης δεδομένων σχετικά με ιατρικές δημοσιεύσεις. Η συλλογή αποτελείται από 348.566 εγγραφές, από 270 ιατρικά περιοδικά κατά το διάστημα 1987-1991. Πάνω σε αυτά τα δεδομένα γίνονται 106 αναζητήσεις και η κάθε μια επιστρέφει μια λίστα συσχετιζόμενων αποτελεσμάτων. Κάθε αποτέλεσμα συνοδεύεται από την αξιολόγηση που του έδωσε ένας πραγματικός χρήστης σε τρεις βαθμίδες: απόλυτα σχετικό, μερικώς σχετικό, άσχετο. Συνολικά συμπεριλαμβάνονται 16.140 ζευγάρια ερωτήματος -αποτελέσματος με τις αντίστοιχες αξιολογήσεις και το αντίστοιχο διάνυσμα χαρακτηριστικών. Κάθε αναζήτηση επιστρέφει περίπου 152 αξιολογημένα αποτελέσματα.

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).

Στο σχήμα (Α΄.2) που δίνεται στο Παράρτημα παρουσιάζονται αναλυτικά οι διαστάσεις του χώρου χαρακτηριστικών (feature space) των διανυσμάτων.

4.1.2 Η συλλογή .GOV

Σε αντίθεση με το αρκετά εξειδικευμένο σύνολο του OHSUMED αυτή η συλλογή δεν περιορίζεται σε κάποιο συγκεκριμένο αντικείμενο, αλλά αφορά ευρύτερα το WWW. Ο στόχος είναι να ελεγχθεί η μέθοδος όταν αναφερόμαστε σε ένα ευρύ φάσμα χωρίς συγκεκριμένο προσανατολισμό συνδεδεμένο με υπερδείκτες όπως είναι το διαδίκτυο. Συνολικά η συλλογή περιλαμβάνει 1,053,110 html σελίδες. Τα είδη των αναζητήσεων που έχουν πραγματοποιηθεί σε αυτό το σύνολο ανήκουν στις εξης κατηγορίες:

1. *Αναζήτηση θέματος:* Σε αυτόν τον τύπο αναζήτησης η αξιολόγηση των αποτελεσμάτων προσομοιάζει την πραγματική αναζήτηση στο διαδίκτυο: τα αποτελέσματα που αξιολογήθηκαν ως *σχετικά* είναι αυτά τα οποία ταιριάζουν περισσότερο *θεματικά* με τις λέξεις κλειδιά της αναζήτησης. Με άλλα λόγια αποτελούν *σελίδες εισόδου* (π.χ. αρχικές σελίδες) στο περιεχόμενο ενδιαφέροντος.

2. *Αναζήτηση αρχικής σελίδας:* Εδώ δίνοντας τη λέξη κλειδί, απαιτούμε την αρχική σελίδα (.GOV) στην οποία αντιστοιχεί (π.χ για αναζήτηση με λέξη κλειδί USGS σχετικό θεωρείται το αποτέλεσμα *www.usgs.gov*)

3. *Αναζήτηση ονόματος:* Εδώ εισάγουμε σαν κλειδί αναζήτησης το πλήρες URL της σελίδας που αναζητούμε. Σχετικό θεωρείται το αποτέλεσμα-σύνδεσμος σε αυτή τη σελίδα.

Ο χώρος των διανυσμάτων χαρακτηριστικών αυτής της συλλογής περιέχει συνολικά 46 διαστάσεις.

4.1.3 Οργάνωση Δεδομένων

Τέλος, το σετ δεδομένων έχει οργανωθεί σε 5 υποσύνολα, που συμβολίζονται ως $S1$, $S2$, $S3$, $S4$ και $S5$. Συνδυάζοντας κάθε φορά διαφορετικά υποσύνολα για να παράγουν τα: (α) σετ εκπαίδευσης (training set), (β) σετ επικύρωσης (validation set) και (γ) σετ αξιολόγησης (test set), οι δημιουργοί του πειραματικού σετ έχουν παράξει 5 ανακατατάξεις (Folds). Το σετ επικύρωσης χρησιμοποιείται για τη βελτιστοποίηση των παραμέτρων του μοντέλου εκπαίδευσης.Οι ανακατατάξεις φαίνονται στον παρακάτω πίνακα 4.1

Ανακατατάξεις	Σετ Εκπαίδευσης	Σετ Επικύρωσης	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	{$S5$, $S1$, $S2$}	$S3$	$S4$

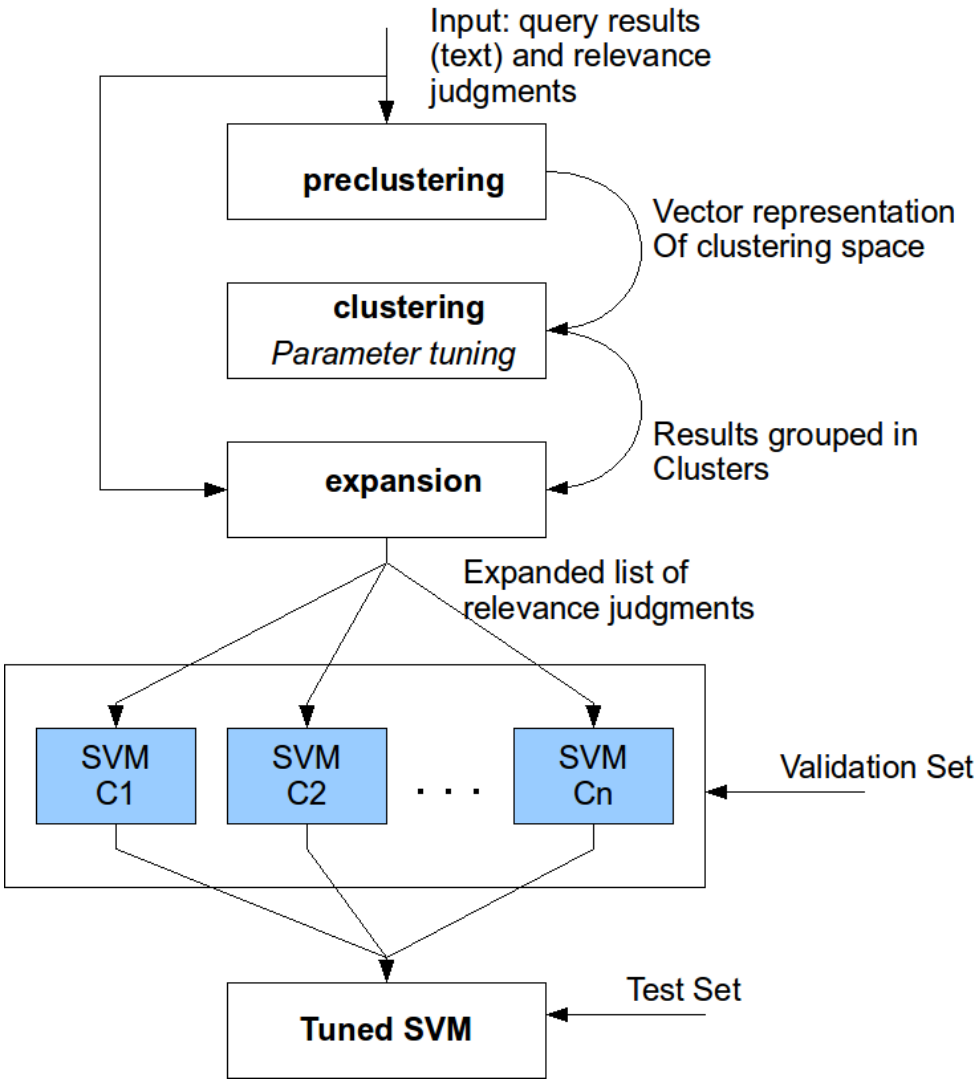
Πίνακας 4.1: Ανακατατάξεις του Σετ Εκπαίδευσης

4.2 Περιβάλλον εργασίας

Το πειραματικό περιβάλλον που σχεδιάσαμε και υλοποιήσαμε απαρτίζεται από τα εξής μέρη:**a) pro-sustadopo'ihsh b) sustadopo'ihsh c)ep'ektash d) ekpa'ideush me SVM**. Η όλη υλοποίηση έγινε σε *Perl* [1]. Διαλέξαμε την *perl* γιατί είναι μια εύχρηστη scripting γλώσσα με δυνατή υποστηρικτική κοινότητα. Επίσης, κάποια εργαλεία που χρησιμοποιήσαμε ήταν γραμμένα σε αυτή τη γλώσσα, οπότε αποφασίσαμε να την διατηρήσουμε ώστε να μην προκύψουν προβλήματα συμβατότητας. Εκ του αποτελέσματος, η επιλογή ήταν πολύ σωστή, η γλώσσα αποδείχτηκε ιδανική μιας και έπρεπε να επεξεργαστούμε μεγάλα αρχεία κειμένου και η perl διαθέτει δυνατά εργαλεία για τέτοιου είδους επεξεργασία. Η ροή των δεδομένων φαίνεται γραφικά στο παρακάτω σχήμα 4.1.

4.2.1 προ-συσταδοποίηση

Σε αυτό το κομμάτι της υλοποίησης σχεδιάσαμε το front end του προγράμματός μας. Πρόκειται για το κομμάτι εκείνο που διαφοροποιείται ανάλογα με το dataset. Η είσοδος σε αυτό το κομμάτι του προγράμματος, όπως φαίνεται και στο σχήμα 4.1 είναι τα κείμενα των εκάστοτε αποτελεσμάτων των αναζητήσεων μαζί με το αρχείο των αξιολογήσεων. Σε αυτή τη φάση πραγματοποιούμε την ένωση των δυο αρχείων και μετατρέπουμε τη λεκτική περιγραφή σε διανυσματική απεικόνιση χρησιμοποιώντας το [6].



Σχήμα 4.1: Ροή δεδομένων

4.2.2 συσταδοποίηση

Τα επεξεργασμένα δεδομένα του προηγούμενου βήματος, εισάγονται σε αυτή τη μονάδα και συσταδοποιούνται. Ο χώρος της συσταδοποίησης έχει προαποφασιστεί στο προηγούμενο βήμα. Εδώ χρησιμοποιούμε το [5] και πειραματιζόμαστε με όλους τους δυνατούς συνδυασμούς των παρακάτω παραμέτρων (αναλύονται στο κεφάλαιο 2):

Αλγόριθμοι Συσταδοποίησης	**Αριθμός Συστάδων**	**Κριτήριο Ομοιότητας**
{rb,rbr,direct,agglo,bagglo,graph,}	{2,5,7,8,9,10,15,20,30}	{h2,e1}

Πίνακας 4.2: Παράμετροι συσταδοποίησης

4.2.3 επέκταση

Εδώ υλοποιούμε τον αλγόριθμο που παρουσιάστηκε στην ενότητα 3.5. Τα κατώφλια k1,k2 παίρνουν τις τιμές που φαίνονται στον παρακάτω πίνακα 4.3

Όνομα	**k1**	**k2**
simple	100	-100
partially expanded	3	-100
partially expanded	2	-100
partially expanded	1	-100
partially expanded	0	-100
fully expanded	2	2
fully expanded	2	1
fully expanded	2	0

Πίνακας 4.3: Παράμετροι συσταδοποίησης

το σύνολο δεδομένων που προκύπτει αποτελείται από το εμπλουτισμένο σύνολο αξιολογήσεων. Αυτό θα αποτελεί το νέο σύνολο εκπαίδευσης.

4.2.4 εκπαίδευση με SVM

Το εμπλουτισμένο πλέον σύνολο δεδομένων εισάγεται στο SVMrank [3]. όπου εκπαιδεύονται μοντέλα με διαφορετικές τιμές της παραμέτρου C (βλ. 2.2.1). Τα μοντέλα αυτά ελέγχονται με τη χρήση του συνόλου επικύρωσης (πίνακας 4.1). Αυτό που παρουσιάζει την καλύτερη ανταπόκριση στο σύνολο επικύρωσης είναι το μοντέλο που θα ελέγξουμε στο με το σύνολο ελέγχου και θα εξάγουμε τα στατιστικά. Η διαδικασία γίνεται ως εξής: εισάγεται στο νευρωνικό δίκτυο το σύνολο ελέγχου και πραγματοποιούνται οι προβλέψεις για την αξιολόγηση των αποτελεσμάτων. Με βάση αυτές τις προβλέψεις, επαναταξινομούνται τα αποτελέσματα της αναζήτησης. Σε αυτό το σημείο αποκαλύπτεται η πραγματική αξιολόγηση και ελέγχεται η ακρίβεια της επαναταξινόμησης των αποτελεσμάτων με βάση τις τρεις μετρικές που ακολουθούν: *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 $NDCG = 1$.

Κεφάλαιο 5

Παρουσίαση Αποτελεσμάτων

Μετά την υλοποίηση και τον πειραματισμό με τα δεδομένα του benchmark εξάγαμε τα παρακάτω αποτελέσματα που δείχνουν την αποτελεσματικότητα της μεθόδου μας. Έχουμε προσπαθήσει να παρουσιάσουμε τα αποτελέσματα των πειραμάτων σε όσο δυνατόν πιο συμπαγή και ευανάγνωστη μορφή.

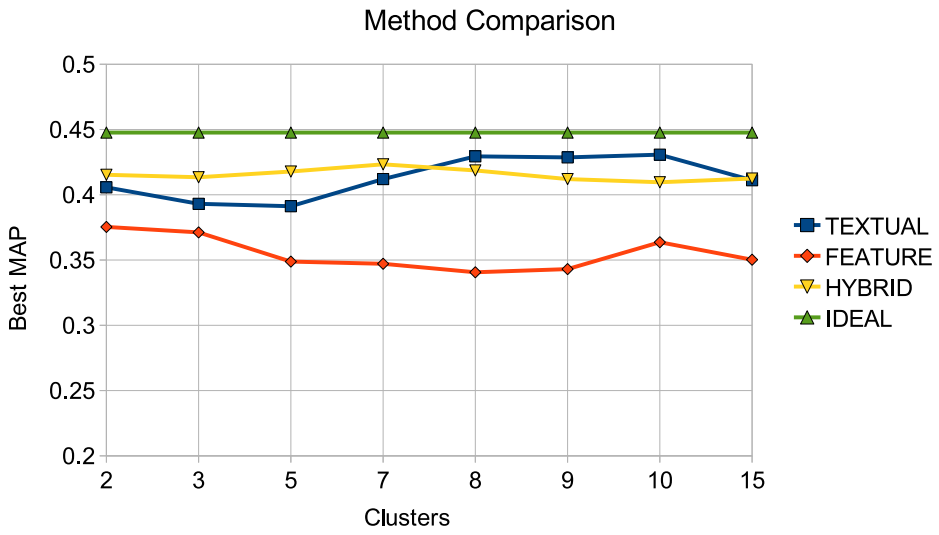
5.1 Σύγκριση χώρων συσταδοποίησης και αριθμού συστάδων

5.1.1 συλλογή OHSUMED

Το παρακάτω σχήμα 5.1 απεικονίζει την ακρίβεια της πρόβλεψης για το σύνολο OHSUMED Συγκεκριμένα, βλέπουμε συγκριτικά τη μετρική *MAP* για τους τρεις χώρους συσταδοποίησης, ανάλογα με το πλήθος των συστάδων που χρησιμοποιήθηκαν. Η γραφική παράσταση απεικονίζει τα αποτελέσματα που εξήχθησαν για την καλύτερη παραμετροποίηση, τόσο κατά την συσταδοποίηση, όσο και κατά την εκπαίδευση του νευρωνικού δικτύου, με βάση τον έλεγχο από το σύνολο επικύρωσης.

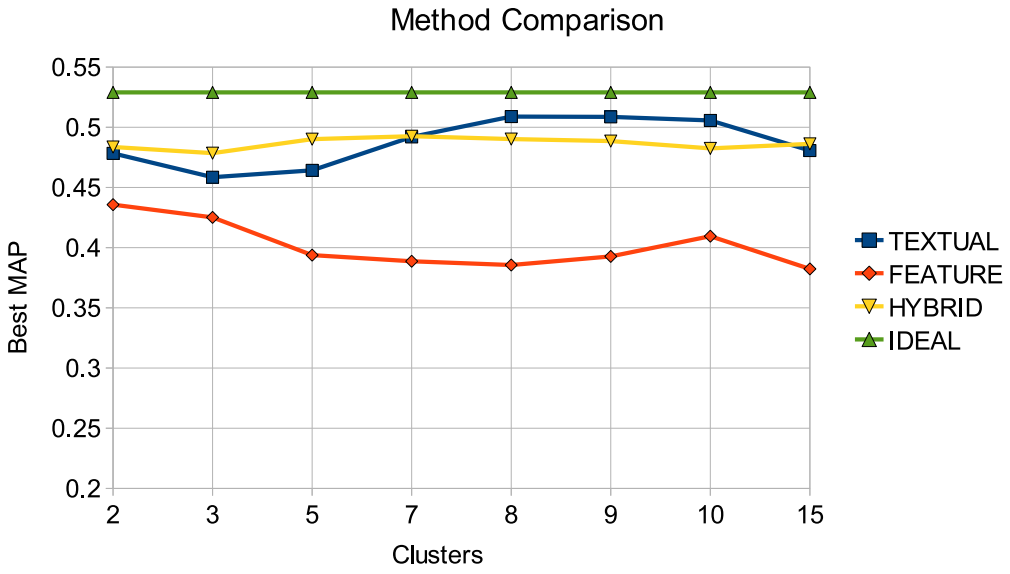
Στο γράφημα παρουσιάζεται η απόκριση του νευρωνικού μοντέλου στο σετ ελέγχου για τους τρεις χώρους (λεκτικής περιγραφής, διανυσμάτων χαρακτηριστικών και υβριδικό) καθώς και για την ιδανική περίπτωση που δίνουμε προς εκπαίδευση του μοντέλου το σύνολο των αξιολογήσεων, όλο δηλαδή το training set. Όπως έχει αναφερθεί η ιδανική περίπτωση είναι μη εφικτή πρακτικά, καθώς ο μέσος χρήστης ποτέ δεν αξιολογεί όλο το σύνολο των αποτελεσμάτων μιας αναζήτησής του.

Βλέπουμε πράγματι ότι στο χώρο της λεκτικής περιγραφής, τα αποτελέσματα της ακρίβειας προσεγγίζουν κατά πολύ την ιδανική προσέγγιση, υπολειπόμενα μόνο κατά ένα 4,4%. Η βέλτιστη απόκριση φαίνεται να είναι για 10 συστάδες. Πράγματι, αυτό ήταν σχετικά αναμενόμενο, καθώς το μικρό πλήθος συστάδων ομαδοποιεί ανόμοια αντικείμενα οπότε η επέκταση αποτυγχάνει. Αντίστοιχα κάνοντας χρήση πολλών συστάδων αυξάνονται οι πιθανότητες μεγάλης διασποράς των αρχικών έγκυρων αξιολογήσεων, κάτι που επιφέρει απόρριψη πολλών συστάδων, και μείωση των δεδομένων εκπαίδευσης. Παρατηρούμε επίσης κατάφωρη υπεροχή



Σχήμα 5.1: Συγκριτική παρουσίαση των μεθόδων συσταδοποίησης και της ιδανικής περίπτωσης για τη μετρική MAP.

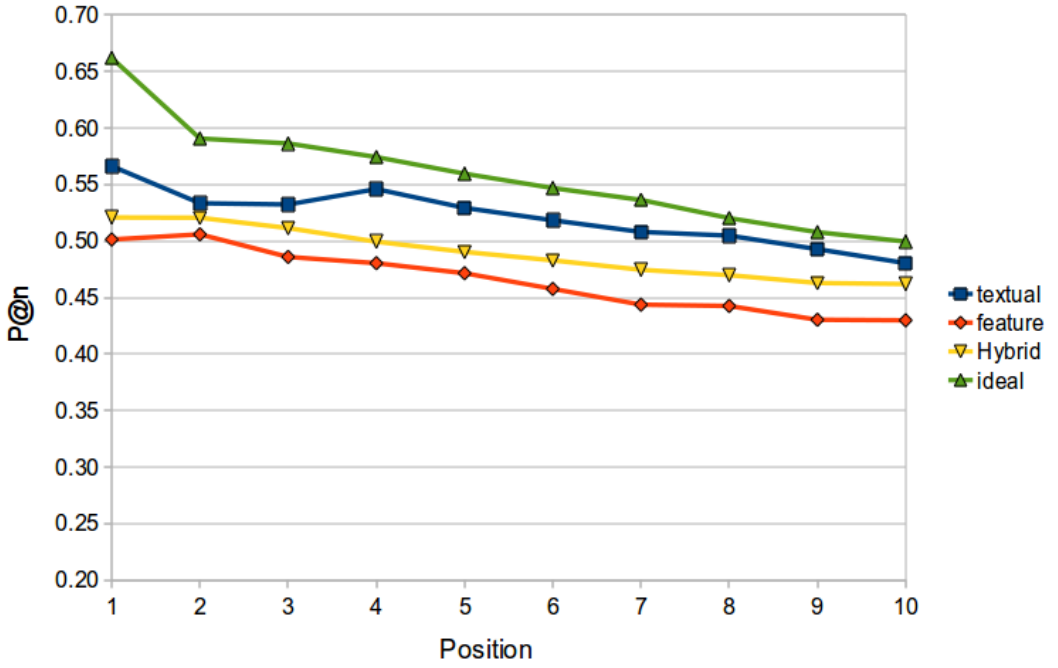
του χώρου λεκτικής περιγραφής έναντι στους άλλους δυο χώρους. Αντίστοιχη εικόνα παρουσιάζει και η μετρική *NDCG* που φαίνεται στο παρακάτω σχήμα 5.2, όπως και η μετρική *P@n* που φαίνεται στο σχήμα 5.3



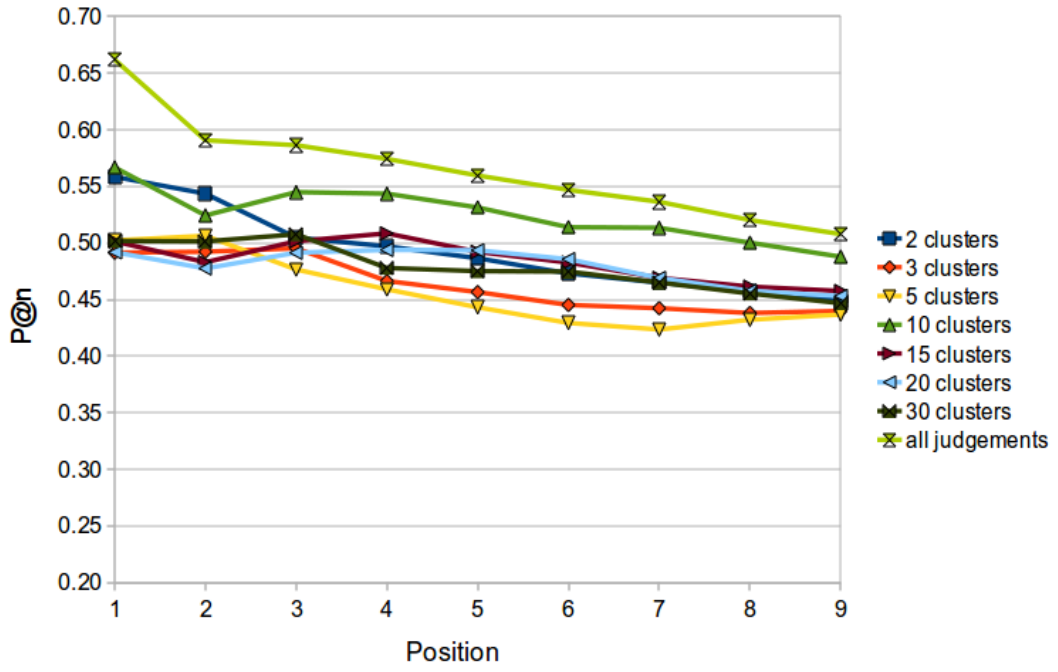
Σχήμα 5.2: Συγκριτική παρουσίαση των μεθόδων συσταδοποίησης και της ιδανικής περίπτωσης για τη μετρική NDCG.

Σε ότι αφορά στον χώρο της λεκτικής περιγραφής, μπορούμε να δούμε συγκριτικά τα αποτελέσματα της μετρικής *Precision @ n(P@n)* στο παρακάτω σχήμα 5.4.

Βλέπουμε ότι η καμπύλη της μεθόδου μας για 10 συστάδες, τείνει να προσεγγίσει την ιδανι-

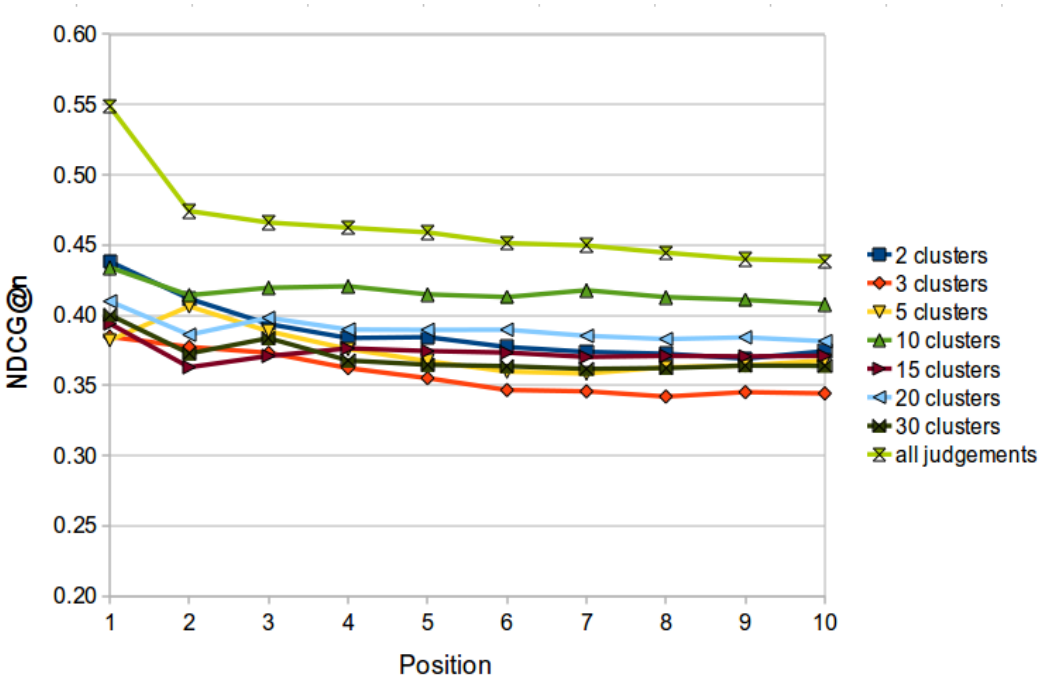


Σχήμα 5.3: Συγκριτική παρουσίαση των μεθόδων συσταδοποίησης και της ιδανικής περίπτωσης για τη μετρική P@n.



Σχήμα 5.4: Συγκριτική παρουσίαση στο χώρο λεκτικής περιγραφής για διαφορετικό πλήθος συστάδων.

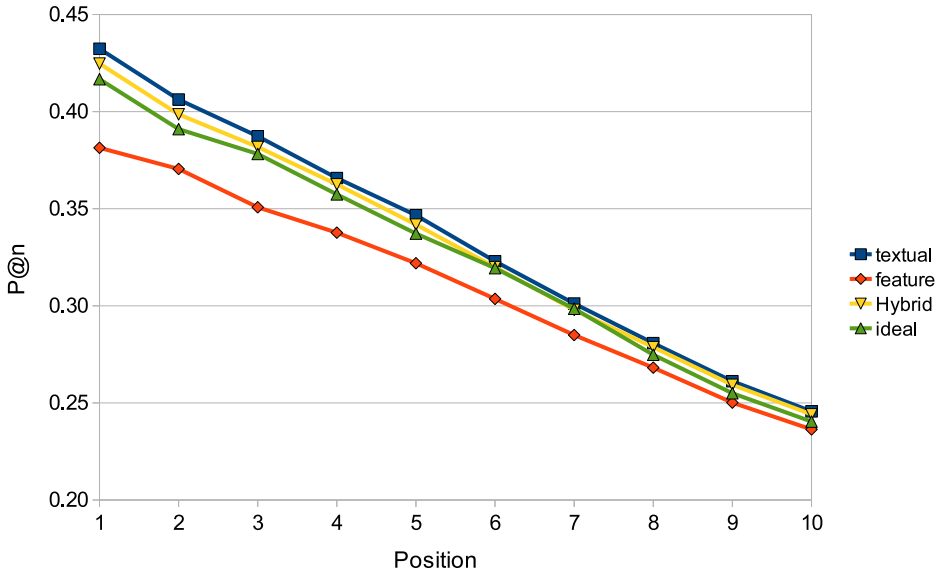
κή μέθοδο, και διαφοροποιείται αρκετά από τις υπόλοιπες καμπύλες. Αντίστοιχη συμπεριφορά παρατηρούμε και για τη μετρική *NDCG@n* που φαίνεται στο σχήμα 5.5.



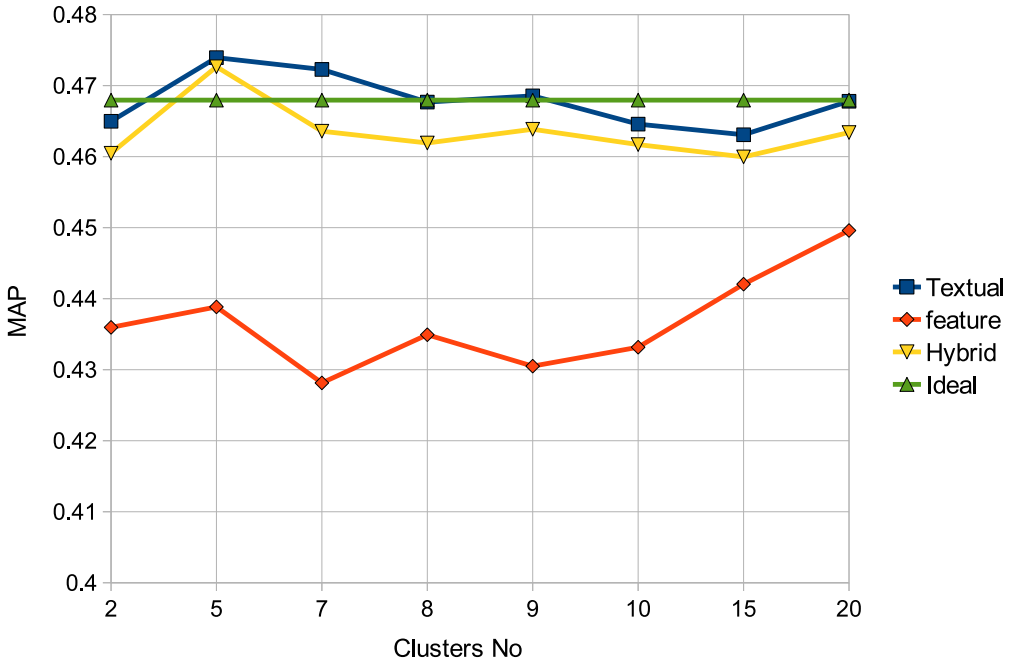
Σχήμα 5.5: Συγκριτική παρουσίαση στο χώρο λεκτικής περιγραφής για διαφορετικό πλήθος συστάδων(NDCG).

5.1.2 συλλογή .GOV

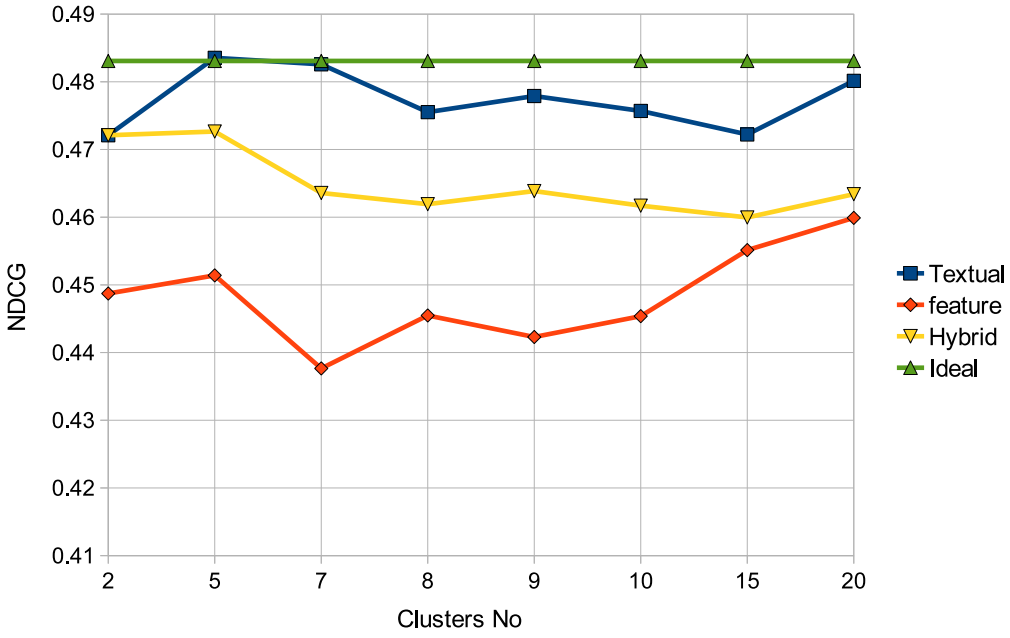
Ανάλογα είναι και τα αποτελέσματα για το δεύτερο dataset, το .GOV dataset. Τα παρακάτω σχήματα παρουσιάζουν τις αντίστοιχες μετρικές.



Σχήμα 5.6: Συγκριτική παρουσίαση των μεθόδων συσταδοποίησης και της ιδανικής περίπτωσης για τη μετρική P@n στο .GOV dataset



Σχήμα 5.7: Συγκριτική παρουσίαση των μεθόδων συσταδοποίησης και της ιδανικής περίπτωσης για τη μετρική MAP στο .GOV dataset



Σχήμα 5.8: Συγκριτική παρουσίαση των μεθόδων συσταδοποίησης και της ιδανικής περίπτωσης για τη μετρική NDCG στο .GOV dataset

Μάλιστα σε αυτό το σύνολο δεδομένων παρατηρούμε ότι σε κάποιες περιπτώσεις η εκπαίδευση που επιτυγχάνεται με τη μέθοδό μας, έχει καλύτερα αποτελέσματα από την ιδανική

(σχήματα 5.7 και 5.8). Πάντως γενικά μπορούμε να πούμε ότι η μέθοδος μας με την κατάλληλη παραμετροποίηση αποκρίνεται πολύ καλά, και κυμαίνεται γύρω από το ιδανικό. Στη σύγκριση των μεθόδων παρατηρούμε ξεκάθαρη υπεροχή της συσταδοποίησης με βάση τη λεκτική περιγραφή, ενώ η συσταδοποίηση στο χώρο των διανυσμάτων χαρακτηριστικών υπολείπεται αρκετά, με την υβριδική μέθοδο να εμφανίζεται κάπου ανάμεσα, έχοντας όμως απορροφήσει τις έντονες ακμές και διακυμάνσεις των δυο άλλων σε ότι αφορά τον αριθμό των συστάδων.

5.2 Σύγκριση αλγόριθμων συσταδοποίησης

Κατά τη διεξαγωγή των πειραμάτων, δοκιμάσαμε πολλούς διαφορετικούς συνδυασμούς αλγορίθμων, κριτηρίων ομοιότητας και πλήθους συστάδων. Στον πίνακα 5.1 που ακολουθεί, παρουσιάζουμε τον καλύτερο συνδυασμό για κάθε αριθμό συστάδων στον χώρο της λεκτικής περιγραφής των αποτελεσμάτων για το σύνολο OHSUMED.

Clusters	**algorithm**	**creterinon**	**MAP**	**NDCG**
2	agglo	h2	0.41	0.48
3	direct	h2	0.39	0.46
4	agglo	h2	0.41	0.48
5	bagglo	e1	0.39	0.46
6	bagglo	h2	0.4	0.47
7	bagglo	e1	0.41	0.49
8	bagglo	h2	0.43	0.51
9	bagglo	h2	0.43	0.51
10	bagglo	h2	0.43	0.51
15	bagglo	e1	0.41	0.48
20	bagglo	e1	0.4	0.48
30	direct	h2	0.41	0.47

Πίνακας 5.1: Αλγόριθμοι συσταδοποίησης για textual

Παρατηρούμε ότι στη μεσαία ζώνη, από 5 έως 20 συστάδες, ο αλγόριθμος biased agglomerative δίνει τα καλύτερα αποτελέσματα. Στη συνέχεια θα χρησιμοποιήσουμε αυτόν για τα υπόλοιπα πειράματά μας, αφού για το συγκεκριμένο τουλάχιστον dataset φαίνεται να επικρατεί σχεδόν καθολικά.

Αντίστοιχο συμπέρασμα δεν μπορεί να βγει για το χώρο των διανυσμάτων χαρακτηριστικών. Όπως βλέπουμε μάλιστα στον παρακάτω πίνακα 5.2, δεν φαίνεται να υπάρχει κάποιος αλγόριθμος που να ξεχωρίζει.

Τέλος, μελετώντας τον υβριδικό χώρο συντάσσουμε τον πίνακα 5.3.Εδώ, για τη μεσαία ζώνη, καλύτερη απόκριση φαίνεται να έχει ο αλγόριθμος graph ενώ ο απλός agglomerative ανταποκρίνεται καλύτερα στις ακραίες ζώνες συστάδων.

Εδώ πρέπει να σημειώσουμε ότι δεν πειραματιστήκαμε αντίστοιχα για το σύνολο (.GOV) καθώς η τόσο αναλυτική πειραματική διαδικασία θεωρήθηκε υπερβολικά χρονοβόρα. Αντίθετα, επεκτείναμε τα συμπεράσματα του πειραματισμού με το σύνολο OHSUMED στην διεξαγωγή

Clusters	algorithm	creterion	MAP	NDCG
2	bagglo	h2	0.38	0.44
3	rbr	e1	0.37	0.43
5	rbr	h2	0.35	0.39
7	agglo	e1	0.35	0.39
8	rb	h2	0.34	0.39
9	direct	e1	0.34	0.39
10	agglo	h2	0.36	0.41
15	graph	e1	0.35	0.38

Πίνακας 5.2: Αλγόριθμοι συσταδοποίησης για feature

Clusters	algorithm	creterion	MAP	NDCG
2	agglo	h2	0.42	0.48
3	agglo	e1	0.41	0.48
5	graph	h2	0.42	0.49
7	graph	e1	0.42	0.49
8	graph	h2	0.42	0.49
9	agglo	e1	0.41	0.49
10	agglo	e1	0.41	0.48
15	agglo	e1	0.41	0.49

Πίνακας 5.3: Αλγόριθμοι συσταδοποίησης για hybrid

των πειραμάτων για το σύνολο .GOV και κρατήσαμε τους καλύτερους αλγορίθμους συσταδοποίησης ανά περίπτωση.

5.3 Σύγκριση μεθόδων επέκτασης

Στα παρακάτω σχήματα απεικονίζουμε με συμπαγή τρόπο τα αποτελέσματα των πειραματισμών μας με διαφορετικές μεθόδους επέκτασης. Ουσιαστικά πρόκειται για διαφοροποίηση των κατωφλιών k1,k2 όπως αναλύθηκε στο προηγούμενο κεφάλαιο. Τα διαγράμματα που ακολουθούν εμπεριέχουν τη σύγκριση ανάμεσα στα ποσοστά των προβλέψεων κατά την επέκταση και τον αντίκτυπο που αυτά έχουν στην έξοδο του νευρωνικού μοντέλου (αυτός ο αντίκτυπος παρουσιάζεται μέσω της μετρικής MAP). Τα ποσοστά έχουν χωριστεί σε τρεις κατηγορίες:

1. σωστά(correct): αυτά για τα οποία η πραγματική αξιολόγηση ταυτίζεται με την αξιολόγηση που πετυχαίνουμε μέσω της μεθόδου της επέκτασης

2. εν μέρη σωστά (partially correct): αυτά για τα οποία η πραγματική αξιολόγηση διαφέρει (κατ΄ απόλυτη) τιμή από την αξιολόγηση μέσω της επέκτασης, ακριβώς ένα. Για παράδειγμα, εν μέρη σωστό θεωρείται ένα αποτέλεσμα που ενώ έχει πραγματική αξιολόγηση 2, μέσω της μεθόδου της επέκτασης του ανατέθηκε η κρίση 1.

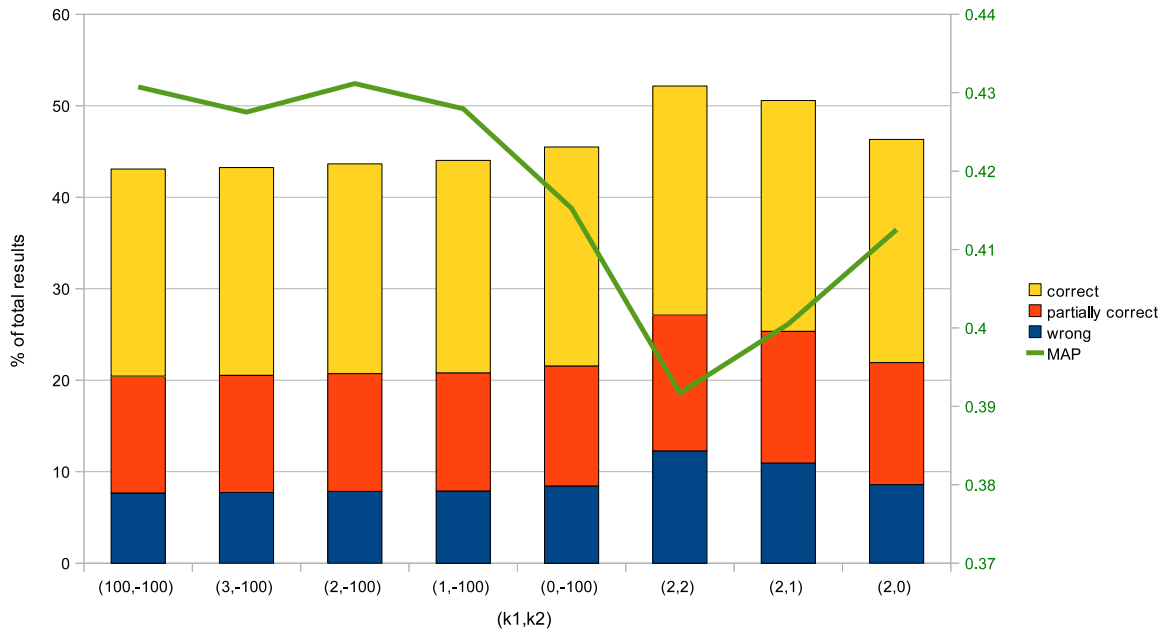
3. λάθος (wrong): αυτά τα αποτελέσματα για τα οποία η πραγματική αξιολόγηση διαφέρει (κατ΄ απόλυτη τιμή) από την αξιολόγηση μέσω της επέκτασης ακριβώς 2. Πρόκειται για

τις περιπτώσεις που η μέθοδος της επέκτασης έχει αποτύχει πλήρως.

Τα αποτελέσματα παρουσιάζονται στη μορφή σύνθετου ραβδογράμματος με καμπύλη για τη μετρική ακρίβειας MAP. Στα διαγράμματα που ακολουθούν έχουμε τοποθετήσει δυο άξονες ψι, τον πρωτεύοντα (αριστερά) και τον δευτερεύοντα (δεξιά). Ο πρωτεύων άξονας, αφορά στα ποσοστά των προβλέψεων (ραβδογράμματα) και μετράει ποσοστιαίες μονάδες. Αντίθετα, ο δευτερεύων άξονας, αφορά στην μετρική ακρίβειας MAP (καμπύλη) έχει εντελώς διαφορετική κλίμακα, και μετράει σε απόλυτες μονάδες (εύρος 0-1). Ο κοινός άξονας χι, αποτελείται από τα ζευγάρια κατωφλιών που παρουσιάστηκαν παραπάνω. Σημειώνουμε ότι ο άξονας χι έχει τη σχέση διάταξης που φαίνεται στον πίνακα 4.3

Ο λόγος για τον οποίο κατασκευάσαμε τα διαγράμματα κατάυτόν τον τρόπο είναι ότι θέλαμε να δώσουμε έμφαση στην συγκριτική απεικόνιση της μεθόδου επέκτασης με την ακρίβεια του μοντέλου εκπαίδευσης που δημιουργήσαμε. Ο πιο αποδοτικός και οικονομικός από πλευράς χώρου τρόπος να το κάνουμε αυτό, ήταν μέσω των σύνθετων διαγραμμάτων που ακολουθούν. Στα διαγράμματα αυτά, φαίνεται ξεκάθαρα η εξάρτηση της ακρίβειας του παραγόμενου μοντέλου από τη μέθοδο επέκτασης.

5.3.1 σύνολο OHSUMED



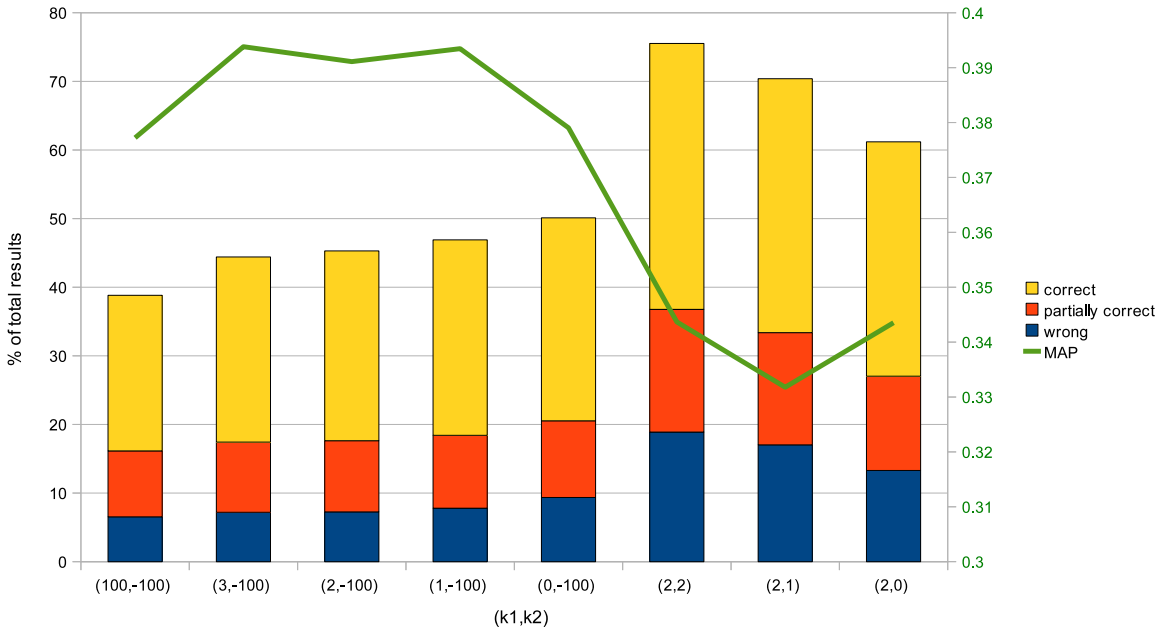
Σχήμα 5.9: Συγκριτική απεικόνιση μεθόδου επέκτασης και αποτελέσματος εκπαίδευσης του SVM. Ο χώρος συσταδοποίησης είναι αυτός της λεκτικής περιγραφής (textual)

Βοηθητικά, παραθέτουμε και τον πίνακα 5.4 που δίνει τις τιμές που αφορούν στο σχήμα 5.9 μαζί με το απόλυτο πλήθος των προβλέψεων. Επεξηγηματικά για τις γραφικές απεικονίσεις, να αναφέρουμε ότι τα ραβδογράμματα δίνουν το ποσοστό των αποτελεσμάτων στα οποία έγινε

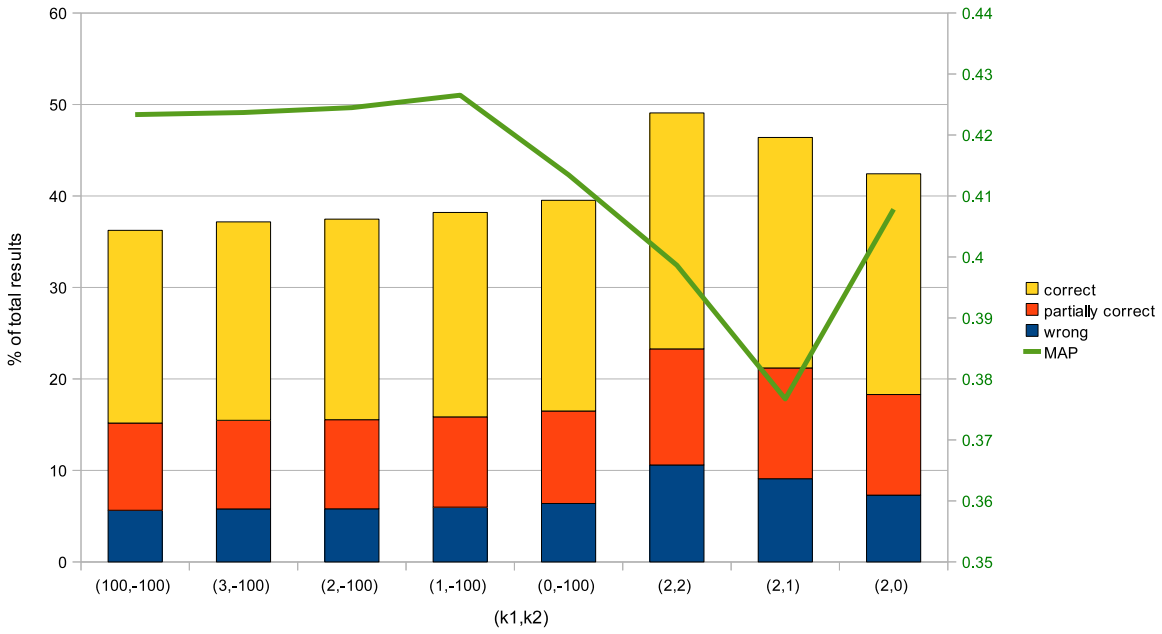
textual	wrong	partially correct	correct	MAP	total predictions
(100,-100)	7.68	12.78	22.63	0.43	4172.6
(3,-100)	7.73	12.81	22.71	0.43	4187.8
(2,-100)	7.85	12.87	22.93	0.43	4226.2
(1,-100)	7.89	12.91	23.22	0.43	4263.2
(0,-100)	8.44	13.12	23.93	0.42	4405
(2,2)	12.28	14.84	25.04	0.39	5056
(2,1)	10.95	14.4	25.23	0.4	4900.2
(2,0)	8.59	13.34	24.39	0.41	4485.8

Πίνακας 5.4: Τιμές που αφορούν στο σχήμα 5.9

επέκταση αξιολόγησης. μέσα σε κάθε ραβδόγραμμα, φαίνεται ο καταμερισμός των αποτελεσμάτων στις κατηγορίες που προαναφέρθηκαν. Για παράδειγμα, *στο σχήμα 5.9 βλέπουμε ότι για την επέκταση με κατώφλια (k1,k2) = (100, -100) καταφέραμε να επεκτείνουμε την αξιολόγηση στο 43% περίπου των αποτελεσμάτων. Αυτό το 43% διαμερίζεται ως εξής: το 7.68% των προβλέψεων απέτυχε, το 12.78% πέτυχε μερικώς, και το υπόλοιπο 22.63% πέτυχε απόλυτα. Αυτή η επέκταση είχε ως αποτέλεσμα, να κατασκευαστεί ένα μοντέλο που όταν ελέγχθηκε με το σύνολο εκπαίδευσης έδωσε μέση ακρίβεια MAP=0,43*



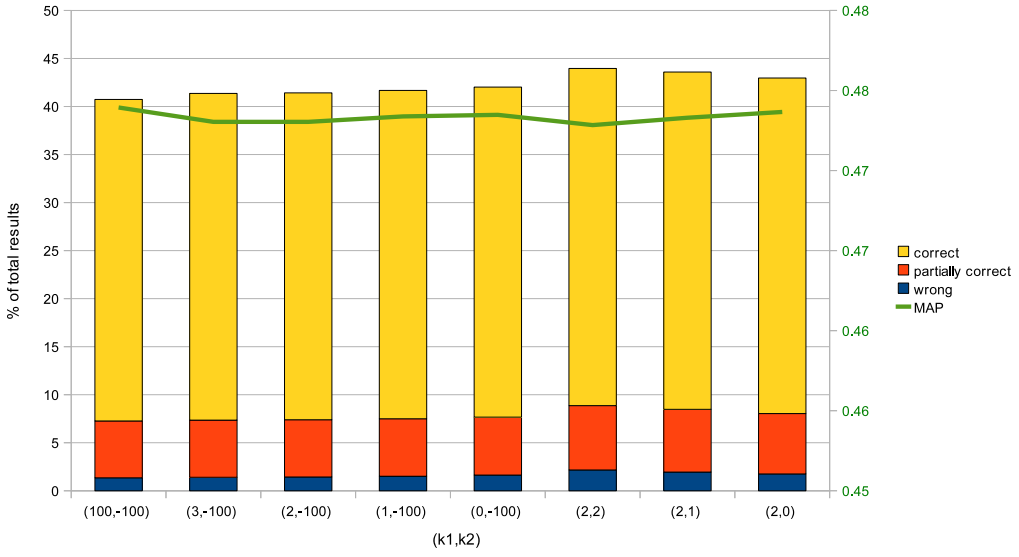
Σχήμα 5.10: Συγκριτική απεικόνιση μεθόδου επέκτασης και αποτελέσματος εκπαίδευσης του SVM. Ο χώρος συσταδοποίησης είναι αυτός των χαρακτηριστικών διανυσμάτων (feature)



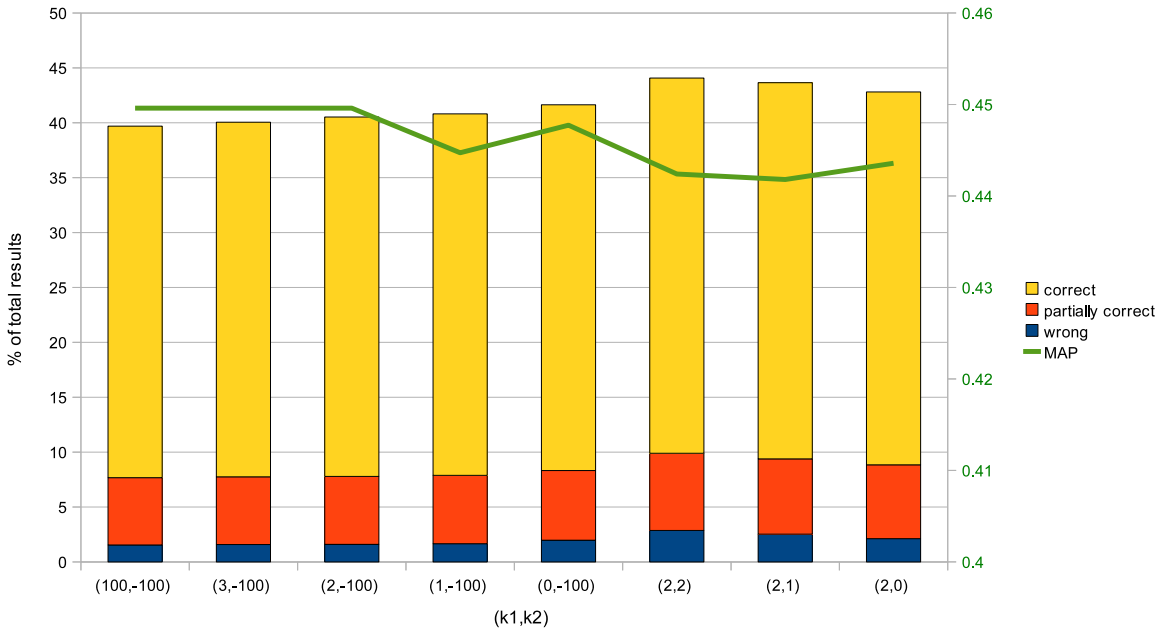
Σχήμα 5.11: Συγκριτική απεικόνιση μεθόδου επέκτασης και αποτελέσματος εκπαίδευσης του SVM. Ο χώρος συσταδοποίησης είναι ο υβριδικός (hybrid)

5.3.2 σύνολο .GOV

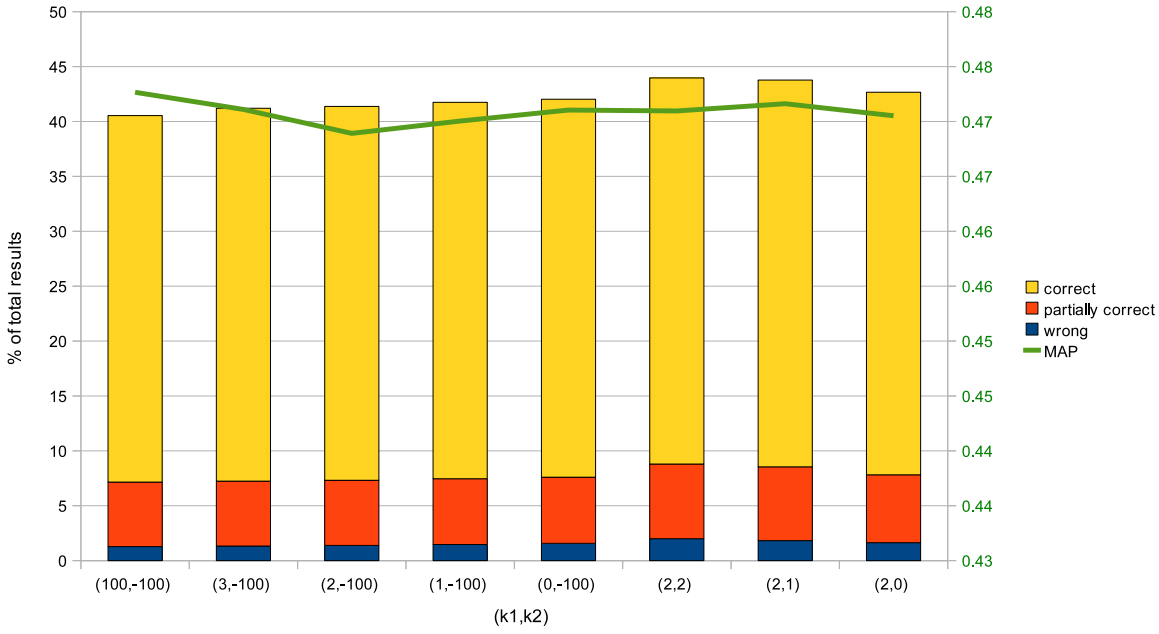
Αντίστοιχα, παραθέτουμε και τα αποτελέσματα του δεύτερου συνόλου, .GOV που δίνονται για τους τρεις διανυσματικούς χώρους συσταδοποίησης, στα σχήματα 5.12 - 5.14



Σχήμα 5.12: Συγκριτική απεικόνιση μεθόδου επέκτασης και αποτελέσματος εκπαίδευσης του SVM για το σύνολο .GOV. Ο χώρος συσταδοποίησης είναι αυτός της λεκτικής περιγραφής (textual)



Σχήμα 5.13: Συγκριτική απεικόνιση μεθόδου επέκτασης και αποτελέσματος εκπαίδευσης του SVM για το σύνολο .GOV. Ο χώρος συσταδοποίησης είναι αυτός της λεκτικής περιγραφής (textual)



Σχήμα 5.14: Συγκριτική απεικόνιση μεθόδου επέκτασης και αποτελέσματος εκπαίδευσης του SVM για το σύνολο .GOV.Ο χώρος συσταδοποίησης είναι ο υβριδικός (hybrid)

5.4 Σύγκριση διαφορετικού πλήθους έγκυρων προβλέψεων χρήστη

Εδώ παρουσιάζουμε τα αποτελέσματα πάνω στον πειραματισμό μας με διαφορετικό πλήθος έγκυρων προβλέψεων χρήστη. Θυμίζουμε εδώ, ότι έγκυρες αξιολογήσεις χρήστη θεωρούνται τα αποτελέσματα εκείνα της κάθε αναζήτησης τα οποία θεωρούμε ότι ο χρήστης έχει αξιολογήσει, και τα οποία χρησιμοποιούμε στη συνέχεια της μεθόδου για την επέκταση των αξιολογήσεων. Σε όλα τα παραπάνω αποτελέσματα, είχαμε διατηρήσει σταθερή αυτή την παράμετρο στο δέκα (10). Δηλαδή, έγκυρα αποτελέσματα χρήστη θεωρούσαμε μόνο τα 10 πρώτα κάθε αναζήτησης. Σε αυτήν την ενότητα, θα δείξουμε την διακύμανση της ακρίβειας του παραγόμενου μοντέλου ανάλογα με το πλήθος των έγκυρων αποτελεσμάτων χρήστη. Για τα πειράματά μας, περιοριζόμαστε στο σύνολο OHSUMED, στους χώρους της λεκτικής περιγραφής και των διανυσμάτων χαρακτηριστικών και στην κατωφλιοποίηση (κ1,κ2) = (100,-100) δηλαδή την *απλή επέκταση* καθώς είναι αυτές που παρουσιάζουν το μεγαλύτερο ενδιαφέρον. Ο παρακάτω πίνακας 5.5 αφορά στο χώρο της λεκτικής περιγραφής.

n	**MAP**	**NDCG**
5 FIRST	0.389	0.443
8 FIRST	0.414	0.481
10 FIRST	0.431	0.506
20 FIRST	0.422	0.500

Πίνακας 5.5: Διακύμανση απόκρισης ακρίβειας με διαφορετικό πλήθος αρχικών έγκυρων αξιολογήσεων, για το χώρο συσταδοποίησης της λεκτικής περιγραφής

Αντίστοιχα ο πίνακας 5.6 αφορά στον χώρο των διανυσμάτων χαρακτηριστικών:

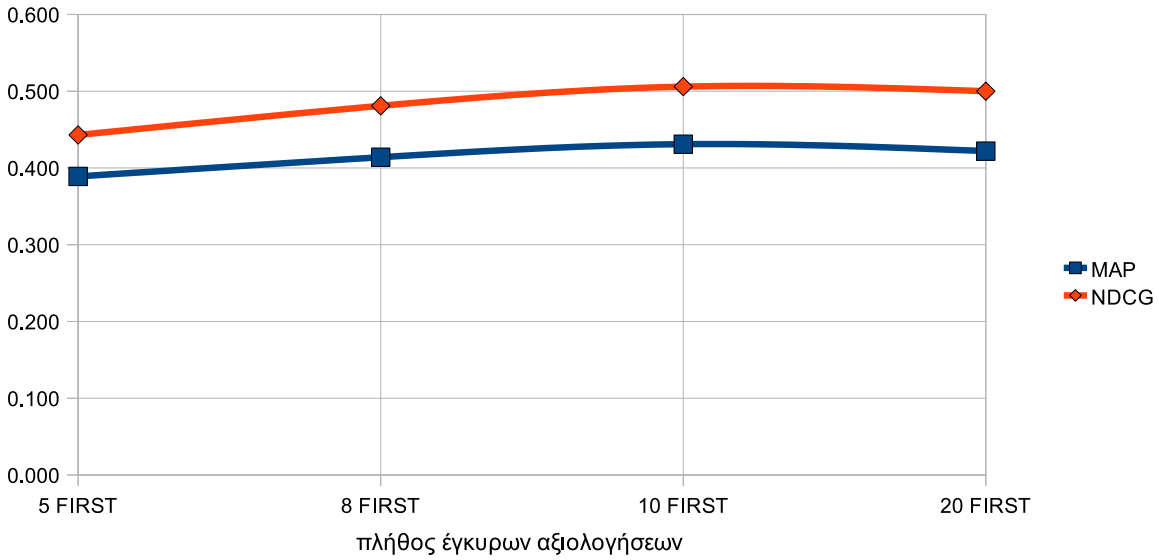
n	**MAP**	**NDCG**
5 FIRST	0.349	0.398
8 FIRST	0.414	0.481
10 FIRST	0.371	0.425
20 FIRST	0.422	0.500

Πίνακας 5.6: Διακύμανση απόκρισης ακρίβειας με διαφορετικό πλήθος αρχικών έγκυρων αξιολογήσεων, για το χώρο συσταδοποίησης των διανυσμάτων χαρακτηριστικών

Στους παραπάνω πίνακες παρουσιάζεται η εξάρτηση της ακρίβειας του μοντέλου (που εκφράζεται μέσω των μετρικών MAP και NDCG) από το πλήθος των αρχικών έγκυρων αποτελεσμάτων χρήστη (n). Η αναμενόμενη συμπεριφορά του μοντέλου θα ήταν αύξηση της ακρίβειας με αύξηση του πλήθους n. Πράγματι, έχοντας περισσότερες έγκυρες αξιολογήσεις χρήστη, λογικό είναι να περιμένει κανείς ότι η μέθοδος της επέκτασης θα φέρει καλύτερα αποτελέσματα αφού το αρχικό υλικό πάνω στο οποίο βασίζεται η επέκταση είναι μεγαλύτερο. Παρόλα αυτά, οι παραπάνω πίνακες δείχνουν ότι τα αποτελέσματα απέχουν από την αναμενόμενη συμπεριφορά. Συγκεκριμένα, στον πίνακα 5.5 παρατηρούμε ότι η αύξηση του πλήθους n πέρα από το δέκα (10) αντί να βελτιώνει την απόκριση του μοντέλου, τη χειροτερεύει. Μια

εξήγηση που μπορεί να δοθεί για αυτό το γεγονός, είναι ότι το μοντέλο που εκπαιδεύουμε για $n = 10$ είναι *κορεσμένο*. Πράγματι, όπως είδαμε νωρίτερα, έχουμε προσεγγίσει αρκετά την ιδανική περίπτωση κάνοντας χρήση μόνο των 10 πρώτων αποτελεσμάτων, οπότε η επιπλέον πληροφορία δεν αρκεί για να βελτιώσει την ακρίβεια. Αντίθετα, τα επιπλέον αποτελέσματα εισάγουν μεγαλύτερη διασπορά στις συστάδες κατά τη συσταδοποίηση, κάτι που ενδεχομένως να μειώνει δραστικά το πλήθος των έγκυρων προβλέψεων. Όπως και να έχει αυτή η συμπεριφορά είναι αξιοσημείωτη (σχήμα 5.15).

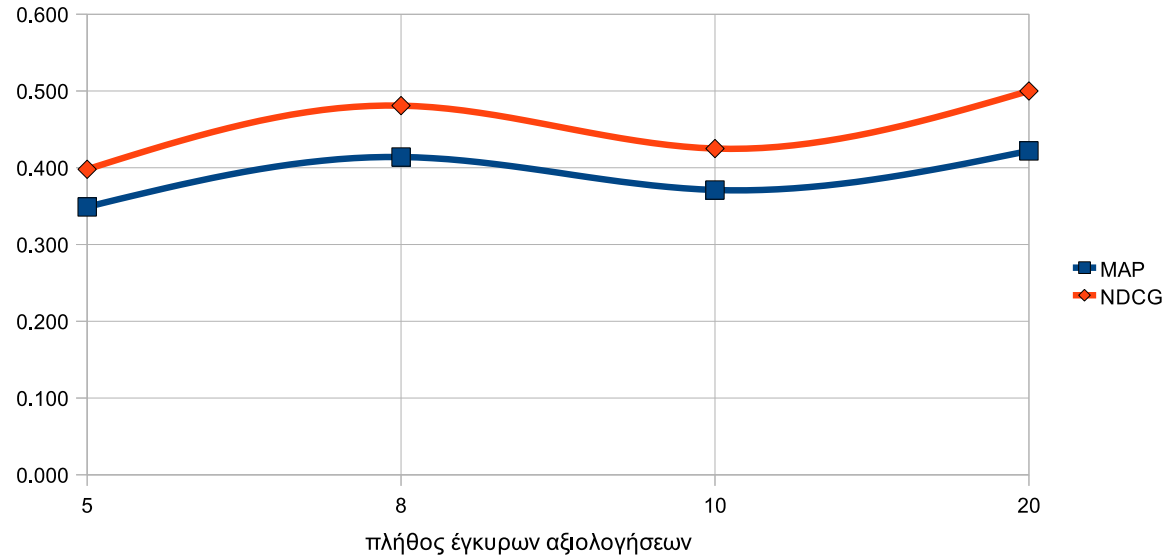
Κάτι αντίστοιχο συμβαίνει και στην περίπτωση των διανυσμάτων χαρακτηριστικών, που φαίνεται στον πίνακα 5.6. Εκεί όμως, η καμπύλη της απόδοσης φαίνεται να παρουσιάζει ένα τοπικό μέγιστο γύρω στο 8 το οποίο εν συνεχεία μειώνεται για να αυξηθεί πάλι μετά το 10 και να μεγιστοποιηθεί εκ νέου στο 20. Παρουσιάζοντας δηλαδή τα παραπάνω αποτελέσματα σε γραφική μορφή, θα είχαμε το ακόλουθο σχήμα 5.16



Σχήμα 5.15: απόκριση ακρίβειας ως προς το πλήθος των έγκυρων αξιολογήσεων n (textual)

5.5 Παρατηρήσεις

Μπορούμε να κάνουμε κάποιες γενικές παρατηρήσεις πάνω στη μορφή των παραπάνω σχημάτων. Αρχικά, παρατηρούμε ότι η τελική ακρίβεια του μοντέλου, εξαρτάται από το πλήθος των δεδομένων εκπαίδευσης όπως ήταν αναμενόμενο. Το αξιοσημείωτο είναι ότι παρά την αύξηση του πλήθους των δεδομένων, η βαρύτητα των λάθος επεκτάσεων φαίνεται να είναι πολύ μεγαλύτερη από αυτήν των δυο άλλων κατηγοριών (σωστών και μερικώς σωστών προβλέψεων). Πράγματι, παρατηρούμε ότι η αύξηση του πλήθους των δεδομένων εκπαίδευσης δεν συμπαρασύρει την καμπύλη ακρίβειας του μοντέλου προς τα πάνω. Αντίθετα, έστω και μικρή διακύμανση στο πλήθος των λάθος προβλέψεων φαίνεται να έχει μεγάλο πραγματικό αντίκτυπο στην απόδοση του μοντέλου. Αυτή η παρατήρηση είναι γενική για όλα τα παραπάνω



Σχήμα 5.16: απόκριση ακρίβειας ως προς το πλήθος των έγκυρων αξιολογήσεων n (fature)

διαγράμματα.

Από τα παραπάνω γραφήματα μπορούμε να εξάγουμε μια σημαντική ιδιότητα που φαίνεται να έχει η υβριδική μέθοδος: Παρατηρούμε ότι ενώ στους υπόλοιπους διανυσματικούς χώρους παρατηρείται αρκετά έντονη διακύμανση, στον υβριδικό χώρο, αυτή η διακύμανση έχει απορροφηθεί σε μεγάλο βαθμό. Βέβαια η σταθεροποίηση τοποθετείται χαμηλότερα από την μέγιστη απόκριση του χώρου λεκτικής περιγραφής αλλά αυτή η ιδιότητα είναι σημαντική.

Επίσης, παρατηρούμε ότι η διακύμανση της ακρίβειας στο σύνολο .GOV είναι πολύ μικρότερη, ακολουθώντας τη μικρή διακύμανση στο πλήθος των προβλέψεων κατά την επέκταση. Ακόμα όμως και για αυτή τη μικρή διακύμανση, συνεχίζει να ισχύει η σημαντική παρατήρηση σχετικά με τη βαρύτητα των λάθος δεδομένων εκπαίδευσης. Παράλληλα παρατηρούμε ότι για το δεύτερο σύνολο δεδομένων, η μέθοδός μας έχει γενικά πολύ καλή απόκριση καθώς το μεγαλύτερο ποσοστό των προβλέψεων (περίπου το 80%) είναι απόλυτα ορθές προβλέψεις. Βέβαια, για αυτό το δεύτερο σύνολο, οι παραλλαγές κατά την επέκταση (διαφοροποίηση των κατωφλιών κ1, κ2) φαίνεται να έχουν μικρό αντίκτυπο.

Αντίθετα στο σύνολο OHSUMED που δοκιμάστηκε πρώτο, οι παραλλαγές παίζουν βαρύνοντα ρόλο. Αυτό φαίνεται ιδιαίτερα στις περιπτώσεις όπου προκλήθηκε τεχνητή αύξηση του πλήθους με αναλογικά μεγάλη αύξηση των λαθών. Εκεί η απόκριση του μοντέλου κατά τον έλεγχό του με το σύνολο ελέγχου, έπεσε πολύ χαμηλά.

Σχετικά με το σύνολο OHSUMED (που λόγω της μεγαλύτερης διακύμανσης των αποτελεσμάτων, έχει μεγαλύτερο ενδιαφέρον) παρατηρούμε ότι το μέγιστο της απόκρισης ακρίβειας διαφοροποιείται από χώρο σε χώρο, σε ότι αφορά στα κατώφλια επέκτασης. Για παράδειγμα, στον χώρο των διανυσμάτων χαρακτηριστικών, βέλτιστη απόδοση φαίνεται να παρουσιάζει η επέκταση με κατώφλια (κ1, κ2) = (3, -100), ενώ στον χώρο της λεκτικής περιγραφής, το μέγιστο τοποθετείται στον συνδυασμό (κ1, κ2) = (2, -100). Τέλος ο υβριδικός χώρος πα-

ρουσιάζει κάποια χαρακτηριστικά εξομάλυνσης σε σχέση με τους άλλους δυο, και εντοπίζει τη μέγιστη ακρίβεια για κατώφλια (κ1, κ2) = (1, -100).

Τέλος, παρατηρούμε ότι η τεχνητή αύξηση που προκλήθηκε με την fully expanded προσέγγιση (βλ. πίνακα 4.3), έχει σαν αποτέλεσμα την μείωση της απόδοσης του παραγόμενου μοντέλου

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Μετά την υλοποίηση και τον διεξοδικό πειραματισμό μας με τις διαφορετικές παραμέτρους που τη συνοδεύουν μπορούμε να αναφέρουμε επιγραμματικά, τα εξής συμπεράσματα:

- Η μέθοδός μας με την κατάλληλη παραμετροποίηση προσεγγίζει αρκετά την ιδανική περίπτωση. Συνολικά, έχουμε συγκρίνει τη μέθοδό μας με την ιδανική περίπτωση εκπαίδευσης καθώς δεν υπάρχει κάποια δουλειά ακριβώς αντίστοιχη ώστε να τη συγκρίνουμε με αυτήν. Έτσι θεωρούμε ότι σε μια ιδανική περίπτωση θα μπορούσαμε να εκπαιδεύσουμε ένα νευρωνικό δίκτυο με το σύνολο των αποτελεσμάτων των προηγούμενων αναζητήσεων, αν ο χρήστης είχε αξιολογήσει αυτό το σύνολο. Στην πράξη βέβαια κάτι τέτοιο είναι ανέφικτο. Με τη μέθοδό μας, χρησιμοποιούμε μόνο τα 10 πρώτα (κατά κανόνα) αποτελέσματα κάθε αναζήτησης και προσπαθούμε βάσει αυτών να εμπλουτίσουμε το σύνολο εκπαίδευσης. Με την κατάλληλη παραμετροποίηση είδαμε ότι η μέθοδός μας προσεγγίζει σε μεγάλο βαθμό την ιδανική, υπολειπόμενη μόλις κατά ένα 4.4%

- Κατά την επέκταση των αξιολογήσεων, πρέπει να δίνεται ιδιαίτερη προσοχή στην περίπτωση λανθασμένων στοιχείων του συνόλου εκπαίδευσης. Είναι προφανές ότι η επέκταση των αξιολογήσεων δεν μπορεί να είναι τέλεια. Σίγουρα θα ενέχει και κάποια λάθη. Είδαμε όμως διεξοδικά στο κεφάλαιο των αποτελεσμάτων, ότι τα λάθος δεδομένα εκπαίδευσης παίζουν βαρύνοντα ρόλο. Έτσι, η αύξηση των δεδομένων εκπαίδευσης πρέπει να γίνεται με προσοχή και με την κατάλληλη παραμετροποίηση, ώστε να μειώνεται όσο το δυνατόν το πλήθος των λανθασμένων προβλέψεων. Είναι γενικά προτιμότερο ένα μικρό σύνολο εκπαίδευσης με λίγα λάθος πρότυπα, παρά ένα μεγάλο που περιέχει πολλά λάθος πρότυπα, παρόλο που μπορεί να υπάρχουν ισόποσα σωστές προβλέψεις.

- Από τους χώρους συσταδοποίησης, αυτός της λεκτικής περιγραφής είναι πιο αποδοτικός. Αντίθετα παρατηρούμε μειωμένη απόδοση στον χώρο των διανυσμάτων χαρακτηριστικών. Αυτό οφείλεται στην επιλογή των διαστάσεων που δεν είναι επαρκώς κατάλληλη για να εκφράσει το κάθε αποτέλεσμα. Βέβαια επιλύοντας το πρόβλημα στον χώρο των διανυσμάτων χαρακτηριστικών, μειώνουμε κατά πολύ την πολυπλοκότητα του μοντέλου

μιας και αυτός ο χώρος έχει μόλις 45 ή 46 διαστάσεις σε αντίθεση με τον χώρο της λεκτικής περιγραφής που αποτελείται από μερικές χιλιάδες διαστάσεις. Αυτό έχει σαν συνέπεια το γεγονός ότι το κόστος (τόσο σε μνήμη όσο και σε χρόνο) της εκτέλεσης της μεθόδου μειώνεται κατά πολύ. Βέβαια πληρώνεται το ισοζύγιο σε ακρίβεια.

- Μια υποσχόμενη λύση είναι η προσέγγιση του υβριδικού χώρου. Σε αυτόν τα αποτελέσματα παρουσιάζουν μια αξιοσημείωτη σταθερότητα, ανεξάρτητα από την παραμετροποίηση. Όπως είδαμε, υπάρχουν πάρα πολλές παράμετροι που επηρεάζουν καίρια την απόδοση της μεθόδου μας. Αν μέσω του υβριδικού χώρου καταφέρουμε να ξεπεράσουμε αυτόν τον σκόπελο, τότε ίσως με μια ζυγισμένη παραλλαγή του υβριδίου μας να βελτιώσει την απόδοση της μεθόδου και να συνδυάσει την απλότητα του χώρου των διανυσμάτων χαρακτηριστικών, με την ακρίβεια του χώρου λεκτικής περιγραφής.

6.2 Μελλοντικές Επεκτάσεις

Τα αποτελέσματα αυτής της διπλωματικής μπορούν να αποτελέσουν το έναυσμα για περαιτέρω δουλειά πάνω στο αντικείμενο της βελτίωσης των αποτελεσμάτων μηχανών αναζήτησης, τουλάχιστον ως προς τέσσερις άξονες. Συγκεκριμένα αναφέρονται τα ακόλουθα:

- Αξιοποίηση των ιδιοτήτων του υβριδικού χώρου συσταδοποίησης που προτάθηκε και υλοποίηση παραλλαγών που βασίζονται σε **σταθμισμένο** συνδυασμό των δυο άλλων χώρων. Με αυτόν τον τρόπο μπορεί να γίνει προσπάθεια, ώστε η υβριδική μέθοδος να βελτιώσει την αποδοτικότητά της, διατηρώντας όμως τη σταθερότητα που παρουσιάζει έναντι της αλλαγής των παραμέτρων.

- Επεξεργασία των διανυσμάτων χαρακτηριστικών με τεχνικές εξόρυξης δεδομένων ώστε να επιλέγονται τα καλύτερα χαρακτηριστικά κατά περίπτωση.

- Επέκταση προσωποποίησης με χρήση *μοτίβων συμπεριφοράς χρήστη* για περαιτέρω προσωποποίηση της αναζήτησης.

- Έμμεση αξιολόγηση, τέλος, των αποτελεσμάτων με επιπλέον κριτήρια (όπως χρόνος παραμονής στην ιστοσελίδα, χρήση του πλήκτρου back για επιστροφή στα αποτελέσματα της αναζήτησης κ.α.)

Βιβλιογραφία

[1] the perl programming language: http://www.perl.org/.

[2] C. Cortes και V. Vapnik. Support-vector networks. machine learning, 20:273-297. 1995.

[3] T. Joachims. Support vector machine for ranking.

[4] T. Joachims. Optimizing search engines using clickthrough data. Στο *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and datamining*, 2002.

[5] Gearge Karypis. Cluto. Στο *WWW http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download*

[6] Gearge Karypis. doc2mat.pl. Στο *WWW http://glaros.dtc.umn.edu/gkhome/files/fs/sw/cluto/doc2mat.*

[7] Ying Zhao και George Karypis. Criterion functions for document clustering experiments and analysis. Στο *WWW http://cs.umn.edu/karypis/publications*, 2003.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

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

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

[34] 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.

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

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

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

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

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

Κεφάλαιο 7

Μεταφράσεις Ξένων όρων

Μετάφραση	Αγγλικός όρος
συσταδοποίηση	clustering
συστάδα	cluster
νευρωνικό δίκτυο	neural network
διανύσμα χαρακτηριστικών	feature vector
σύνολο δεδομένων	dataset
σύνολο εκπαίδευσης	training set
σύνολο επικύρωσης	validation set
ανάκτηση πληροφορίας	information retrieval
μηχανές διανυσμάτων υποστήριξης	support vector machines
προεπεξεργασία	preclustering
επέκταση	expansion
αξιολόγηση	judgment

Παράρτημα Α΄

Διανύσματα Χαρακτηριστικών

Column in Output	Description
1	TF(Term frequency) of body
2	TF of anchor
3	TF of title
4	TF of URL
5	TF of whole document
6	IDF(Inverse document frequency) of body
7	IDF of anchor
8	IDF of title
9	IDF of URL
10	IDF of whole document
11	TF*IDF of body
12	TF*IDF of anchor
13	TF*IDF of title
14	TF*IDF of URL
15	TF*IDF of whole document
16	DL(Document length) of body
17	DL of anchor
18	DL of title
19	DL of URL
20	DL of whole document
21	BM25 of body
22	BM25 of anchor
23	BM25 of title
24	BM25 of URL
25	BM25 of whole document
26	LMIR.ABS of body
27	LMIR.ABS of anchor
28	LMIR. ABS of title
29	LMIR. ABS of URL
30	LMIR.ABS of whole document
31	LMIR.DIR of body
32	LMIR.DIR of anchor
33	LMIR.DIR of title
34	LMIR.DIR of URL
35	LMIR.DIR of whole document
36	LMIR.JM of body
37	LMIR.JM of anchor
38	LMIR.JM of title
39	LMIR.JM of URL
40	LMIR.JM of whole document
41	PageRank
42	Inlink number
43	Outlink number
44	Number of slash in URL
45	Length of URL
46	Number of child page

Σχήμα Α΄.1: Διανύσματα Χαρακτηριστικών στη συλλογη .GOV.

ID	Feature Description
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ in 'title'
2	$\sum_{q_i \in q \cap d} \log\left(c(q_i, d) + 1\right)$ in 'title'
3	$\sum_{q_i \in q \cap d} \frac{c(q_i, d)}{\lvert d \rvert}$ in 'title'
4	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} + 1\right)$ in 'title'
5	$\sum_{q_i \in q \cap d} \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in 'title'
6	$\sum_{q_i \in q \cap d} \log\left(\log\left(\frac{\lvert C \rvert}{df(q_i)}\right)\right)$ in 'title'
7	$\sum_{q_i \in q \cap d} \log\left(\frac{\lvert C \rvert}{c(q_i, C)} + 1\right)$ in 'title'
8	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right) + 1\right)$ in 'title'
9	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in 'title'
10	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} \cdot \frac{\lvert C \rvert}{c(q_i, C)} + 1\right)$ in 'title'
11	BM25 score in 'title'
12	log(BM25 score) in 'title'
13	LMIR with DIR smoothing in 'title'
14	LMIR with JM smoothing in 'title'
15	LMIR with ABS smoothing in 'title'
16	$\sum_{q_i \in q \cap d} c(q_i, d)$ in 'abstract'
17	$\sum_{q_i \in q \cap d} \log\left(c(q_i, d) + 1\right)$ in 'abstract'
18	$\sum_{q_i \in q \cap d} \frac{c(q_i, d)}{\lvert d \rvert}$ in 'abstract'
19	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} + 1\right)$ in 'abstract'
20	$\sum_{q_i \in q \cap d} \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in 'abstract'
21	$\sum_{q_i \in q \cap d} \log\left(\log\left(\frac{\lvert C \rvert}{df(q_i)}\right)\right)$ in 'abstract'
22	$\sum_{q_i \in q \cap d} \log\left(\frac{\lvert C \rvert}{c(q_i, C)} + 1\right)$ in 'abstract'
23	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right) + 1\right)$ in 'abstract'
24	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in 'abstract'
25	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} \cdot \frac{\lvert C \rvert}{c(q_i, C)} + 1\right)$ in 'abstract'
26	BM25 score in 'abstract'
27	log(BM25 score) 'abstract'
28	LMIR with DIR smoothing in 'abstract'
29	LMIR with JM smoothing in 'abstract'
30	LMIR with ABS smoothing in 'abstract'
31	$\sum_{q_i \in q \cap d} c(q_i, d)$ in 'title + abstract'
32	$\sum_{q_i \in q \cap d} \log\left(c(q_i, d) + 1\right)$ in 'title + abstract'
33	$\sum_{q_i \in q \cap d} \frac{c(q_i, d)}{\lvert d \rvert}$ in 'title + abstract'
34	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} + 1\right)$ in 'title + abstract'
35	$\sum_{q_i \in q \cap d} \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in 'title + abstract'
36	$\sum_{q_i \in q \cap d} \log\left(\log\left(\frac{\lvert C \rvert}{df(q_i)}\right)\right)$ in 'title + abstract'
37	$\sum_{q_i \in q \cap d} \log\left(\frac{\lvert C \rvert}{c(q_i, C)} + 1\right)$ in 'title + abstract'
38	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right) + 1\right)$ in 'title + abstract'
39	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in 'title + abstract'
40	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i, d)}{\lvert d \rvert} \cdot \frac{\lvert C \rvert}{c(q_i, C)} + 1\right)$ in 'title + abstract'
41	BM25 score in 'title + abstract'
42	log(BM25 score) 'title + abstract'
43	LMIR with DIR smoothing 'title + abstract'
44	LMIR with JM smoothing in 'title + abstract'
45	LMIR with ABS smoothing in 'title + abstract'

Σχήμα Α΄.2: Διανύσματα Χαρακτηριστικών στη συλλογη .OHSUMED.