ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αποδοτικό Ξεφύλλισμα σε Ιατρικές Βάσεις Δεδομένων

(Efficient Browsing In Medical Databases)

Διπλωματική Εργασία

Σάββας Πέτρου

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Ευριπίδης Πετράκης (Αν. Καθηγητής) Επιβλέπων

Μιχαήλ Γ. Λαγουδάκης (Επίκουρος Καθηγητής)

Αικατερίνη Μανιά (Επίκουρη Καθηγήτρια)

Χανιά

Ιούλιος 2008

Περιεχόμενα

Κατάλογος Σχημάτων

Περίληψη

Το ξεφύλλισμα σε εκτεταμένες ιατρικές βάσεις δεδομένων είναι μία από τις μεθόδους για αναζήτηση πληροφορίας. Η αναζήτηση για κείμενα που περιέχουν συγκεκριμένους ιατρικούς όρους δεν επιστρέφει πάντα όλο το διαθέσιμο υλικό. Οι συγγραφείς των διαφόρων άρθρων χρησιμοποιούν τον δικό τους προσωπικό τρόπο αναφοράς και συχνά μια δική τους ορολογία. Αυτό καθιστά την αναζήτηση ένα σημαντικό ζήτημα καθώς ο όγκος των δεδομένων αυξάνεται συνεχώς με γοργούς ρυθμούς. Στην εργασία αυτή προτείνεται μια προσέγγιση στο πρόβλημα αυτό με την χρησιμοποίηση της πληροφορίας που δίνεται μέσω των παραπομπών.

Με την χρησιμοποίηση του γράφου παραπομπών μπορεί να εξαχθεί πληροφορία που αφορά την διασύνδεση άρθρων αγνοώντας το περιεχόμενό τους. Δύο άρθρα που δεν μοιράζονται κοινούς όρους δεν σημαίνει απόλυτα ότι δεν αναφέρονται στο ίδιο θέμα. Η χρησιμοποίηση ενός άρθρου για την συγγραφή ενός άλλου υποδηλώνει μια σχέση μεταξύ τους. Έχοντας τον γράφο παραπομπών μιας ιατρικής συλλογής κειμένων εκτελούνται διάφοροι μέθοδοι ομαδοποίησης για την εξαγωγή γενικών ομάδων. Στην συνέχεια γίνεται ταυτοποίηση των ομάδων αυτών με βάση το περιεχόμενο των κειμένων που τα απαρτίζουν. Τέλος εφαρμόζεται ένα μέτρο σημασιολογικής ομοιότητας μεταξύ των όρων που εισάγει ο χρήστης και των ομάδων από τις διάφορες μεθόδους ομαδοποίησης για την επιλογή και ταξινόμηση των αποτελεσμάτων.

Ευχαριστίες

Η εργασία αυτή έχει εκτελεστεί στα πλαίσια της διπλωματικής μου εργασίας. Θεωρώ υποχρέωσή μου να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέπων καθηγητή μου, κ.Ευριπίδη Πετράκη, για την ανάθεση της παρούσας εργασίας και για την βοήθεια και υποστήριξη που μου πρόσφερε καθ' όλη την διάρκεια της διεξαγωγής της. Επίσης θα ήθελα να ευχαριστήσω τον Άγγελο Ηλιαουτάκη και τον Ευθύμη Δρυμώνα για την πολύτιμή τους βοήθεια όταν την χρειάστηκα.

Τέλος θα ήθελα να ευχαριστήσω θερμά τους φίλους και την οικογένεια μου για την στήριξη και συμπαράστασή τους κατά τη διάρκεια της εκπόνησης αυτής της εργασίας.

Κεφάλαιο 1

Εισαγωγή

Η ανάκτηση πληροφορίας είναι ένας τομέας της πληροφορικής που βρίσκεται υπό μεγάλη ανάπτυξη τα τελευταία χρόνια. Ο μεγάλος όγκος πληροφορίας που διατίθεται τόσο στο διαδίκτυο όσο και σε άλλες πηγές είναι τεράστιος. Μέθοδοι για το αποδοτικό ξεφύλλισμα μέσα σε μεγάλες συλλογές κειμένων είναι ένας από τους πλέον πιο σημαντικούς σκοπούς των σημερινών μηχανών αναζήτησης. Η αναζήτηση με κλασσικά μοντέλα αναζήτησης (Vector Space Model [14], Boolean Model [15]), γίνεται με απλή λεξικογραφική σύγκριση όρων, κάτι που δεν είναι αρκετό για την εύρεση *σχετικής* πληροφορίας. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με την χρήση της πληροφορίας που δίνεται μέσα από τα ίδια τα κείμενα. Σήμερα όλες σχεδόν οι μηχανές αναζήτησης χρησιμοποιούν τους συνδέσμους που βρίσκονται στις ιστοσελίδες για την αναζήτηση σχετικών σελίδων. Η ύπαρξη συνδέσμων σε μια ιστοσελίδα υποδηλώνει συνήθως την θεματική της ομοιότητα με άλλες. Σκοπός της παρούσας εργασίας είναι να χρησιμοποιήσει μια παρόμοια προσέγγιση για την ομαδοποίηση μιας ιατρικής συλλογής κειμένων. Η χρήση των παραπομπών ενός άρθρου μπορούν να προσδιορίσουν σχέσεις μεταξύ ολόκληρων κειμένων. Μελετώντας τις παραπομπές ενός κειμένου εξάγεται σημαντική πληροφορία όσων αφορά την θεματική ενότητα των εμπλεκόμενων άρθρων. Η συγγραφή ενός άρθρου συνήθως συνιστά την χρησιμοποίηση άλλων άρθρων που αφορούν το ίδιο θέμα. Με την δημιουργία αυτών των θεματικών ομάδων μπορεί να επιτευχθεί αποδοτικότερο ξεφύλλισμα της συλλογής μέσω της έξυπνης ομαδοποίησης σχετικών άρθρων.

Αρχικά, με την χρήση των παραπομπών δημιουργείται ο *γράφος παραπομπών (citation*

graph). Ο γράφος αποτελεί μια αναπαράσταση της συνδεσιμότητας των κειμένων μέσα στη συλλογή. Σκοπός της δημιουργίας του είναι η ομαδοποίηση των άρθρων με βάση την συνδεσιμότητά τους. Για την ομαδοποίηση έχουν χρησιμοποιηθεί δύο αλγόριθμοι *ανάλυσης συνδέσμων (link analysis)* και ένας κλασσικός αλγόριθμος με χρησιμοποίηση ομοιότητας περιεχομένου. Οι αλγόριθμοι είναι οι εξής :

- Edge-Betweenness : Η ιδέα του αλγορίθμου βασίζεται στην εφαρμογή ενός μέτρου *κεντρικότητας* για την επιλογή και αφαίρεση ακμών. Γειτονικές ομάδες σε ένα γράφο ενώνονται συνήθως χαλαρά με πολύ λίγες σε αριθμό ακμές. Αφαιρώντας τις ακμές δημιουργούνται ανεξάρτητοι, ασύνδετοι γράφοι που αποτελούν την ομαδοποίηση.

- Markov Clustering Algorithm : Ο αλγόριθμος βασίζεται στην θεωρία ροών (flows) μέσα σε ένα γράφο. Αρχικά δίνεται ένα βάρος σε κάθε ακμή του γράφου. Με τις επαναληπτικές εκτελέσεις του αλγορίθμου τα βάρη μεταβάλλονται έτσι ώστε η ροή να μεταφερθεί από τις αραιές περιοχές στις πιο πυκνές. Οι επαναλήψεις σταματάνε όταν η ροή στον γράφο σταθεροποιηθεί [16].

- BIC-Means : Οι πιο πάνω αλγόριθμοι χρησιμοποιούν τον γράφο για να εξάγουν την ομαδοποίηση. Η συγκεκριμένη μέθοδος είναι μια κλασσική μέθοδος ομαδοποίησης που χρησιμοποιεί ομοιότητα περιεχομένου για την εξαγωγή των ομάδων. Είναι μια παραλλαγή του γνωστού αλγόριθμου K-Means. Βασικό της χαρακτηριστικό είναι η ύπαρξη μέτρου για την απόφαση διάσπασης μιας ομάδας. Ο αλγόριθμος δεν είναι εξαντλητικός και εκτιμά την ποιότητα των ομάδων καθώς τις δημιουργεί με βάση το Bayesian Information Criterion (BIC) [11].

Στη συνέχεια της εργασίας γίνεται μια ταυτοποίηση των ομάδων (Topic Distillation). Για την επίτευξη αυτού του βήματος χρησιμοποιήθηκαν ιατρικοί όροι για την περιγραφή των κειμένων. Οι όροι εξάχθηκαν με την μέθοδο AMTEx [1]. Η μέθοδος χρησιμοποιεί τον θησαυρό ορολογίας MeSH και την τεχνική εξαγωγής πολυλεκτικών όρων C/NC-

value [5] για την επιλογή ιατρικών όρων που περιγράφουν ένα κείμενο. Τέλος, έχοντας τις ταυτότητες των ομάδων και το άνυσμα από όρους που δίνει ο χρήστης, εφαρμόζεται ένα μέτρο ομοιότητας για την επιλογή και ταξινόμηση των ομάδων. Όπως αναφέρθηκε προηγουμένως τα κλασσικά μέτρα ομοιότητας αγνοούν το σημασιολογικό περιεχόμενο των όρων και εφαρμόζουν απλή λεξικογραφική ομοιότητα. Μια πιο έξυπνη προσέγγιση στο θέμα της ομοιότητας ανυσμάτων είναι η σημασιολογική ομοιότητα (Semantic Similarity). Η τεχνική Rada Mihalcea [12] είναι μια από τις τεχνικές που χρησιμοποιεί σημασιολογική ομοιότητα και είναι η τεχνική που χρησιμοποιείτε και στην παρούσα εργασία.

Το υπόλοιπο κείμενο είναι διαχωρισμένο ως εξής:

Κεφάλαιο 2 - Υπόβαθρο : Επεξηγούνται οι αλγόριθμοι και μέθοδοι που χρησιμοποιήθηκαν για τις ομαδοποιήσεις, την εξαγωγή ιατρικών όρων, την ταυτοποίηση των ομάδων και την σημασιολογική ομοιότητα.

Κεφάλαιο 3 - Σχετικές Εργασίες : Αναφέρονται κάποια ήδη υπάρχον προγράμματα που βασίζονται στην ιδέα της χρήσης του γράφου παραπομπών για την εύρεση σχετικών κειμένων.

Κεφάλαιο 4 - Υλοποίηση Συστήματος : Αναλυτική περιγραφή των βημάτων που ακολουθήθηκαν για την υλοποίηση του συστήματος.

Κεφάλαιο 5 - Αποτελέσματα - Μελλοντική Εργασία : Αναφέρονται τα αποτελέσματα και τα συμπεράσματα που εξάχθηκαν από την εργασία. Ακόμη αναφέρονται σημεία στα οποία υπάρχει μελλοντική εργασία.

Παράρτημα Α΄ : Τεχνικές πληροφορίες που αφορούν κυρίως τις βιβλιοθήκες Java που χρησιμοποιήθηκαν και τα εργαλεία για αποθήκευση των ευρετηρίων και του γράφου παραπομπών.

Κεφάλαιο 2

Υπόβαθρο

Σε αυτό το κεφάλαιο θα επεξηγηθούν οι σχετικές εργασίες που χρησιμοποιούνται για την διεξαγωγή της τρέχουσας εργασίας. Αναλυτικά θα επεξηγηθούν οι μέθοδοι ομαδοποίησης του γράφου παραπομπών (Citation Graph Clustering), η μέθοδος για την εξαγωγή όρων από τα κείμενα και η μέθοδος για την σημασιολογική ομοιότητα ανυσμάτων όρων.

2.1 Μέθοδοι Ομαδοποίησης (Clustering)

Για την ομαδοποίηση του γράφου παραπομπών (Citation Graph) χρησιμοποιήθηκαν τρεις μέθοδοι. Οι μέθοδοι *Edge-Betweenness* και *Markov Clustering Algorithm* χρησιμοποιούν Ανάλυση Συνδέσμων (Link Analysis) ενώ η μέθοδος *BIC-Means* Ομοιότητα Περιεχομένου (Content Similarity).

2.1.1 Edge-Betweenness Clustering

Η μέθοδος αυτή είναι γνωστή σαν Girvan-Newman [8]. Η κεντρική ιδέα της μεθόδου βασίζεται στην εύρεση ακμών οι οποίες είναι λιγότερο κεντρικές σε ένα γράφο και στην αφαίρεσή τους. Τις ακμές δηλαδή που είναι **μεταξύ** (between) ομάδων. Η ομαδοποίηση του γράφου επιτυγχάνεται με την σταδιακή αφαίρεση ακμών από τον αρχικό γράφο. Ο συνολικός αριθμός των ακμών που αφαιρούνται εξαρτάται από διάφορα κριτήρια ανάλογα με την υλοποίηση. Αυτά τα κριτήρια μπορεί να είναι ο τελικός αριθμός ομάδων ή κάποιος συγκεκριμένος αριθμός ακμών.

Η αρχική προσέγγιση της έννοιας *Betweenness* μελετήθηκε στο παρελθόν σαν ένα μέτρο *κεντρικότητας* (centrality). Δηλαδή κατά πόσο επηρεάζει ένας κόμβος κάποιο δίκτυο. Για κάθε κόμβο το μέτρο *vertex betweenness* ορίζεται σαν ο αριθμός από συντομότερα μονοπάτια μεταξύ ζευγαριών κόμβων, που περνάνε από τον συγκεκριμένο κόμβο. Ο αλγόριθμος Girvan-Newman επεκτείνει αυτό τον ορισμό στην περίπτωση των ακμών. Ορίζει σαν *edge betweenness* τον αριθμό από συντομότερα μονοπάτια μεταξύ ζευγαριών από κόμβους, που περνάνε από την συγκεκριμένη ακμή. Σε περίπτωση που ένα ζευγάρι κόμβων έχει περισσότερα από ένα συντομότερα μονοπάτια, τότε κάθε μονοπάτι παίρνει από ένα βάρος, τέτοιο ώστε το άθροισμα του βάρους όλων τον μονοπατιών να είναι ίσο με μονάδα.

Σε ένα γράφο που περιέχει κοινότητες ή ομάδες που ενώνονται χαλαρά με πολύ λίγες ακμές μεταξύ τους, όλα τα μονοπάτια μεταξύ των διαφορετικών ομάδων περνάνε πάνω από αυτές τις λίγες ακμές. Κατά συνέπεια, οι ακμές που ενώνουν τις ομάδες θα έχουν μεγάλο *edge betweenness*. Αφαιρώντας τις ακμές αυτές, οι ομάδες θα διαχωριστούν και θα δημιουργηθεί η ομαδοποίηση του γράφου [10].

Η εκτέλεση του αλγόριθμου γίνεται σε δύο μόνο βήματα που επαναλαμβάνονται :

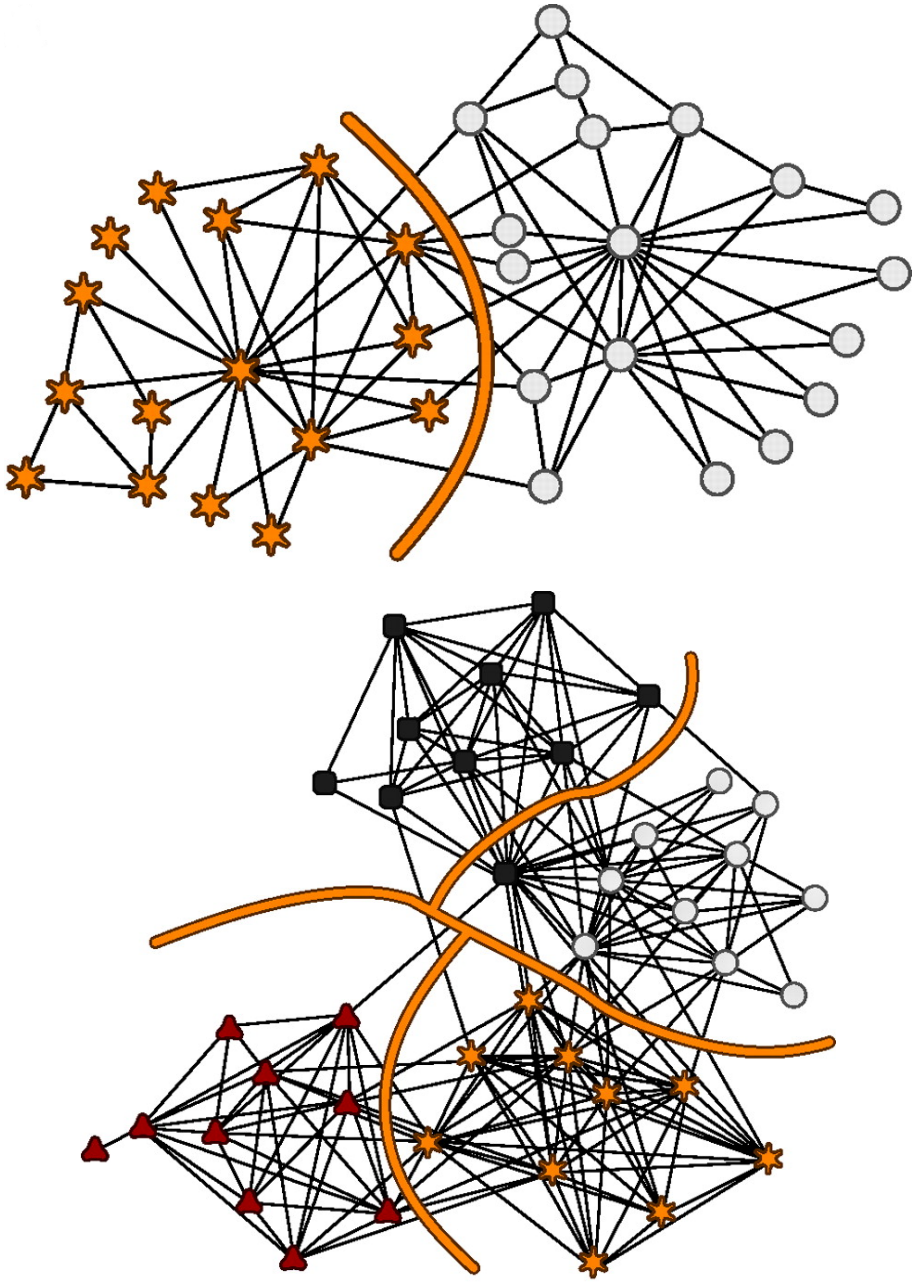
1. Υπολογισμός της *κεντρικότητας* όλων των ακμών. Αυτό επιτυγχάνεται με την εύρεση των συντομότερων μονοπατιών για όλα τα ζεύγη κόμβων.
2. Αφαίρεση της ακμής με την μεγαλύτερη *κεντρικότητα*

Σχήμα 2.1: Βήματα εκτέλεσης αλγόριθμου Edge-Betweenness

Η υπολογιστική πολυπλοκότητα του αλγόριθμου είναι $O(kmn)$, όπου k ο αριθμός των ακμών που θέλουμε να αφαιρεθούν, m ο συνολικός αριθμός ακμών και n ο συνολικός αριθμός κόμβων. Σε περίπτωση που ο γράφος είναι *αραιός* (sparse) η υπολογιστική πολυπλοκότητα πλησιάζει στο $O(kn^2)$ και για γράφους με δυνατή δομή ομάδων επιτυγχάνεται ακόμη μικρότερη πολυπλοκότητα [9].

Στο Σχήμα 2.2 φαίνεται μια αναπαράσταση του διαχωρισμού των ομάδων με την

μέθοδο. Οι ακμές που τέμνονται από την διαχωριστική γραμμή είναι οι ακμές που θα αφαιρεθούν μετά από επαναληπτικές εκτελέσεις του αλγορίθμου. [1].



Σχήμα 2.2: Παράδειγμα ομαδοποίησης

[1] http://www.pnas.org/cgi/content/full/104/18/7327

2.1.2 Markov Cluster Algorithm (MCL)

Ο αλγόριθμος είναι γρήγορος, εξελικτικός (scalable) και ανεπίβλεπτος (unsupervised). Η ιδέα της λειτουργίας του βασίζεται στην προσομοίωση της ροής σε ένα γράφο. Δημιουργήθηκε από τον Stijn van Dongen στο Centre for Mathematics and Computer Science στην Ολλανδία.

Η μεθοδολογία του συγκεκριμένου αλγορίθμου είναι αρκετά απλοϊκή και βασίζεται σε δύο απλές αλγεβρικές πράξεις πινάκων. Η αρχή λειτουργίας του απλή και ελκυστική. Δεν υπάρχουν διαδικασίες υψηλού επιπέδου για την δημιουργία, συνένωση ή διάσπαση των ομάδων.

Η πρώτη πράξη που χρησιμοποιεί ο αλγόριθμος είναι η *επέκταση* (expansion). Σε μαθηματικό επίπεδο η πράξη αυτή συμπίπτει με την λειτουργία του πολλαπλασιασμού πινάκων. Η επέκταση μοντελοποιεί την **διάδοση** της ροής μέσα στο γράφο. Η δεύτερη πράξη είναι ο *πληθωρισμός* (inflation). Αντίστοιχη μαθηματική πράξη σε πίνακες είναι η ύψωση του πίνακα σε μια δύναμη και στη συνέχεια η διαγώνια κλιμάκωση (diagonal scaling). Ο *πληθωρισμός* μοντελοποιεί την **συστολή** της ροής μέσα στο γράφο. Η μέθοδος αυτή προκαλεί την συσσώρευση της ροής μέσα σε φυσικές ομάδες και την αραίωση της ροής μεταξύ διαφορετικών ομάδων.

Ο κεντρικός σκελετός του επαναληπτικού αλγόριθμου είναι διαμορφωμένος με την εναλλαγή του πολλαπλασιασμού και πληθωρισμού ενός πίνακα σε ένα βρόγχο επανάληψης. Στην k-οστή επανάληψη αυτού του βρόγχου υπολογίζονται δύο πίνακες με ετικέτες T_{2k} και T_{2k+1}. Ο πίνακας T_{2k} υπολογίζεται με την ύψωση του πίνακα T_{2k-1} σε μια δύναμη e_k. Ο πίνακας T_{2k+1} υπολογίζεται με την απεικόνιση του πίνακα T_{2k} στο Γ_{rk}.

Ορισμός 1.

*Δεδομένου ενός πίνακα $M \in \Re^{k \times l}, M \geq 0$ και ενός μη αρνητικού αριθμού r, ο πίνακας που προκύπτει από την κλιμάκωση της κάθε στήλης του M με τον συντελεστή δύναμης r καλείται $\Gamma_r M$, και Γ_r καλείται **συντελεστής πληθωρισμού** σε δύναμη r. Ο τύπος*

που δίνει το $\Gamma_r : \Re^{k\times l} \to \Re^{k\times l}$ *είναι :*

$$(\Gamma_r M_{pq}) = (M_{pq})^r / \sum_{i=1}^{k} (M_{iq})^r) \tag{2.1}$$

όπου $p, 1 \leq p \leq k$ *και* $q, 1 \leq q \leq l$ *είναι δείκτες στα στοιχεία του πίνακα και* r *μια μη μηδενική τιμή.*

Πιο κάτω φαίνεται ένα παράδειγμα εκτέλεσης της πράξης του πληθωρισμού σε ένα πίνακα άνυσμα. Με βάση την ανάλυση της απόδειξης της μεθόδου [16], έχουν επιλεγεί οι τιμές $e = 2$ και $r = 2$.

$$\begin{pmatrix} 0 \\ 3 \\ 0 \\ 1 \\ 2 \end{pmatrix} \begin{matrix} \to 0^2/(0^2+3^2+0^2+1^2+2^2) \\ \to 3^2/(0^2+3^2+0^2+1^2+2^2) \\ \to 0^2/(0^2+3^2+0^2+1^2+2^2) \\ \to 1^2/(0^2+3^2+0^2+1^2+2^2) \\ \to 2^2/(0^2+3^2+0^2+1^2+2^2) \end{matrix} \begin{pmatrix} 0 \\ 9/14 \\ 0 \\ 1/14 \\ 4/14 \end{pmatrix}$$

Τα τελευταία στοιχεία για την συμπλήρωση των βημάτων του αλγορίθμου είναι ο τρόπος δημιουργίας και αρχικοποίησης του αρχικού πίνακα και η εξαγωγή της ομαδοποίησης από τον τελικό πίνακα. Για να γίνει αυτό θα οριστούν κάποιες έννοιες σημαντικές για την κατανόηση των βημάτων.

Ορισμός 2. *Θέτουμε* $G = G(V, w)$ *ένα μη αρνητικό, άκυκλο, ταυτοδύναμο (idempotent)* [2] *γράφο τάξης* N*, όπου* $V = \{1, ..., N\}$ *το σύνολο των κόμβων. Ο κόμβος* $\alpha \in V$ *καλείτε* ***κόμβος έλξης*** *εάν* $M_{\alpha\alpha} \neq 0$*. Αν το* α *είναι* ***κόμβος έλξης*** *τότε και ο ομάδα των γειτόνων του καλείται* ***σύστημα έλξης****.*

Ορισμός 3. *Θέτουμε* M *μια μη αρνητική στήλη του πίνακα. Θέτουμε* $G = (V, w)$ *τον γράφο. Θέτουμε* V_x *το σύνολο των* ***κόμβων έλξης*** *(Ορισμός 2) μεγέθους* d*. Θέτουμε* $E = \{E_1, ..., E_d\}$ *τις ομάδες του γράφου. Ορίζουμε την σχέση* ν *στο* $E \times V$ *σαν* $\nu(E, \alpha) = 1$ *αν* $\exists \beta \in E$ *για το οποίο υπάρχει ακμή που ενώνει* α *με* β*, και* $\nu(E, \alpha) = 0$

[2] http://en.wikipedia.org/wiki/Idempotent_matrix

αλλιώς. Το σύνολο των ομάδων $CL_M = \{C_1, ..., C_d\}$ *που σχετίζονται με τον πίνακα, έχει* d *στοιχεία. Η* i*-οστή ομάδα* $C_i, i = 1, ..., d$ *ορίζεται από τον πιο κάτω τύπο :*

$$C_i = \{\upsilon \in V \mid \nu(E_i, \upsilon) = 1\} \tag{2.2}$$

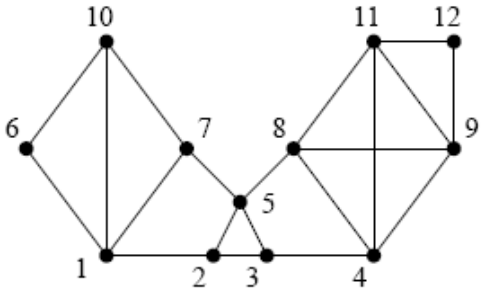
Η δημιουργία και αρχικοποίηση του αρχικού πίνακα γίνεται ως εξής. Όπως φαίνεται από το Σχήμα 2.3 αρχικά γίνεται αρίθμηση των κόμβων και στη συνέχεια δημιουργία ενός τετραγωνικού πίνακα τάξης ίσης με τον συνολικό αριθμό κόμβων του γράφου. Η i-οστή στήλη συμπληρώνεται θέτοντας την τιμή *1/(αριθμός ακμών* i*-οστού κόμβου + 1)*, στις γραμμές της στήλης που υποδηλώνουν ότι υπάρχει ακμή με τον i-οστό κόμβο. Το *+1* αντιπροσωπεύει τον επαναληπτικό βρόγχο του κόμβου (loop). Η κύρια διαγώνιος είναι μη μηδενική και παίρνει την ίδια τιμή με τα υπόλοιπα στοιχεία της στήλης λόγω του επαναληπτικού βρόγχου.

Χρησιμοποιώντας τον Ορισμό 3 στον τελικό πίνακα παίρνουμε την ομαδοποίηση. Οι ομάδες είναι οι εξής :

$$\begin{aligned} C_1 &= \{\mathbf{1}, 6, 7, 10\} \\ C_2 &= \{\mathbf{5}, 2, 3\} \\ C_3 &= \{\mathbf{9}, 4, 8, \mathbf{11}, 12\} \\ C_4 &= \{\mathbf{11}, 4, 8, \mathbf{9}, 12\} \end{aligned}$$

Παρατηρείται ότι οι δύο τελευταίες ομάδες είναι όμοιες. Αυτό συμβαίνει γιατί η ομάδα αυτή διαθέτει δύο *κόμβους έλξης*. Τον κόμβο 9 και 11. Δεχόμαστε μια από τις δύο ομάδες και αγνοούμε την άλλη. Οι κόμβοι έλξης των υπολοίπων ομάδων φαίνονται με έντονα γράμματα.

Στο Σχήμα 2.3 φαίνονται όλα τα βήματα που περιγράφηκαν μέχρι στιγμής.



Γράφος

$$\begin{pmatrix}
0.200 & 0.250 & -- & -- & -- & 0.333 & 0.250 & -- & -- & 0.250 & -- & -- \\
0.200 & 0.250 & 0.250 & -- & 0.200 & -- & -- & -- & -- & -- & -- & -- \\
-- & 0.250 & 0.250 & 0.200 & 0.200 & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & 0.250 & 0.200 & -- & -- & -- & 0.200 & 0.200 & -- & 0.200 & -- \\
-- & 0.250 & 0.250 & -- & 0.200 & -- & 0.250 & 0.200 & -- & -- & -- & -- \\
0.200 & -- & -- & -- & -- & 0.333 & -- & -- & -- & 0.250 & -- & -- \\
0.200 & -- & -- & -- & 0.200 & -- & 0.250 & -- & -- & 0.250 & -- & -- \\
-- & -- & -- & 0.200 & 0.200 & -- & -- & 0.200 & 0.200 & -- & 0.200 & -- \\
-- & -- & -- & 0.200 & -- & -- & -- & 0.200 & 0.200 & -- & 0.200 & 0.333 \\
0.200 & -- & -- & -- & -- & 0.333 & 0.250 & -- & -- & 0.250 & -- & -- \\
-- & -- & -- & 0.200 & -- & -- & -- & 0.200 & 0.200 & -- & 0.200 & 0.333 \\
-- & -- & -- & -- & -- & -- & -- & -- & 0.200 & -- & 0.200 & 0.333
\end{pmatrix}$$

Αρχικός πίνακας

$$\begin{pmatrix}
1.000 & -- & -- & -- & -- & 1.000 & 1.000 & -- & -- & 1.000 & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & 1.000 & 1.000 & -- & 1.000 & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & 0.500 & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & 0.500 & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & --
\end{pmatrix}$$

Τελικός πίνακας

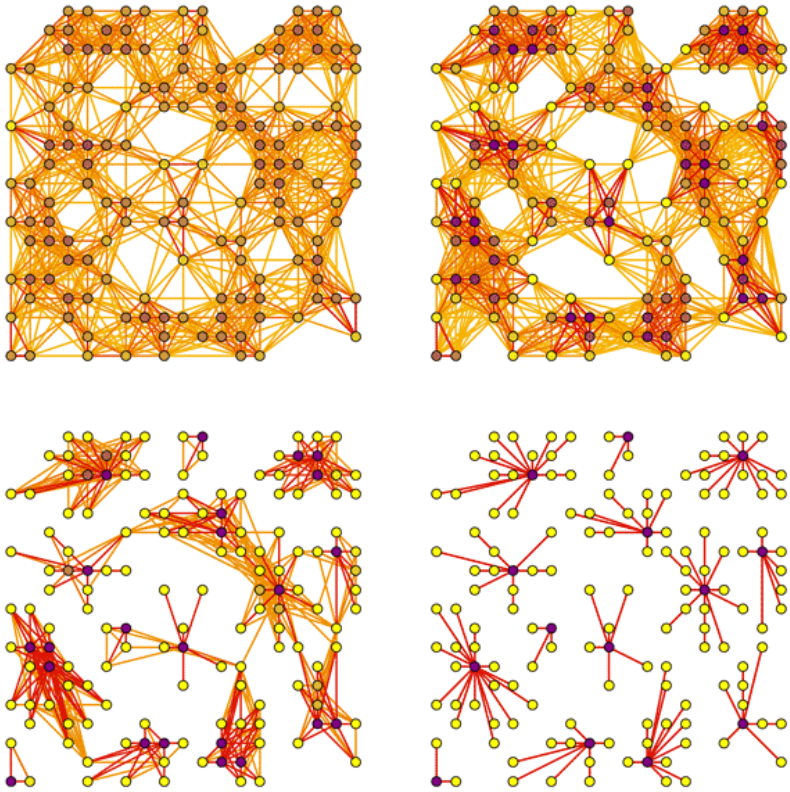
Σχήμα 2.3: Παράδειγμα ομαδοποίησης Markov

Τα βασικά βήματα εκτέλεσης του αλγορίθμου φαίνονται στο Σχήμα 2.4 :

1. **Βήμα 1 :** Δημιουργία αρχικού πίνακα και αρχικοποίηση τιμών.
2. **Βήμα 2 :** Ύψωση του πίνακα σε δύναμη e (στην περίπτωσή μας αυτή η δύναμη είναι ίση με 2). Αυτός είναι ο πίνακας με ετικέτα T_{2k}.
3. **Βήμα 3 :** Υπολογισμός της απεικόνισης του πίνακα T_{2k}. Ο πίνακας αυτός έχει ετικέτα T_{2k+1}
4. **Βήμα 4 :** Έλεγχος πίνακα T_{2k+1} αν είναι ταυτοδύναμος. Σε περίπτωση ταυτοδυναμίας εκτελείται το Βήμα 5 αλλιώς επαναλαμβάνονται τα βήματα 2 με 4.
5. **Βήμα 5 :** Εξαγωγή ομαδοποίησης.

Σχήμα 2.4: Βήματα εκτέλεσης αλγόριθμου Markov Clusterer

Στο Σχήμα 2.5 φαίνεται μια γραφική αναπαράσταση του αλγορίθμου.



Σχήμα 2.5: Markov flow pictorial

2.1.3 BIC-Means

Στις παραπάνω μεθόδους χρησιμοποιήθηκε Ανάλυση Συνδέσμων για την δημιουργία της ομαδοποίησης. Η συγκεκριμένη μέθοδος είναι μια παραλλαγή της μεθόδου 'K-Μέσων' (K-Means) και έχει υλοποιηθεί από τον Νικόλαο Χουρδάκη στα πλαίσια της μεταπτυχιακής του εργασίας στο Πολυτεχνείο Κρήτης [11].

Στη νέα αυτή προσέγγιση χρησιμοποιείται η *Επαυξητική μέθοδος K-μέσων (Incremental K-Means)*, για ανανέωση του κέντρου (centroid) μιας ομάδας μόλις ένα κείμενο προστεθεί σε αυτήν. Η τελική μέθοδος "BIC-Means" είναι ιεραρχική, δηλαδή δημιουργείται μια ιεραρχία από ομάδες, κειμένων στη συγκεκριμένη περίπτωση, εφαρμόζοντας επαναληπτικά την επαυξητική μέθοδο K-μέσων στη συλλογή κειμένων PMC. Σημαντικό πλεονέκτημα της μεθόδου είναι το ότι δεν είναι εξαντλητική. Με λίγα λόγια δεν είναι αναγκαία η επανάληψη των βημάτων έως ότου οι ομάδες να περιέχουν μόνο ένα κείμενο (singleton clusters). Για την επίτευξη αυτού, ο BIC-Means ενσωματώνει το Bayesian Information Criterion (BIC) ή Schwarz Criterion. Χρησιμοποιείται για να σταματήσει τις διασπάσεις των ομάδων σε ανώτερα επίπεδα της ιεραρχίας όταν η περαιτέρω διάσπασή τους δεν οδηγεί σε καλύτερη ομαδοποίηση. Η μέθοδος τερματίζει όταν εξεταστούν όλες οι υποψήφιες προς διάσπαση ομάδες και δεν υπάρχει άλλη υποψήφια.

Bayesian Information Criterion (BIC)

Ο υπολογισμός του κριτηρίου BIC γίνεται με τον πιο κάτω τρόπο.

Ορίζεται σαν D η συλλογή από κείμενα $\{x_1, x_2, ...\}$. Το D μπορεί να χωριστεί σε $D_1, D_2, ..., D_k$ ανεξάρτητα κομμάτια. Στην περίπτωση του Bisecting K-Means το $K = 2$. Ορίζεται ότι μ_j είναι το κέντρο της j^{th} ομάδας $(i < j < K)$. Ορίζεται ότι το (i) είναι ο δείκτης του κέντρου της ομάδας που είναι πιο κοντά στο (i) κείμενο. Ορίζεται σαν $D_j \subseteq D$ η ομάδα από κείμενα που έχουν μ_j το κοντινότερο τους κέντρο. Ορίζεται $R = |D|$ και $R_j = |D_j|$. Ο αριθμός των διαστάσεων είναι M. Παρατηρείται ότι η αρχική συλλογή των κειμένων χωρίζεται σε δύο ομάδες.

Το κριτήριο BIC για το μοντέλο M_j, στη περίπτωσή μας η ομάδα του πατέρα ή οι δύο ομάδες παιδία, δίνεται από την εξίσωση :

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} logR \tag{2.3}$$

όπου το $\hat{l}_j$ είναι το log-likelihood των κειμένων σύμφωνα με το μοντέλο M_j, δεδομένου ότι $p_j = K(M+1)$ είναι ο αριθμός των ανεξάρτητων παραμέτρων του M_j.

Σύμφωνα με την Εξίσωση 2.3 το κριτήριο αποτελείται από δύο μέτρα. Το πρώτο μέτρο είναι το log-likelihood που χρησιμοποιείται σαν μέτρο για την συνεκτικότητα της ομάδας, κάτι που μπορεί να υποδείξει στην συνέχεια αν μια ομάδα πρέπει να διασπαστεί ή όχι. Με το μέτρο αυτό υπολογίζεται ακόμη πόσο κοντα στο κέντρο βρίσκονται τα κείμενα σε μια συγκεκριμένη ομάδα. Πιο αναλυτικά, δεδομένης κάποιας ομάδας από κείμενα, δοσμένα από μια Gaussian κατανομή $N(\mu, \sigma^2)$, το log-likelihood είναι η πιθανότητα μια γειτονιά από κείμενα να ακολουθούν αυτή την κατανομή. Το δεύτερο μέτρο βοηθάει στην διόρθωση της κατανομής των κειμένων. Λόγω της πολυπλοκότητα του μοντέλου που χρησιμοποιήθηκε, μερικά από τα κείμενα της ομάδας, εκτός από την Gaussian κατανομή μπορεί να ακολουθούν και άλλες κατανομές. Για το λόγο αυτό το δεύτερο μέτρο τείνει να φτιάξει το πρόβλημα.

Το Maximum Likelihood Estimate (MLE) για την διακύμανση δίνεται από:

$$\hat{\sigma}^2 = \frac{1}{R_n - K} \sum_i ||x_i - \mu_{(i)}||^2 \tag{2.4}$$

όπου R_n δίνει τον αριθμό των κειμένων στην ομάδα D_n. Δεδομένης μιας ομάδας από κείμενα, $P_{(x_i)}$ είναι η πιθανότητα το κείμενο x_i να ακολουθεί την Gaussian κατανομή $N(\mu, \sigma^2)$ που δίνεται από την ομάδα που ανήκει.

$$P(x_i) = \frac{R_n}{R} \frac{1}{\sqrt{2\pi}\hat{\sigma}^M} exp(-\frac{1}{2\hat{\sigma}^2}||x_i - \mu_{(i)}||^2) \tag{2.5}$$

Κατά συνέπεια, το log-likelihood των κειμένων σε μια ομάδα μπορεί να υπολογιστεί με βάση τον λογάριθμο του αποτελέσματος του πολλαπλασιασμού των πιθανοτήτων των κειμένων που το αποτελούν.

$$\begin{aligned} \hat{l}(C_i) &= log \prod_i \hat{P}(x_i) \\ &= \sum_i (log \frac{1}{\sqrt{2\pi}\hat{\sigma}^d} - \frac{1}{2\hat{\sigma}^2}||x_i - \mu_{(i)}||^2 + log \frac{R}{R_n}) \\ &= -\frac{R_n}{2} log(2\pi) - \frac{R_n M}{2} log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n log(R_n) - R_n log(R) \end{aligned} \tag{2.6}$$

Για να επεκταθεί ο τύπος στην Εξίσωση 2.3 για όλα τα κέντρα και όχι μόνο για ένα, χρησιμοποιείται το γεγονός ότι το log-likelihood όλων των κειμένων που ανήκουν σε κάποιο κέντρο είναι το άθροισμα των επιμέρους log-likelihood κέντρων αυτών. Κατά συνέπεια η Εξίσωση 2.3 μετατρέπεται στην πιο κάτω :

$$BIC(M_j) = \sum_{j=1}^{K} \hat{l}(C_j) - \frac{p_j}{2} log(R) \tag{2.7}$$

Ο αριθμός των ανεξάρτητων παραμέτρων του p_j είναι το άθροισμα των πιο κάτω :

- $K - 1$ πιθανότητα κλάσης
- $M * K$ συντεταγμένες κέντρων
- ένα για την εκτίμηση της διακύμανσης

Με την διακύμανση (Εξίσωση 2.4) υπολογίζεται ο μέσος όρος του τετραγώνου της απόστασης του κάθε κειμένου από το κέντρο της ομάδας. Αυτό είναι ένα μέτρο της συνεκτικότητας της ομάδας. Με τον υπολογισμό του BIC εκτιμάται το πόσο κοντά στο κέντρο είναι τα κείμενα που ανήκουν σε μια συγκεκριμένη ομάδα.

Τα βήματα του αλγορίθμου για την δημιουργία της ομαδοποίησης είναι τα πιο κάτω:

- **Δεδομένα :** $K = 2$ για το Incremental K-Means $S : (d_1, d_2, ..., d_n)$.
- **Έξοδος :** Μια ιεραρχική ομαδοποίηση της συλλογής.
- **Βήμα 1 :** Ενώνονται όλα τα κείμενα σαν αρχική ενιαία ομάδα. Προστίθεται η ομάδα στη λίστα με τις ομάδες που περιμένουν για διάσπαση.
- **Βήμα 2 :** Επιλέγεται μια ομάδα για να διασπαστεί από την λίστα με τις ομάδες που περιμένουν διάσπαση.
- **Βήμα 3 :** Γίνεται η διάσπαση της ομάδας σε δύο υποομάδες. Για τον διαχωρισμό εκτελείται ο αλγόριθμος Incremental Bisecting K-Means.
- **Βήμα 4 :** Υπολογισμός του κριτηρίου BIC για τον πατέρα κα για τα παιδία (καινούργιες ομάδες που έχουν δημιουργηθεί). Με βάση τα κριτήρια έχουμε να διαλέξουμε μεταξύ δύο περιπτώσεων :
  - Αν το κριτήριο BIC του πατέρα είναι μικρότερο από το κριτήριο των δύο καινούργιων ομάδων τότε δεχόμαστε την καινούργια ομαδοποίηση. Προστίθενται οι δύο καινούργιες ομάδες στη λίστα με τις ομάδες που περιμένουν διάσπαση.
  - Σε περίπτωση που το κριτήριο του πατέρα είναι μεγαλύτερο από αυτό των παιδιών δεν δεχόμαστε την καινούργια ομαδοποίηση και κρατάμε σαν ομάδα τον πατέρα. Αφαιρείται ο πατέρας από την λίστα.
- **Βήμα 5 :** Επανάληψη βημάτων 2, 3, 4 έως ότου δεν μείνει καμιά ομάδα στη λίστα με τις ομάδες που περιμένουν διάσπαση

Σχήμα 2.6: Βήματα μεθόδου BIC-Means

2.2 Εξαγωγή όρων MeSH

Μετά την ομαδοποίηση του γράφου ακολουθεί η ταυτοποίηση των ομάδων. Στο σημείο αυτό πρέπει να δοθεί μια περιγραφή της ομάδας με ιατρικούς όρους. Για την διαδικασία εξαγωγής όρων από τα κείμενα χρησιμοποιήθηκε η μέθοδος AMTEx. Η μέθοδος αναπτύχθηκε στο Intelligence Systems Laboratory του Πολυτεχνείου Κρήτης.

Η μέθοδος εκτελεί αυτόματη εξαγωγή όρων από μεγάλες συλλογές κειμένων. Η AMTEx συνδυάζει τον θησαυρό ορολογίας MeSH της Εθνικής Βιβλιοθήκης Ιατρικής των Ηνωμένων Πολιτειών Αμερικής (U.S. National Library of Medicine (NLM)) και την καθιερωμένη μέθοδο για εξαγωγή όρων C/NC-value method.

2.2.1 Θησαυρός Ορολογίας MeSH

Ο θησαυρός ορολογίας MeSH (Medical Subject Headings) είναι μια ταξονομία από ιατρικούς και βιολογικούς όρους και έννοιες που προτάθηκαν από την Εθνική Βιβλιοθήκη Ιατρικής των Ηνωμένων Πολιτειών (U.S. National Library of Medicine). Η οργάνωσή τους έχει γίνει χρησιμοποιώντας Extensive Markup Language (XML) [3]. Τα αρχεία του θησαυρού διατίθενται σε μορφή XML στη ηλεκτρονική σελίδα της Εθνικής Βιβλιοθήκης Ιατρικής (NLM)[4]. Οι όροι αυτοί είναι οργανωμένοι σε ιεραρχίες IS-A, όπου η πιο γενικοί όροι, όπως το *"'chemicals and drugs"*, εμφανίζονται σε πιο ψηλά επίπεδα από τους πιο συγκεκριμενους όρους, όπως το *"'aspirin"*. Το MeSH είναι οργανωμένο σε 15 ταξονομίες και περιέχει περισσότερους από 22,000 όρους. Ένας όρος μπορεί να εμφανιστεί σε περισσότερες από μια ταξονομίες. Η περιγραφή του κάθε όρου γίνετε μέσα από διαφορετικές ιδιότητες. Οι πιο σημαντικές είναι οι πιο κάτω :

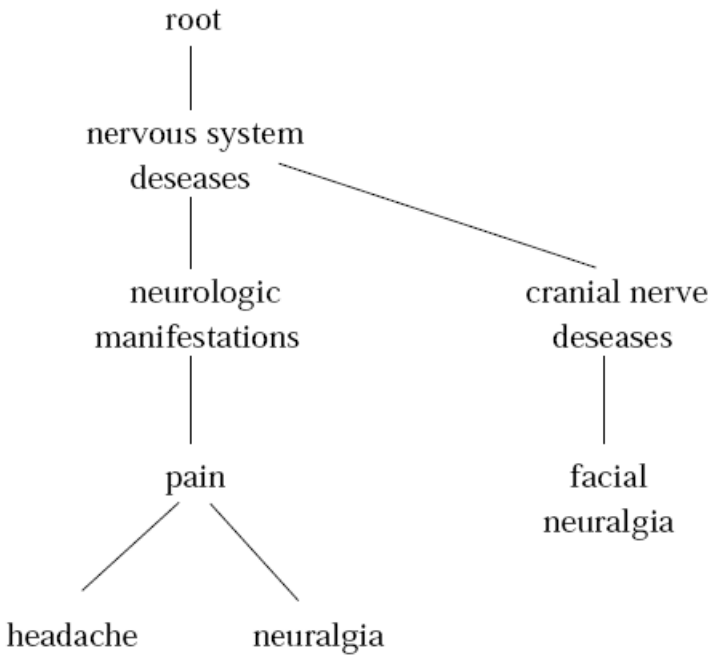
MeSH Heading : Το όνομα του όρου/έννοιας. Χρησιμοποιείται για την δεικτοδότηση των κειμένων στο MEDLINE

Scope Note : Περιγραφή σε μορφή κειμένου του όρου/έννοιας

[3] http://www.w3.org/XML/
[4] http://www.nlm.nih.gov/

Entry Terms : Συνώνυμα του όρου/έννοιας

Ακόμη κάθε όρος MeSH χαρακτηρίζεται από ένα αριθμό δέντρου που υποδεικνύει την ακριβή θέση του όρου στην δεντρική ταξονομία. Για παράδειγμα το κωδικό όνομα του όρου *"Chemicals and drugs"* είναι "D01.029". Στο σχήμα πιο Σχήμα 2.7 φαίνεται μια κατάτμηση της ιεραρχίας IS-A.



Σχήμα 2.7: Κατάτμηση της ιεραρχίας IS-A του MeSH

2.2.2 Η μέθοδος C/NC-value

Αποτελεί μια υβριδική μέθοδο για την εξαγωγή όρων. Η μέθοδος αυτή δεν έχει περιορισμούς θέματος και μπορεί να εφαρμοστεί σε οποιαδήποτε συλλογή κειμένων. Συνδυάζει στατιστική και γλωσσική πληροφορία για την εξαγωγή πολυλεκτικών και εμφωλιασμένων όρων. Σε αυτή τη μέθοδο το κείμενο πρώτα περνάει από μια διαδικασία για διαχωρισμό όρων (tokenizing) και στη συνέχεια από μια διαδικασία ταυτοποίησης των όρων αυτών με την χρήση ενός ταυτοποιητή για μέρη του λόγου (part-of-speech tagger). Τέλος εφαρμόζεται ένας σύνολο από κανόνες και γλωσσικά φίλτρα για την αναγνώριση των υποψήφιων όρων. Τα τρία αυτά φίλτρα είναι τα πιο κάτω :

- N^+N
- $(A|N)^+N$
- $((A|N)^+|((A|N)^*(NP)?)(A|N)^*)N$

όπου N είναι ουσιαστικό , A είναι επίθετο και P αντιπροσωπεύει τον εμπρόθετο προσδιορισμό. Προφανώς τα γλωσσικά φίλτρα έχουν μεγάλο αντίκτυπο στην ακρίβεια (precision) και την ανάκληση (recall) του συστήματος. Χρησιμοποιώντας ένα αρκετά κλειστό φίλτρο όπως το πρώτο, θα αυξηθεί η ακρίβεια και θα μειωθεί η ανάκληση. Αντιθέτως χρησιμοποιώντας ένα σχετικά ανοικτό φίλτρο θα αυξηθεί η ανάκληση και θα μειωθεί η ακρίβεια [5]. Η υλοποίηση της μεθόδου που έχει χρησιμοποιηθεί υλοποιεί και τα τρία πιο πάνω φίλτρα. Η λίστα με τα υποψήφια ουσιαστικά περνάει από ένα φίλτρο για λέξεις χωρίς ιδιαίτερη σημασία (stoplist). Το στατιστικό μέρος της μεθόδου, το οποίο καθορίζει την γειτονιά των όρων για τις υποψήφιες φράσεις, έχει σαν σκοπό να πάρει πιο ακριβή όρους από αυτούς που μπορούν να ληφθούν από την απλή μέθοδο της συχνότητας εμφάνισης. Συγκεκριμένα τους όρους που μπορούν να εμφανιστούν εμφωλιασμένοι μέσα σε μεγαλύτερους όρους, παράδειγμα ο όρος *"enzyme inhibitors"* που βρίσκεται εμφωλιασμένος στον όρο *"Angiotensin-converting enzyme inhibitors"* . Το μέτρο που χρησιμοποιείται για τον πιο πάνω διαχωρισμό είναι το C-value. Ορίζεται σαν η σχέση

της αθροιστικής συχνότητας της εμφάνισης μιας ακολουθίας λέξεων μέσα στο κείμενο, σε σχέση με την συχνότητα εμφάνισης της ίδιας ακολουθίας σε μεγαλύτερες ακολουθίες όρων στο ίδιο κείμενο. Αναλόγως με το αν ο όρος είναι εμφωλιασμένος ή όχι το C-value υπολογίζεται ως εξής :

$$\text{C-value} = \begin{cases} log_2|\alpha| f(\alpha), \\ log_2|\alpha| (f(\alpha) - \frac{1}{P(T_\alpha)} \sum_{b \in T_\alpha} f(b)) \end{cases} \tag{2.8}$$

Στον πιο πάνω τύπο το πρώτο C-value απευθύνεται για τους όρους που δεν είναι εμφωλιασμένοι και το δεύτερο για τους εμφωλιασμένους. Το α αντιπροσωπεύει την ακολουθία των λέξεων που προτείνεται σαν όρος, το $|\alpha|$ είναι το μήκος της ακολουθίας, $f(\alpha)$ είναι η συχνότητα εμφάνισης του όρου σε όλη την συλλογή (τόσο σαν ανεξάρτητος όρως όσο και σαν εμφωλιασμένος όρος σε μεγαλύτερες ακολουθίες), T_α υποδεικνύει την ομάδα από τους όρους που έχουνε εξαχθεί και περιλαμβάνουν το α και $P(T_\alpha)$ είναι ο αριθμός των όρων αυτών. Ο αλγόριθμος C-value παράγει μια λίστα από προτεινόμενους όρους, ταξινομημένους κατά φθίνουσα σειρά όσων αφορά το μέγεθος *likelihood*.

Το NC-value λαμβάνει υπόψη τα συμφραζόμενα για κάθε όρο και δίνει κάποιο βάρος στα ρήματα, επίθετα και ουσιαστικά που εμφανίζονται στους υποψήφιους πολυλεκτικούς όρους. Ο παράγοντας 'βάρος' μιας λέξης w είναι μεγαλύτερος για λέξεις που τείνουν να εμφανίζονται μαζί με πολυλεκτικούς όρους και δίνεται από τον πιο κάτω τύπο:

$$weight(w) = \frac{t(w)}{n} \tag{2.9}$$

όπου $t(w)$ είναι ο αριθμός των πολυλεκτικών όρων στους οποίους εμφανίζεται η λέξη w και n ο αριθμός όλων των όρων. Τέλος το NC-value ορίζεται από τον πιο κάτω τύπο:

$$\text{NC-value} = 0.8 \cdot \text{C-value} + 0.2 \cdot CF(\alpha) \tag{2.10}$$

Στον πιο πάνω τύπο α είναι ο προτεινόμενος όρος, το C-value (α) υπολογίζεται με βάση την Εξ. 2.8 και το $CF(\alpha)$ υπολογίζεται με τον τύπο πιο κάτω:

$$CF(\alpha) = \sum_{w \in C_\alpha} f_\alpha(w) \cdot weigth(w) \tag{2.11}$$

όπου το C_α είναι η ομάδα από τις συμφραζόμενες λέξεις του όρου α, w είναι μια συγκεκριμένη λέξη μέσα στο C_α, $weight(w)$ είναι το βάρος της λέξης w και $f(\alpha)$ είναι η συχνότητα της λέξης σαν συμφραζόμενη του όρου α.

Η μέθοδος C/NC-value έχει δοκιμαστεί επιτυχώς σε πολλές διαφορετικές συλλογές κειμένων όπως μοριακή βιολογία [13], ιατρικά αρχεία [5], βιοιατρικά κείμενα και άρθρα για επιστήμη υπολογιστών [2].

2.2.3 Μέθοδος εξαγωγής όρων AMTEx

Πιο πάνω έχει γίνει η περιγραφή των αρχών στις οποίες βασίστηκε η δημιουργία και λειτουργία της μεθόδου εξαγωγής όρων AMTEx. Έχοντας υπόψη τα πιο πάνω ακολουθεί η περιγραφή των βημάτων που ακολουθεί η μέθοδος για να πετύχει τον σκοπό της.

1. **Εξαγωγή πολυλεκτικών όρων :** Εφαρμόζεται η μέθοδος C/NC-value για την εξαγωγή των όρων. Όπως αναφέρθηκε και προηγουμένως αυτή η μέθοδος είναι ανεξάρτητη από το γενικό θέμα της συλλογής. Χρησιμοποιούνται τα φίλτρα που αναφέρθηκαν.

2. **Ταξινόμιση όρων :** Γίνεται εκτίμηση των εξαγώμενων υποψήφιων όρων πρώτα βάση κριτηρίου C-value και στη συνέχεια βάση κριτηρίου NC-value. Η ταξινόμηση γίνετε με φθίνουσα σειρά. Οι όροι που είναι πιο ψηλά είναι και πιο σημαντικοί.

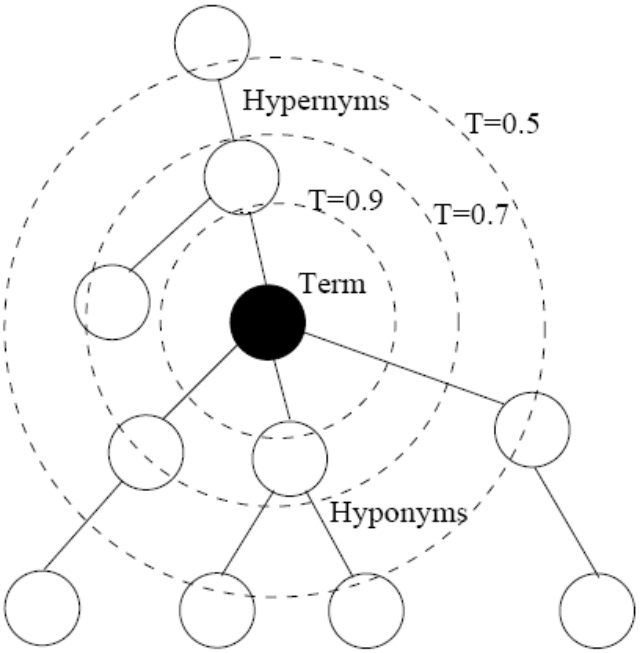
3. **Αντιστοίχηση όρων :** Οι υποψήφιοι όροι αντιστοιχήθηκαν στο θησαυρό ορολογίας MeSH.

4. **Εξαγωγή μονολεκτικών όρων :** Η μέθοδος C/NC-value έχει την τάση

να παράγει πολυλεκτικούς όρους. Σε πολλές περιπτώσεις αυτοί οι πολυλεκτικοί όροι εμπεριέχουν μονολεκτικούς όρους MeSH. Η λίστα με τους υποψήφιους όρους περνάει από επεξεργασία και προστίθενται στην τελική λίστα και οι μονολεκτικοί όροι.

5. **Παραλλαγές όρων :** Παραλλαγές όρων προστίθενται στην λίστα με τους όρους μέσω του πεδίου Entry Terms που υπάρχει στο MeSH.

6. **Επέκταση όρων :** Η λίστα με τους όρους επαυξάνεται με όρους που είναι σημασιολογικά όμοι με αυτούς που ήδη εξάχθηκαν. Όπως φαίνεται καθαρά στο Σχήμα 2.8 θέτοντας ένα κατώτατο όριο (threshold) επιλέγονται οι όροι που είναι σημασιολογικά κοντά στους όρους που υπάρχουν ήδη στη λίστα.



Σχήμα 2.8: Επέκταση όρων με την χρήση του MeSH

2.3 Εννοιολογική Ομοιότητα (Semantic Similarity)

Σημαντικό παράγοντα στην απόδοση μιας αναζήτησης είναι η σύγκριση της ερώτησης του χρήστη με την ήδη υπάρχουσα περιγραφή των κειμένων που υπάρχει στο ευρετήριο. Στη συγκεκριμένη διπλωματική εργασία επιλέχθηκε να χρησιμοποιήσουμε την σημασιολογική και εννοιολογική προσέγγιση της ομοιότητα (Semantic Similarity) των όρων της ερώτησης με τις ομάδες κειμένων που έχουν δημιουργηθεί. Για τον σκοπό αυτό επιλέχθηκε η μέθοδος της Rada Mihalcea σαν ιδέα σύγκρισης και σαν μέτρο ομοιότητας των όρων επιλέχθηκε το μέτρο Li et al. [7].

2.3.1 Μέτρο σύγκρισης Li et Al.

Τα μέτρα σύγκρισης χωρίζονται σε τέσσερις μεγάλες κατηγορίες : (1) μέτρα σύγκρισης που βασίζονται στο πόσο κοντά είναι οι δύο όροι μεταξύ τους σε μια ταξονομία, (2) μέτρα που ελέγχουν πόση κοινή πληροφορία έχουνε οι δύο όροι, (3) μέτρα που ελέγχουν τις ιδιότητες των όρων και (4) μέτρα που χρησιμοποιούν ένα συνδυασμό των πιο πάνω ιδεών. Στη δική μας περίπτωση επιλέχθηκε ένα μέτρο της κατηγορίας (1).

Η ονομασία του μέτρου είναι Li et al. Το εν λόγω μέτρο, το οποίο αποκομίστηκε διαισθητικά και εμπειρικά, συνδυάζει το μήκος L των συντομότερων μονοπατιών μεταξύ των όρων $c_1.c_2$ και το βάθος μέσα στην ταξονομία του πιο κοινού επακριβείς όρου c, H. Ο τύπος του μέτρου δίνεται από το πιο κάτω μη-γραμμικό τύπο :

$$sim_{Li}(c_1, c_2) = e^{-\alpha L} \cdot \frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \tag{2.12}$$

όπου $\alpha \geq 0$ και $\beta > 0$ είναι παράμετροι για κλιμάκωση της συνεισφοράς του μήκους του συντομότερου μονοπατιού και του βάθους αντίστοιχα. Βασιζόμενοι στην παραπομπή [7] οι βέλτιστες τιμές για τις παραμέτρους είναι $\alpha = 0.2$ και $\beta = 0.6$. Αυτό το μέτρο έχει βασικό κίνητρο το γεγονός ότι η πηγή πληροφορίας έχει άπειρο μέγεθος ενώ οι άνθρωποι συγκρίνουν την ομοιότητα των όρων με πεπερασμένο διάστημα, μεταξύ του

εντελώς όμοιου και του εντελώς ξένου. Διαισθητικά η μετατροπή μεταξύ άπειρου και πεπερασμένου διαστήματος είναι μη γραμμική. Σημαντικό να αναφέρουμε είναι ότι το πιο πάνω μέτρο έχει διάστημα μεταξύ 1, για το εντελώς όμοιο, και 0, για το εντελώς ξένο.

2.3.2 Μέθοδος Rada Mihalcea

Έχοντας καλύψει το θέμα της σύγκρισης όρων τους μπορούμε να προχωρήσουμε στην σύγκριση ανυσμάτων όρων. Μια τέτοια μέθοδος έχει αναπτυχθεί από την Rada Mihalcea και Courtney Corley [12]. Η υλοποίηση της μεθόδου απαιτεί δύο μέτρα. Η περιγραφή του πρώτου έχει γίνει στην προηγούμενη παράγραφο. Το δεύτερο είναι ένα μέτρο *ιδιαιτερότητας*. Η ιδιαιτερότητα ενός όρου μπορεί να περιγραφεί χρησιμοποιώντας την *ανάστροφη συχνότητα κειμένου* (inverse document frequency, idf). Το μέτρο αυτό υπολογίζεται διαιρώντας τον συνολικό αριθμό κειμένων της συλλογής με τον αριθμό των κειμένων που περιέχουν τον όρο. Αυτό το μέτρο έχει επιλεγεί λόγω προηγούμενης χρησιμοποίησής του που είχε αποδείξει την αποδοτικότητά του σαν μέτρο ιδιαιτερότητας[6].

Δεδομένου του μέτρου σύγκρισης όρων και του μέτρου ιδιαιτερότητας μπορεί να επεξηγηθεί η ιδέα λειτουργίας της μεθόδου Rada Mihalcea. Δίνονται σαν δεδομένα δύο ανύσματα όρων T_1, T_2. Αρχικά, για κάθε λέξη w στο T_1 βρίσκεται η λέξη στο T_2 που έχει την μεγαλύτερη σημασιολογική ομοιότητα $maxSim(w, T_1)$, σύμφωνα με το μέτρο σύγκρισης όρων που έχουμε αναφέρει πιο πάνω. Στη συνέχεια εφαρμόζεται η ίδια ακριβώς μεθοδολογία για τον προσδιορισμό της πιο σημασιολογικά όμοιας λέξης w από το άνυσμα T_2 με λέξεις στο άνυσμα T_1. Οι ομοιότητες των λέξεων ισοσταθμίζονται με βάση το μέτρο ιδιαιτερότητας της λέξης, προστίθενται μεταξύ τους και κανονικοποιούνται με βάση το μήκος του ανύσματος. Στο τέλος τα αποτελέσματα ομοιότητας αθροίζονται και διαιρούνται δια δύο για να βρεθεί ο μέσος όρος ομοιότητας. Πιο κάτω είναι ο τύπος που

δίνει το αποτέλεσμα της ομοιότητας των ανυσμάτων :

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in \{T_1\}} (maxSim(w, T_2) * idf(w))}{\sum_{w \in \{T_1\}} idf(w)} + \frac{\sum_{w \in \{T_2\}} (maxSim(w, T_1) * idf(w))}{\sum_{w \in \{T_2\}} idf(w)}\right) \tag{2.13}$$

Αυτό το μέτρο ομοιότητας παίρνει τιμές από 0 μέχρι 1, με 1 ορίζονται όμοια ανύσματα και με 0 ανύσματα με καμιά σημασιολογική ομοιότητα.

Κεφάλαιο 3

Σχετικές Εργασίες

Στο διαδύκτιακό χώρο υπάρχουνε αρκετές μηχανές αναζήτησης οι οποίες χρησιμοποιούν πληροφορία που σχετίζεται με τις παραπομπές των κειμένων για να πετύχουν καλύτερα αποτελέσματα σε αναζητήσεις επιστημονικών άρθρων. Μερικές από τις πιο γνωστές είναι το CiteSeer [1] το Google Scholar [2] και το Scirus [3].

3.1 CiteSeer

Το CiteSeer είναι μια δημόσια μηχανή αναζήτησης και συνάμα μια ψηφιακή βιβλιοθήκη για επιστημονικά και ακαδημαϊκά έγγραφα. Δημιουργήθηκε από τους ερευνητές Steve Lawrence, Kurt Bollacker και Lee Giles στο NEC Research Institute. Σκοπός του CiteSeer είναι να αναζητήσει και να συγκεντρώσει πληροφορία για ακαδημαϊκά και επιστημονικά έγγραφα στο διαδίκτυο και να αυτοματοποιήσει την διαδικασία ευρετηρίασης των παραπομπών τους. Έχοντας αυτή την πληροφορία μπορούν να εκτελεστούν ερωτήσεις που αφορούν τις παραπομπές των κειμένων ή και τα ίδια τα κείμενα. Η ιστοσελίδα αυτής της μηχανής στεγάζεται στο College of Information Sciences and Technology στην Pennsylvania και διαθέτει περισσότερα από 700,000 κείμενα, κυρίως στον τομέα των ηλεκτρονικών υπολογιστών, της επιστήμης πληροφορίας και της εφαρμοσμένης μηχανικής.

[1] http://citeseer.ist.psu.edu/
[2] http://scholar.google.com/
[3] http://www.scirus.com/srsapp/

Η βάση δεδομένων του CiteSeer δεν έχει ανανεωθεί εκτενώς από το 2005 λόγω των περιορισμών στην αρχιτεκτονική του συστήματος. Είναι μια απεικόνιση ενός δείγματος των ερευνητικών άρθρων που αφορούν επιστήμη υπολογιστών και πληροφορίας αλλά είναι περιορισμένο σε κάλυψη. Έχει γίνει ευρετηρίαση μόνο σε άρθρα ανοικτής πρόσβασης που διατίθενται δωρεάν στο κοινό, συνήθως μέσα από τις ιστοσελίδες των συγγραφέων τους.

Η επόμενη γενεά του CiteSeer ονομάζεται CiteSeerX και χρηματοδοτείτε από τον οργανισμό National Science Foundation και τον οργανισμό Microsoft Research. Σκοπός της καινούργιας αρχιτεκτονικής είναι η βελτιστοποίηση του CiteSeer σαν μηχανή αναζήτησης αλλά και σαν ψηφιακή βιβλιοθήκη. Ένα παράδειγμα της καινούργιας αυτής μηχανής είναι η προσθήκη της έννοιας *συνεισφορά (contribution)*. Η έννοια αυτή κάνει τη μηχανή το πρώτο αυτοματοποιημένο ευρετήριο για συνεισφορές. Ακόμη το CiteSeerX θα υλοποιεί καινούργιους αλγορίθμους για την εξαγωγή οντοτήτων και ένα πρότυπο, επεκτάσιμο, ιεραρχικό σύστημα αρχιτεκτονικής βασισμένο σε εργαλεία ανοικτού λογισμικού όπως το Lucene και άλλα εργαλεία του οργανισμού Apache. Λόγω αυτού, το CiteSeerX θα προάγει την δημιουργία και άλλων παρόμοιων εργαλείων αναζήτησης.

3.2 Google Scholar

Τον Νοέμβριο του 2004 η εταιρία Google ανακοίνωσε την δημιουργία της μηχανής αναζήτησης Google Scholar. Σκοπός της είναι να διαθέσει στους χρήστες ένα απλό τρόπο για την αναζήτηση επιστημονικών άρθρων. Από αυτό το ευρετήριο μπορεί οποιοσδήποτε να αναζητήσει σε πολλές πηγές όπως βιβλία, περιγραφές άρθρων, διπλωματικές εργασίες, ακαδημαϊκά δημοσιεύματα, επιστημονικές κοινότητες, πανεπιστήμια και άλλους οργανισμούς με ακαδημαϊκό περιεχόμενο.

Το Google Scholar βοηθάει στην αναγνώριση των πιο σχετικών άρθρων μέσα σε επιστημονικές συλλογές. Η αναζήτηση μέσα από επιστημονικά άρθρα στο Google Scholar αποδίδει πολύ ικανοποιητικά όταν η ερώτηση περιέχει όρους που είναι συγκεκριμένοι

αρκετά για να προσδιορίσουν το θέμα και επίσης γενικοί αρκετά για να μην αγνοήσουν τα σχετικά άρθρα. Οι συγγραφείς χρησιμοποιούν διαφορετική ορολογία για να αναφερθούν στο ίδιο θέμα, κάτι που συνήθως συμβαίνει όταν ένα θέμα είναι ακόμη πρόσφατο. Αυτό κάνει την μέθοδο αναζήτησης λιγότερο αποδοτική. Για το σκοπό αυτό έγινε μια προσθήκη στο Google Scholar για να μπορούν να συσχετιστούν τα άρθρα. Για τα αποτελέσματα της αναζήτησης γίνεται μια προσπάθεια να προσδιοριστούν αυτόματα τα σχετικά άρθρα μέσα από το ευρετήριο παραπομπών. Η λίστα με αυτά τα άρθρα είναι διαθέσιμη στο σύνδεσμο Related Articles που βρίσκεται δίπλα από κάθε αποτέλεσμα. Η λίστα με τα σχετικά άρθρα είναι ταξινομημένη με βάση δύο κριτήρια. Αρχικά ανάλογα με το πόσο όμοια είναι τα σχετικά άρθρα σε σχέση με τα αρχικά αποτελέσματα και κατά δεύτερο πόσο σχετικά είναι με την αρχική ερώτηση.

Η ανάκτηση ομάδων από σχετικά άρθρα και βιβλία είναι συχνά ένας πολύ καλός τρόπος για αρχάριους να αποκτήσουν γνώσεις για κάποιο συγκεκριμένο θέμα. Εντούτοις και οι πιο έμπειροι χρήστες μπορεί πολλές φορές να ξαφνιαστούν από την ανακάλυψη σχετικών άρθρων πάνω στο θέμα ειδίκευσης τους.

3.3 Scirus

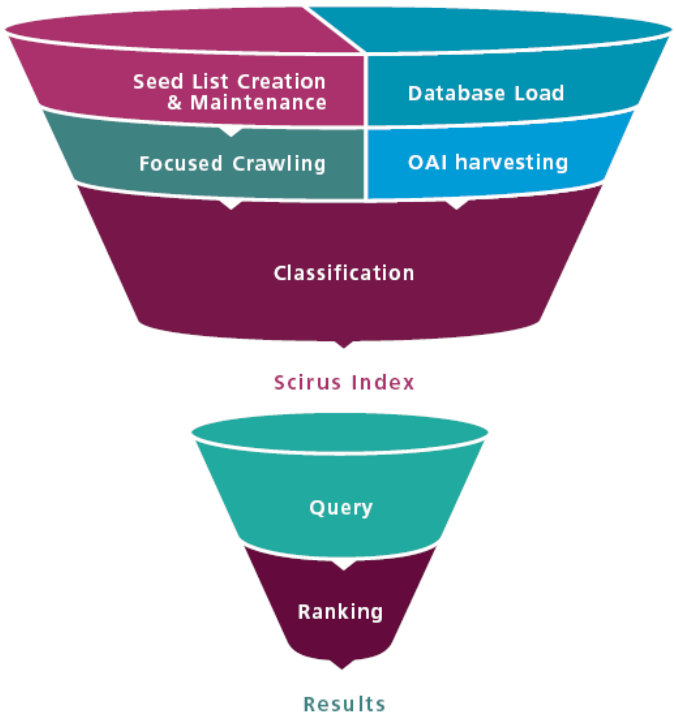
Scirus είναι μια περιεκτική μηχανή αναζήτησης επιστημονικών άρθρων. Η λειτουργία της είναι παρόμοια με αυτή του CiteSeer και του Google Scholar, δηλαδή είναι εστιασμένη σε επιστημονική πληροφορία. Σε αντίθεση με το CiteSeer δεν είναι μόνο για πληροφορία που αφορά επιστήμη υπολογιστών. Επίσης επιστρέφονται αποτελέσματα που δεν περιέχουν το πλήρες κείμενο. Το Scirus είναι ιδιοκτησία της εταιρίας Elsevier.

Τα αποτελέσματα του Scirus [3] προέρχονται από όλο το διαδικτυακό χώρο, συμπεριμβανομένου και ιστοσελίδων που έχουνε περιορισμένη πρόσβαση. Αυτές οι ιστοσελίδες συνήθως δεν υπάρχουνε στα ευρετήρια των περισσότερων μηχανών αναζήτησης, κάτι που δίνει ένα μεγάλο πλεονέκτημα στο Scirus έναντι των υπολοίπων μηχανών. Το Scirus καλύπτει προς το παρών πάνω από 450 εκατομμύρια ιστοσελίδες με επιστημονικό

περιεχόμενο. Το ευρετήριο απαρτίζεται από τις πιο κάτω ιστοσελίδες :

- 150 εκατ. ιστοσελίδες .edu
- 54 εκατ. ιστοσελίδες .org
- 9 εκατ. ιστοσελίδες .ac.uk
- 52 εκατ. ιστοσελίδες .com
- 36 εκατ. ιστοσελίδες .gov
- Περισσότερες από 143 εκατ. ιστοσελίδες από πανεπιστήμια σε όλο τον κόσμο

Στο Σχήμα 3.1 φαίνεται η ιεραρχία της αρχιτεκτονικής της μηχανής Scirus.



Σχήμα 3.1: Ιεραρχία αρχιτεκτονικής της μηχανής αναζήτης Scirus

Κεφάλαιο 4

Υλοποίηση Συστήματος

Σε αυτό το κεφάλαιο θα γίνει η περιγραφή των βημάτων που εκτελέστηκαν για την διεκπεραίωση αυτής της εργασίας. Σε γενικές γραμμές η εργασία μπορεί να χωριστεί σε τέσσερα βήματα. Την επεξεργασία των κειμένων για την δημιουργία του γράφου παραπομπών, την ομαδοποίηση με τις διάφορες μεθόδους, την ταυτοποίηση των ομάδων και την ανάκτηση και ταξινόμηση των αποτελεσμάτων. Ακόμη θα γίνει περιγραφή του τρόπου αποθήκευσης των διάφορων ευρετηρίων στα σημεία που δημιουργούνται.

4.1 Επεξεργασία συλλογής κειμένων

Η συλλογή κειμένων που χρησιμοποιήθηκε προέρχεται από την Εθνική Βιβλιοθήκη Ιατρικής των Ηνωμένων Πολιτειών Αμερικής (NLM)[1]. Επιλέχθηκε η συγκεκριμένη συλλογή λόγω της διαθεσιμότητας όλης της πληροφορίας για κάθε άρθρο. Τα άρθρα αυτά είναι διαθέσιμα στο κοινό σε μορφή XML και περιέχουν όλη την πληροφορία του άρθρου *(full text)*. Υπάρχει επίσης διαθέσιμο το XML Schema με το οποίο γίνεται δυνατή η κατανόηση την διάρθρωση της πληροφορίας μέσα στα αρχεία XML.

Αρχικά η συλλογή κειμένων PMC περιείχε 57,355 κείμενα. Από αυτά κρατάμε μόνο τα κείμενα που έχουνε παραπομπές σε κείμενα που υπάρχουν στην συλλογή μας. Σκοπός είναι να σχηματίσουμε ένα γράφο στον οποίο όλοι οι κόμβοι (άρθρα στην συγκεκριμένη περίπτωση) έχουν όλη την πληροφορία τους διαθέσιμη. Η προσθήκη άρθρων με

[1] http://www.nlm.nih.gov/

μερική πληροφορία θα καταστήσει αδύνατη την επεξεργασία στα επόμενα στάδια. Για παράδειγμα η προσθήκη άρθρων χωρίς το κυρίως κείμενο θα δημιουργήσει πρόβλημα στην ταυτοποίηση των ομάδων.

Χρησιμοποιώντας τη βιβλιοθήκη *SAX* [2] (Simple API for XML) της Java εξάγεται η χρήσιμη για την εφαρμογή πληροφορία. Η πληροφορία αυτή περιλαμβάνει :

- **PMID :** Αναγνωριστικό που δίνεται από την Εθνική Βιβλιοθήκη Ιατρικής Ηνωμένων Πολιτειών για την μονοσήμαντη ταυτοποίηση των άρθρων στην συλλογή.
- **Abstract :** Περίληψη του άρθρου.
- **Body :** Το κυρίως κείμενο του άρθρου.
- **Keywords :** Λέξεις κλειδιά που δίνουν μια γενική ιδέα για το τι περιέχει το άρθρο. Αυτές οι λέξεις μπορεί να προέρχονται από λεξικά, οντολογίες ή άλλες πηγές και δεν είναι διαθέσιμες για όλα τα άρθρα.
- **Citations :** Οι παραπομπές του άρθρου.

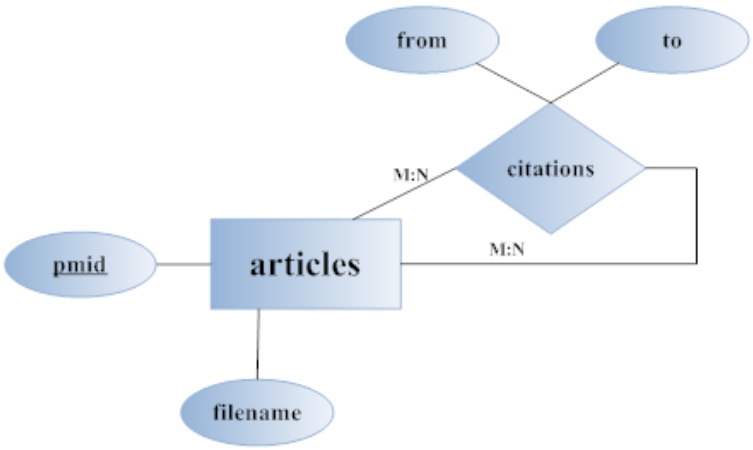
Μετά την επεξεργασία επιλέχθηκαν 11,712 κείμενα τα οποία ενώνουν 29,607 σχέσεις. Το επόμενο στάδιο μετά την εξαγωγή της πληροφορίας είναι η αποθήκευσή της σε μια βάση δεδομένων για την ευκολότερη χρησιμοποίησή της. Για αυτό το σκοπό χρησιμοποιήθηκαν δύο τρόποι αποθήκευσης. Μια σχεσιακή βάση δεδομένων σε MySQL για την αποθήκευση του γράφου παραπομπών και ένα ανεστραμμένο ευρετήριο σε Lucene για την αποθήκευση της υπόλοιπης πληροφορίας.

4.1.1 Γράφος Παραπομπών (Citation Graph)

Δημιουργήθηκε μια βάση δεδομένων για να κρατήσει την πληροφορία που αφορά αποκλειστικά τον γράφο παραπομπών. Δηλαδή ποιες παραπομπές διαθέτει κάθε άρθρο

[2] http://www.saxproject.org/

(forward links). Με την αποθήκευση αυτής της πληροφορίας γίνεται διαθέσιμη η δυνατότητα να βρούμε πια άρθρα έχουνε χρησιμοποιήσει ένα άρθρο για την εγγραφή τους (backward links). Εδώ να επισημανθεί ότι έχουν αποθηκευτεί *μόνο* οι παραπομπές που αναφέρονται σε άρθρα που υπάρχουν στην τρέχον συλλογή. Η δομή της βάσης αποτελείτε από : (1) ένα πίνακα που περιέχει για κάθε άρθρο το PMID και το όνομα του XML αρχείου και (2) ένα πίνακα που κρατάει την πληροφορία των ακμών του γράφου, δηλαδή τις παραπομπές. Στο Σχήμα 4.1 φαίνεται γραφικά η δομή της βάσης αυτής.



Σχήμα 4.1: Γράφημα της σχεσιακής βάσης σε MySQL

4.1.2 Ανεστραμμένο Ευρετήριο (Inverted Index)

Για την αποθήκευση της υπόλοιπης πληροφορίας δημιουργήθηκε ένα ανεστραμμένο ευρετήριο με την χρήση του πακέτου Lucene [3]. Ο κυρίως λόγος που οδήγησε στην επιλογή αυτού του πακέτου είναι η πλούσια σε περιεχόμενο συλλογή εργαλείων που διαθέτει. Ένα από τα πιο σημαντικά εργαλεία κάνει δυνατή την επεξεργασία κατά την αποθήκευση των διαφόρων πεδίων που απαρτίζουν ένα άρθρο. Εξάγεται πληροφορία που θα μας είναι χρήσιμη για την συνέχεια. Η πληροφορία περιλαμβάνει διαχωρισμό λέξεων με συγκεκριμένα φίλτρα ανά πεδίο, κάτι που θα βοηθήσει στην αναζήτηση όρων. Γίνεται επίσης ευρετηρίαση και δημιουργία πινάκων με τις συχνότητες εμφάνισης των όρων του

[3] http://lucene.apache.org/java/docs/index.html

κειμένου. Περισσότερες πληροφορίες για το πακέτο αναφέρονται στο Παράρτημα Α΄.2. Περαιτέρω από την πληροφορία στην οποία αναφερθήκαμε πιο πάνω αποθηκεύονται σε αυτό το ευρετήριο και οι όρους που εξάχθηκαν από την επεξεργασία των άρθρων με την μέθοδο AMTEx. Παρακάτω περιγράφονται αναλυτικά ονομαστικά τα πεδία και το περιεχόμενό τους στο ανεστραμμένο ευρετήριο :

- ***pmid :*** Αναγνωριστικό άρθρου που δίνεται από το MEDLINE για την μονοσήμαντη δεικτοδότηση των άρθρων
- ***body :*** Το κυρίως κείμενο του άρθρου. Αυτό το πεδίο περνάει από διαδικασία διαχωρισμού λέξεων. Αυτό δημιουργεί όπως προαναφέρθηκε ένα πίνακα με τις συχνότητες εμφάνισης όρων
- ***abstract :*** Η περίληψη του άρθρου. Το πεδίο περνάει από επεξεργασία διαχωρισμού λέξεων
- ***article-title :*** Ο τίτλος του άρθρου
- ***keywords :*** Οι λέξεις κλειδιά που αναγράφονται στα XML αρχείο του άρθρου
- ***amtex-terms :*** Οι όροι που εξάγηκαν με την μέθοδο AMTEx. Οι όροι είναι διαχωρισμένοι με ';'. Το πεδίο έχει περάσει επίσης από διαδικασία διαχωρισμού όρων

4.2 Ομαδοποίηση γράφου παραπομπών

Έχοντας εξαγάγει όλη την χρήσιμη πληροφορία, επόμενο στάδιο επεξεργασίας είναι οι μέθοδοι ομαδοποίησης του γράφου παραπομπών. Η αναπαράσταση του γράφου στην μνήμη για την εκτέλεση των μεθόδων γίνεται με την βοήθεια του πακέτου JUNG (Java Universal Network/Graph) [4] [4].

[4] http://jung.sourceforge.net/

Οι υλοποιήσεις των αλγορίθμων έχουν γίνει σε Java [5] και MATLAB [6]. Συγκεκριμένα η υλοποίηση του αλγορίθμου Edge-Betweenness είναι διαθέσιμη μέσα από το πακέτο JUNG. Η υλοποίηση του αλγορίθμου BIC-Means είναι μια παραλλαγή του κώδικα από την μεταπτυχιακή εργασία του Νικόλαου Χουρδάκη και είναι υλοποιημένη σε Java. Οι αλλαγές είναι μικρές και αφορούν την προσαρμογή του κώδικα στην τρέχον συλλογή. Δεν έχει γίνει καμία τροποποίηση όσον αφορά την λογική ή τους τύπους πάνω στους οποίους στηρίζεται η μέθοδος. Η μέθοδος Markov Clustering είναι από κοινού υλοποιημένη και στις δύο γλώσσες προγραμματισμού.

Για τον αλγόριθμο Edge-Betweenness δεν χρειάστηκε να υλοποιηθεί κάποιο μέρος του. Το πακέτο JUNG έχει ενσωματωμένες όλες τις συναρτήσεις που απαιτούνται για τα βήματα του επαναληπτικού αλγορίθμου. Πέραν από την εκτέλεση του επαναληπτικού αλγορίθμου διαθέτει συναρτήσεις για την ανάκτηση των ακμών που αφαιρούνται, τα βάρη των ακμών μετά από κάθε επανάληψη καθώς και την τελική ομαδοποίηση μετά το πέρας της εκτέλεσης. Ακόμη υπάρχουν έτοιμες συναρτήσεις για την αποθήκευση του γράφου σε μορφή αρχείου XML για να είναι δυνατή η συνέχιση της εκτέλεση σε περίπτωση διακοπής της διαδικασίας. Στο αρχείο αυτό αποθηκεύεται το τρέχον στιγμιότυπο του γράφου, δηλαδή δεν υπάρχουν οι ακμές που αφαιρέθηκαν μέχρι την στιγμή εκείνη. Σε κάθε επανάληψη υπολογίζονται τα συντομότερα μονοπάτια όλων των ζευγών από κόμβους. Σε περίπτωση που η αφαίρεση μιας ακμής διαχωρίζει δύο ομάδες, τότε αποθηκεύονται οι δύο ξεχωριστοί πλέον γράφοι σε διαφορετικά αρχεία και η επεξεργασία συνεχίζεται ανεξάρτητα. Δεν υπάρχει λόγος να γίνουν άσκοποι υπολογισμοί για μονοπάτια μεταξύ ζευγαριών που βρίσκονται σε ξεχωριστούς γράφους. Τα μονοπάτια αυτά είναι ανύπαρκτα.

Όπως προαναφέρθηκε για την μέθοδο BIC-Means έχει χρησιμοποιηθεί ένα μεγάλο μέρος του ήδη υπάρχον κώδικα σε Java. Οι τροποποιήσεις που έγιναν αφορούν τα σημεία στα οποία ο αλγόριθμος χρειάζεται πληροφορία από τα κείμενα της συλλογής. Η

[5] http://www.java.com/en/
[6] http://www.mathworks.com/

πληροφορία αφορά τα ανύσματα όρων του κάθε κειμένου, συχνότητες εμφάνισης των όρων, τόσο ανά κείμενο όσο και μέσα σε όλη τη συλλογή, το βάρος των όρων και την ανάστροφη συχνότητα κειμένου (inverse document frequency, idf). Οι τιμές των παραμέτρων αυτών μπορούν εύκολα να ανακτηθούν μέσα από το ανάστροφο ευρετήριο Lucene.

Όσων αφορά την υλοποίηση του αλγορίθμου Markov Clustering, έγινε σε τρία μέρη. Το πρώτο μέρος είναι η δημιουργία του αρχικού πίνακα, το δεύτερο η εκτέλεση του επαναληπτικού αλγορίθμου και το τρίτο η εξαγωγή της τελικής ομαδοποίησης. Το πρώτο και τρίτο μέρος έχει υλοποιηθεί σε Java και το δεύτερο μέρος σε MATLAB. Έχει επιλεγεί αυτός η διάρθρωση γιατί η MATLAB είναι ένα πολύ καλό πακέτο για πράξεις πινάκων. Υπάρχει διαθέσιμη όλη την λειτουργικότητα που θα χρειαστεί για τον υπολογισμό των πινάκων στα επαναληπτικά βήματα. Ακόμη ένα πλεονέκτημα της συγκεκριμένης υλοποίησης είναι το γεγονός ότι η MATLAB λειτουργεί με πολυεπεξεργασία (multi-threading), κάτι που θα επιταχύνει πολύ τον υπολογισμό της τελικής ομαδοποίησης. Οι πράξεις που χρησιμοποιεί ο αλγόριθμος (πολλαπλασιασμός, ύψωση σε δύναμη) μπορούν να εκτελεστούν σε πολλαπλούς επεξεργαστές για καλύτερη απόδοση.

Με το πέρας της εκτέλεσης των αλγορίθμων ομαδοποίησης δημιουργείτε μια βάση δεδομένων για την αποθήκευσή τους. Σε αυτό το σημείο της εργασίας επιλέχθηκε για την αποθήκευση το πακέτο Berkeley DB. Όπως εξηγείτε και στο Παράρτημα Α΄.6, η συγκεκριμένη βάση δεδομένων δεν είναι σχεσιακή αλλά λειτουργεί με λεξικά. Δίνεται ένα κλειδί (key) σε κάθε ομάδα και αποθηκεύεται σε αυτό το κλειδί η λίστα με τα αναγνωριστικά (PMID) των κειμένων που την αποτελούν (value). Με αυτό τον τρόπο δημιουργούνται ζεύγη από (key, value) που αντιπροσωπεύουν τις ομάδες. Κάθε μέθοδος έχει τη δική της ξεχωριστή βάση δεδομένων (Database).

4.3 Ταυτοποίηση Ομάδων

Μετά το τέλος της ομαδοποίησης σειρά έχει η ταυτοποίηση των ομάδων που δημιουργήθηκαν από τις μεθόδους. Για να γίνει εφικτό θα χρησιμοποιηθούν τρεις παράμετροι.

- Οι ομάδες από την κάθε μέθοδο
- Η βάση δεδομένων με την πληροφορία για κάθε άρθρο
- Ένα λεξικό με τους σημαντικότερους όρους όλης της συλλογής μαζί με τα NC-Value τους

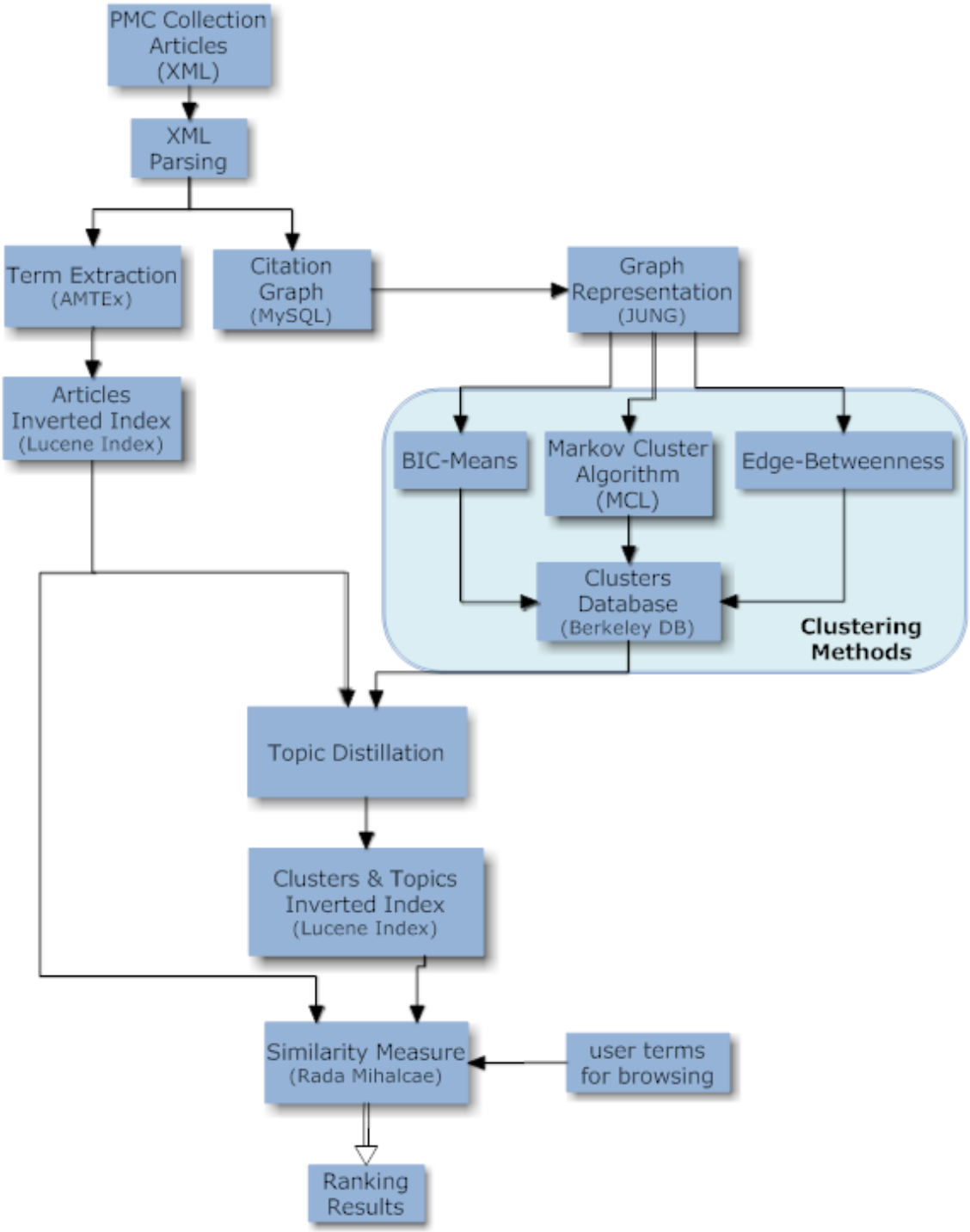
Οι δύο πρώτοι παράμετροι έχουνε ήδη δημιουργηθεί και βρίσκονται αποθηκεμένοι σε βάσεις δεδομένων. Αυτό που πρέπει να επεξηγηθεί είναι η χρησιμότητα της τρίτης παραμέτρου, το λεξικό με τους σημαντικότερους όρους σε ολόκληρη τη συλλογή.

Η μεθοδολογία που εφαρμόστηκε για την ταυτοποίηση των ομάδων είναι η εξής. Δημιουργούνται τα κέντρα των ομάδων για να βρεθούν οι όροι που τα απαρτίζουν. Τα κέντρα των ομάδων είναι ανύσματα που περιλαμβάνουν τους όρους των κειμένων που τα αποτελούν. Για την συγκεκριμένη συλλογή κειμένων επιλέχθηκαν να χρησιμοποιηθούν σαν όροι κειμένων, οι ιατρικοί όροι που εξάχθηκαν από την επεξεργασία των κειμένων με την μέθοδο AMTEx. Αυτοί οι όροι είναι όρου του θησαυρού ορολογίας MeSH. Το καθένα άρθρο περιγράφεται από περίπου δέκα με είκοσι τέτοιους όρους. Η ένωση όλων των όρων της κάθε ομάδας δίνει ένα πολύ μεγάλο σε αριθμό άνυσμα από όρους (περισσότερους από 100). Για την μείωση του αριθμού των όρων εφαρμόστηκε ένα μέτρο *σημαντικότητας* για το φιλτράρισμα των ανυσμάτων.

Με βάση την ιδέα αυτή έγινε επεξεργασία όλης της συλλογής με τη μέθοδο AMTEx. Δίνοντας στη μέθοδο όλα τα κείμενα της συλλογής σαν ενιαίο κείμενο, επιστρέφεται στην έξοδο ένας αριθμό από όρους, MeSH όρους πάντα, που περιγράφουν την συλλογή. Αυτοί είναι οι πιο σημαντικοί όροι της συλλογής και είναι οι όροι που θα παραμείνουν στις ταυτότητες των ομάδων. Η μέθοδος C/NC-value δεν επιστρέφει **όλους** τους πολυλεκτικούς όρους από το κείμενο εισόδου αλλά εφαρμόζει ένα κατώφλι για να επιλέξει

τους πιο σημαντικούς. Το κριτήριο αυτό είναι μια τιμή του C-value [5]. Το συγκεκριμένο κατώφλι υπολογίζεται πειραματικά μέσα από το άρθρο της μεθόδου AMTEx[1]. Το λεξικό που δημιουργήθηκε περιέχει 2500 MeSH όρους. Αυτοί είναι οι όροι που θα χρησιμοποιηθούν για την ταυτοποίηση των ομάδων. Φιλτράροντας τις ομάδες μειώθηκαν οι όροι των ταυτοτήτων σε 30 με 40.

Αναγκαία σε αυτό το σημείο είναι η δημιουργία ενός ανεστραμμένου ευρετηρίου. Σε αυτό το ευρετήριο θα βρίσκεται η πληροφορία που αφορά τις ταυτότητες των ομάδων όπως επίσης και τα αναγνωριστικά των κειμένων που τις αποτελούν. Με την προσθήκη των αναγνωριστικών σε αυτό το ευρετήριο γίνεται περιττή η παρουσία της βάσης δεδομένων σε Berkeley DB. Όλη η πληροφορία που αφορά τις ομάδες βρίσκεται μαζεμένη στο ανεστραμμένο ευρετήριο. Με τη χρήση του ευρετηρίου γίνεται εύκολη η ανάκτηση των ομάδων με κριτήριο ένα όρο ή ένα συγκεκριμένο αναγνωριστικό. Ο συνδυασμός του με το ήδη υπάρχον ανεστραμμένο ευρετήριο που διαθέτει την πληροφορία ανά κείμενο, θα αποτελέσει τον κορμό για την υλοποίηση του επόμενου βήματος. Στο παρακάτω σχήμα φαίνεται γραφικά η δομή του όλου συστήματος μέχρι τώρα.



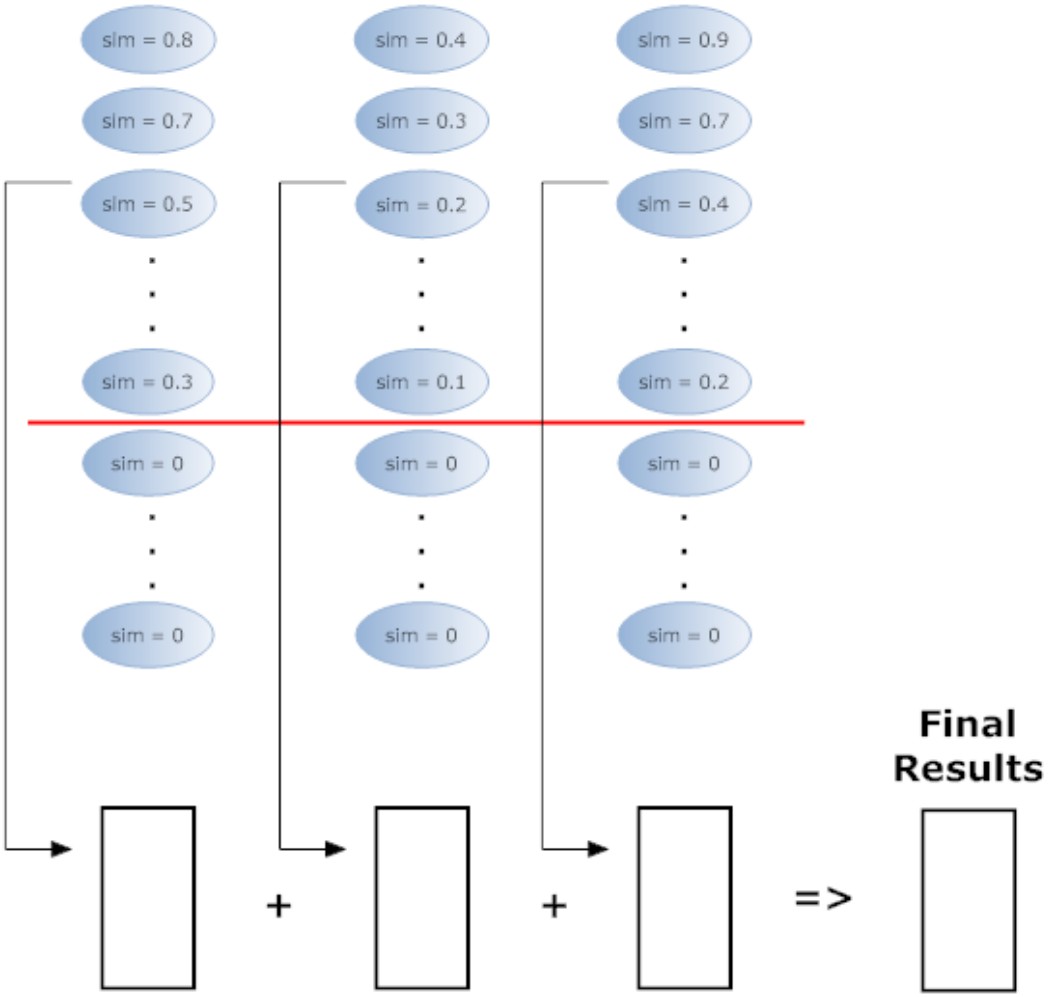
Σχήμα 4.2: Δομή συστήματος

4.4 Ανάκτηση και Ταξινόμηση Αποτελεσμάτων

Το τελικό στάδιο της εργασίας προϋποθέτει την εισαγωγή όρων από τον χρήστη για το ξεφύλλισμα της συλλογής. Υπάρχουν δύο μεθοδολογίες για τον υπολογισμό ομοιότητας μεταξύ ανυσμάτων όρων. Η πρώτη είναι η κλασσική μέθοδος της απλής λεξικογραφικής σύγκρισης όρων, γνωστή σαν Vector Space Model. Η δεύτερη είναι η σημασιολογική ομοιότητα των ανυσμάτων, γνωστή σαν Semantic Similarity. Στην παρούσα εργασία χρησιμοποιήθηκε η δεύτερη μεθοδολογία. Επιλέχθηκε η συγκεκριμένη μεθοδολογία γιατί πολλοί ιατρικοί όροι έχουν μεγάλη σημασιολογική ομοιότητα και η αυστηρή μέθοδος της λεξικογραφικής ομοιότητας συντελεί σε χάσιμο πληροφορίας. Δύο άρθρα που αναφέρονται στο ίδιο ιατρικό θέμα είναι δυνατόν να χρησιμοποιούν διαφορετική ορολογία. Παραδείγματος χάριν ο όρος *ache* και ο όρος *pain*, οι οποίοι σημασιολογικά είναι συνώνυμοι, λεξικογραφικά είναι διάφοροι. Χρησιμοποιώντας μια αυστηρή μέθοδο ομοιότητας όπως το Vector Space Model χάνεται σημαντική πληροφορία. Οι ταυτότητες των ομάδων είναι ανύσματα από όρους που περιγράφουν το θέμα μιας ομάδας. Η έννοια της περιγραφής προϋποθέτει ένα σημασιολογικό χαρακτήρα παρά ένα αυστηρό κριτήριο.

Στο Κεφάλαιο 2.3 έχει επεξηγηθεί αναλυτικά η ιδέα της μεθόδου Rada Mihalcea για τον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ ανυσμάτων με την βοήθεια του μέτρου ομοιότητας όρων Li et Al.. Χρησιμοποιείτε η συγκεκριμένη μέθοδος για τον υπολογισμό της ομοιότητας των όρων του χρήστη με τις ομάδες. Στη συνέχεια γίνεται ταξινόμηση των ομάδων, για κάθε μια από τις τρεις μεθόδους ξεχωριστά, με βάση αυτή την ομοιότητα. Τα κείμενα μέσα στις ομάδες δεν είναι ταξινομημένα. Εφαρμόζοντας την ίδια τεχνική υπολογίζεται η ομοιότητα των όρων του χρήστη με τα κείμενα των ομάδων. Επιτυγχάνεται έτσι μια εσωτερική ταξινόμηση στις ομάδες. Η ταξινόμηση εσωτερικά των ομάδων εφαρμόζεται *μόνο* στις ομάδες που παρουσιάζουν κάποια ομοιότητα με τους όρους του χρήστη. Ομάδες μηδενικής ομοιότητας αγνοούνται. Έχοντας ταξινομηθεί όλες οι ομάδες εκτελείτε μια συγχώνευση (merge)) των ομάδων αυτών ανα μέθοδο. Αυτό δημιουργεί μια ταξινομημένη λίστα κειμένων ανά μέθοδο.

Για την παρουσίαση των τελικών αποτελεσμάτων εφαρμόστηκε μια λογική για τον συνδυασμό των τριών ομαδοποιήσεων. Οι ταξινομημένες λίστες προστίθενται μεταξύ τους αλγεβρικά και δημιουργείτε μια τελική ταξινόμηση από κείμενα. Η ομοιότητα για κείμενα που βρίσκονται σε περισσότερες από μία λίστες είναι αθροιστική ούτως ώστε να ευνοηθούν στην ταξινόμηση. Κείμενα που έχουν επιλεγεί και από τις τρεις ομαδοποιήσεις σαν σχετικά έχουν μεγαλύτερες πιθανότητες ομοιότητας. Στο παρακάτω σχήμα φαίνεται γραφικά η ιδέα του συνδυασμού των αποτελεσμάτων.



Σχήμα 4.3: Γραφική αναπαράσταση ταξινόμησης αποτελεσμάτων

4.5 Παραδείγματα ξεφυλλίσματος

Παρακάτω θα παρατεθούν παραδείγματα ξεφυλλίσματος της συλλογής ιατρικών κειμένων που χρησιμοποιήθηκε. Το παρόν σύστημα τρέχει σε κονσόλα και τα αποτελέσματα εμφανίζονται σε μορφή λίστας. Στα αποτελέσματα φαίνεται ο βαθμός ομοιότητας με τους όρους που εισάγει ο χρήστης, ο τίτλος του άρθρου καθώς και το αναγνωριστικό. Σε μερικά από τα αποτελέσματα παρατίθεται και μέρος από την περίληψη του άρθρου για να γίνει πιο εμφανής η ομοιότητα με τους όρους. Για κάθε ερώτηση παρατίθενται τα αποτελέσματα τόσο του συστήματος που υλοποιήθηκε όσο και τα αποτελέσματα κλασσικής αναζήτησης με την χρήση του μοντέλου λεξικογραφικής ομοιότητας VSM (Vector Space Model) για να γίνει σύγκριση μεταξύ τους.

Παράδειγμα 1 - VSM

```
Enter your query (terms separated with comma ",") : meningioma,carcinoid tumor

Results found : 14

(1) Document PMID : 16140629                Similarity with query : 0.4068941
Title : Aminolevulinic Acid Dehydratase Polymorphism and Risk of Brain
        Tumors in Adults

(2) Document PMID : 16224098                Similarity with query : 0.3296196
Title : Transformation of expression intensities across generations of Affymetrix
        microarrays using sequence matching and regression modeling

(3) Document PMID : 16643657                Similarity with query : 0.3296196
Title : A stable gene selection in microarray data analysis

(4) Document PMID : 17090325                Similarity with query : 0.3051705
Title : Automated recognition of malignancy mentions in biomedical literature

(5) Document PMID : 17217530                Similarity with query : 0.3051705
Title : The role of PDGF in radiation oncology

(6) Document PMID : 16207359                Similarity with query : 0.2884172
Title : Searching for differentially expressed gene combinations

(7) Document PMID : 17125495                Similarity with query : 0.2877175
Title : CNS progenitor cells and oligodendrocytes are targets of
        chemotherapeutic agents
```

```
(8) Document PMID : 16232314                Similarity with query : 0.2497766
Title : Automatic extraction of candidate nomenclature terms using the
        doublet method

(9) Document PMID : 16872493                Similarity with query : 0.2472147
Title : Selection of suitable reference genes for accurate normalization
        of gene expression profile studies in non-small cell lung cancer


(10) Document PMID : 17573973               Similarity with query : 0.2472147
Title : Selecting dissimilar genes for multi-class classification,
        an application in cancer subtyping
...
```

Παράδειγμα 1 - Semantic Similarity

```
Enter your query (terms separated with comma ",") : meningioma,carcinoid tumor

Results found : 220

(1) Document PMID : 17125495                Similarity with query : 0.3735754
Title : CNS progenitor cells and oligodendrocytes are targets of
        chemotherapeutic agents
Abstract :  Background Chemotherapy in cancer patients can be associated
            with serious short- and long-term adverse neurological effects,
            such as leukoencephalopathy and cognitive impairment, even when
            therapy is delivered systemically.

(2) Document PMID : 16140629                Similarity with query : 0.3040373
Title : Aminolevulinic Acid Dehydratase Polymorphism and Risk of Brain
        Tumors in Adults

(3) Document PMID : 12184810                Similarity with query : 0.2503105
Title : A prediction-based resampling method for estimating the number
        of clusters in a dataset
Abstract :  An important statistical problem associated with tumor
            classification is the identification of new tumor classes using
            gene-expression profiles.

(4) Document PMID : 16464241                Similarity with query : 0.1960949
Title : The use of microarray technologies in clinical oncology

...

(10) Document PMID : 16393652               Similarity with query : 0.1913266
Title : Mortality among Workers Exposed to Polychlorinated Biphenyls (PCBs)
        in an Electrical Capacitor Manufacturing Plant in Indiana: An Update
Abstract :  An Indiana capacitor-manufacturing cohort (n = 3,569) was exposed
            to polychlorinated biphenyls (PCBs) from 1957 to 1977. The original
            study of mortality through 1984 found excess melanoma
            and brain cancer;
```

```
(11) Document PMID : 15488140              Similarity with query : 0.1808978
Title : Melanoma-restricted genes


(12) Document PMID : 12782499              Similarity with query : 0.1797314
Title : Childhood leukemia: electric and magnetic fields as possible
        risk factors.

(13) Document PMID : 17185272              Similarity with query : 0.1782502
Title : Fonofos Exposure and Cancer Incidence in the Agricultural Health Study

...

(25) Document PMID : 15310396              Similarity with query : 0.1673389
Title : Frequency of cancer in children residing in Mexico City and treated
        in the hospitals of the Instituto Mexicano del Seguro Social (1996
```

Παράδειγμα 2 - VSM

```
Enter your query (terms separated with comma ",") : health surveys,incidence,
                                                    data collection

Results found : 158

(1) Document PMID : 15927082               Similarity with query : 0.4884273
Title : Neighborhood size and local geographic variation of health
        and social determinants

(2) Document PMID : 15035669               Similarity with query : 0.4749014
Title : Spatial correlations of mapped malaria rates with environmental
        factors in Belize, Central America

(3) Document PMID : 15312213               Similarity with query : 0.4702217
Title : Prevalence and incidence of severe sepsis in Dutch intensive care units

(4) Document PMID : 17178001               Similarity with query : 0.4682878
Title : Population distribution and burden of acute gastrointestinal
        illness in British Columbia, Canada

(5) Document PMID : 15566585               Similarity with query : 0.3593081
Title : An international sepsis survey: a study of doctors' knowledge
        and perception about sepsis

(6) Document PMID : 15574197               Similarity with query : 0.3417369
Title : Current practices in spatial analysis of cancer data: data
        characteristics and data sources for geographic studies of cancer

...
```

```
(15) Document PMID : 17640371               Similarity with query : 0.3124050
Title : Demographic determinants of acute gastrointestinal illness in
        Canada: a population study
...

(32) Document PMID : 15151704               Similarity with query : 0.0349299
Title : Comprehensive Geriatric Assessment (CGA) in general practice: Results
        from a pilot study in Vorarlberg, Austria
...
```

Παράδειγμα 2 - Semantic Similarity

```
Enter your query (terms separated with comma ",") : health surveys,incidence,
                                                    data collection

Results found : 598

(1) Document PMID : 15312213                Similarity with query : 0.5978842
Title : Prevalence and incidence of severe sepsis in Dutch intensive care units

(2) Document PMID : 16635265                Similarity with query : 0.5196444
Title : A steep decline of malaria morbidity and mortality trends in
        Eritrea between 2000 and 2004: the effect of combination
        of control methods

(3) Document PMID : 15202945                Similarity with query : 0.5120939
Title : The relationship between the Plasmodium falciparumparasite ratio
        in childhood and climate estimates of malaria transmission in Kenya

(4) Document PMID : 15282030                Similarity with query : 0.5110154
Title : Malaria morbidity and immunity among residents of villages with
        different Plasmodium falciparum transmission intensity
        in North-Eastern Tanzania

(5) Document PMID : 17559638                Similarity with query : 0.5098310
Title : Effect of meteorological factors on clinical malaria risk among
        children: an assessment using village-based meteorological stations
        and community-based parasitological survey

...

(20) Document PMID : 17425787               Similarity with query : 0.3781487
Title : Boys are more stunted than girls in Sub-Saharan Africa: a
        meta-analysis of 16 demographic

...

(25) Document PMID : 16542492               Similarity with query : 0.3728857
Title : The epidemiology of severe sepsis in England, Wales and
        Northern Ireland, 1996 to 2004: secondary analysis of a high quality
        clinical database, the ICNARC Case Mix Programme Database
```

```
(26) Document PMID : 1026402                Similarity with query : 0.3697736
Title : Mortality and cancer morbidity in a group of Swedish VCM and
        PCV production workers.

...

(29) Document PMID : 16904006               Similarity with query : 0.3649434
Title : Determinants of fruit and vegetable consumption among children and
        adolescents: a review of the literature. Part I: quantitative studies
...
```

Παράδειγμα 3 - VSM

```
Enter your query (terms separated with comma ",") : natural disasters,morbidity

Results found : 51

(1) Document PMID : 16966092                Similarity with query : 1.4419633
Title : Airborne Mold and Endotoxin Concentrations in New Orleans, Louisiana,
        after Flooding, October through November 2005

(2) Document PMID : 11940456                Similarity with query : 1.2029324
Title : The potential impact of flooding on confined animal feeding operations
        in eastern North Carolina.

(3) Document PMID : 1026402                 Similarity with query : 0.8489382
Title : Mortality and cancer morbidity in a group of Swedish VCM and PCV
        production workers.

(4) Document PMID : 15238290                Similarity with query : 0.7518327
Title : Guest Editorial: Fertilizers, Water Quality, and Human Health

(5) Document PMID : 16504156                Similarity with query : 0.6766494
Title : Variable hydrology and salinity of salt ponds in the
        British Virgin Islands

(6) Document PMID : 11834466                Similarity with query : 0.60029
Title : Biologic effects of oil fly ash.
Abstract :  Epidemiologic studies have demonstrated increased human
            morbidity and mortality

(7) Document PMID : 11675274                Similarity with query : 0.5659588
Title : Violence: an unrecognized environmental exposure that may contribute
        to greater asthma morbidity in high risk inner-city populations.

...

(15) Document PMID : 11675274               Similarity with query : 0.5659588
Title : Violence: an unrecognized environmental exposure that may contribute
        to greater asthma morbidity in high risk inner-city populations.

...
```

```
(25) Document PMID : 15686592              Similarity with query : 0.4502175
Title : Postpartum maternal morbidity requiring hospital admission in
        Lusaka, Zambia

(26) Document PMID : 12537591              Similarity with query : 0.4244691
Title : Economic evaluation of the benefits of reducing acute
        cardiorespiratory morbidity associated with air pollution
...
```

Παράδειγμα 3 - Semantic Similarity

```
Enter your query (terms separated with comma ",") : natural disasters,morbidity

Results found : 696

(1) Document PMID : 16700905               Similarity with query : 0.7519017
Title : Spatio-temporal analysis of the role of climate in inter-annual
        variation of malaria incidence in Zimbabwe

(2) Document PMID : 16504156               Similarity with query : 0.7410334
Title : Variable hydrology and salinity of salt ponds in the British
        Virgin Islands
Abstract :  Caribbean salt ponds are unique wetlands that have received
            little scientific attention.

(3) Document PMID : 11737915               Similarity with query : 0.7299459
Title : The World Trade Center Attack: Similarities to the 1988 earthquake
        in Armenia: time to teach the public life-supporting first aid?

(4) Document PMID : 16283932               Similarity with query : 0.3168838
Title : Neglected diseases of neglected populations: Thinking to reshape the
        determinants of health in Latin America and the Caribbean

(5) Document PMID : 16635265               Similarity with query : 0.6046882
Title : A steep decline of malaria morbidity and mortality trends in
        Eritrea between 2000 and 2004: the effect of combination of
        control methods

(6) Document PMID : 1026402                Similarity with query : 0.5039700
Title : Mortality and cancer morbidity in a group of Swedish VCM and
        PCV production workers.

...

(20) Document PMID : 15693992              Similarity with query : 0.4433899
Title : 11 March 2004: The terrorist bomb explosions in Madrid, Spain
Abstract :  At 07:39 on 11 March 2004, 10 terrorist bomb explosions occurred
            almost simultaneously in four commuter trains in Madrid, Spain,
            killing 177 people instantly and injuring more than 2000. There
            were 14 subsequent in-hospital deaths...
...
```

4.6 Σχολιασμός Αποτελεσμάτων

Στην προηγούμενη ενότητα παρατίθενται παραδείγματα από το ξεφύλλισμα της συλλογής. Οι όροι που δόθηκαν είναι γενικοί και απευθύνονται σε θεματικές ενότητες για τις οποίες υπάρχουν διαθέσιμα άρθρα στην παρούσα συλλογή. Τέτοιου είδους ερωτήσεις μπορούν να αξιολογήσουν τα αποτελέσματα της ομαδοποίησης με την χρήση μέτρων ακρίβειας και ανάκλησης (precision and recall).

Με βάση τα παραδείγματα παρατηρείται ότι η ομαδοποίηση των κειμένων έχει ένα καλό μέτρο ακρίβειας. Και στις τρεις ερωτήσεις ανακτήθηκαν κείμενα που η σχέση τους με τους όρους της ερώτησης φαίνεται με βάση τον τίτλο. Σε μερικές περιπωσεις χρειάστηκε να παρατεθούν και μέρη από την περίληψη των άρθρων για να φανεί καθαρά ότι απευθύνονται στο ίδιο θέμα. Τόσο η αναζήτηση με την υλοποίηση της εργασίας όσο και με την κλασσική μέθοδο VSM επιστρέφει αποτελέσματα που περιέχουν τους όρους που δίνει ο χρήστης.

Σημαντική παρατήρηση, που αποτελεί επίσης και διαφορά των δύο μεθόδων, είναι η ανάκτηση κειμένων με σημασιολογική ομοιότητα. Στην περίπτωση της μεθόδου VSM ανακτώνται μόνο τα κείμενα που περιέχουν τουλάχιστον ένα από τους όρους του χρήστη. Στην υλοποίηση της εργασίας παρατηρείται ότι ανακτώνται αποτελέσματα που έχουν σημασιολογική ομοιότητα με την ερώτηση. Στην πρώτη ερώτηση ο όρος *meningioma* έχει σχέση ομοιότητας με τον όρο *leukemia*. Ακόμη ο όρος *carcinoid tumor* έχει σχέση ομοιότητας με τον όρο *melanoma*. Αποτελέσματα με τέτοιους όρους δεν ανακτώνται με την μέθοδο VSM. Ίδια συμπεράσματα μπορούν να εξαχθούν με βάση την δεύτερη και τρίτη ερώτηση επίσης. Η δεύτερη ερώτηση περιέχει τον όρο *incidence* που σχετίζεται με τον όρο *mortality*. Στο τρίτο παράδειγμα ο όρος *natural disasters* ανταποκρίνεται σε κάθε είδους φυσική καταστροφή. Τα αποτελέσματα του ξεφυλλίσματος περιέχουν αποτελέσματα από διάφορα καιρικά φαινόμενα όπως σεισμούς, κλιματικές αλλαγές και άλλα. Επίσης ο όρος *morbidity* σχετίζεται με τον όρο *deaths* κάτι που φαίνεται στο αποτέλεσμα (20).

Συμπεράσματα για την ποιότητα της ομαδοποίησης μπορούν ακόμη να εξαχθούν παρατηρώντας τις ταυτότητες των ομάδων. Σε κάθε μέθοδο μελετήθηκαν οι MeSH όρου των ταυτοτήτων στις ομάδες. Μπορεί εύκολα να παρατηρηθεί οτι οι κοινοί όροι στις ταυτότητες είναι μεμονωμένοι. Αυτό υποδηλώνει την ύπαρξη διαφορετικών θεμάτων σε κάθε μια ομάδα, που είναι ο απώτερος σκοπός μιας ομαδοποίησης. Ενθαρρυντικό είναι ακόμη το γεγονός οτι ο αριθμός των ομάδων που δημιουργούνται με τις τρεις διαφορετικές μεθόδους είναι παραπλήσιος.

Παρακάτω αναγράφονται τα περιεχόμενα του ανεστραμμένου ευρετηρίου που περιέχει τις ομαδοποιήσεις των μεθόδων. Ο αριθμός των εγγραφών υποδηλώνει τον συνολικό αριθμό ομάδων που δημιουργούνται από κάθε μια μέθοδο. Το νούμερο στο όνομα της βάσης δεδομένων για την μέθοδο Edge-Betweenness υποδηλώνει τον αριθμό των ακμών που έχουν αφαιρεθεί.

```
    Database Information
    --------------------
(1) Database Name : BIC-MEANS
    Records count : 101

(2) Database Name : Edge-Betweenness-2700
    Records count : 117

(3) Database Name : MCL
    Records count : 98
```

Κεφάλαιο 5

Συμπεράσματα - Μελλοντική εργασία

Στο παρόν κεφάλαιο θα αναφερθούν κάποια συμπεράσματα που έχουν εξαχθεί. Στη συνέχεια θα αναφερθούν διάφορα σημεία της εργασίας στα οποία μπορεί να γίνει κάποια επέκταση ή αλλαγή που ίσως επηρεάσει τα αποτελέσματα.

5.1 Συμπεράσματα

Σκοπός της εργασίας είναι να δημιουργήσει το υπόβαθρο της ιδέας για την ομαδοποίησης του γράφου παραπομπών και ένα ευρετήριο για την εκτέλεση αναζήτησης υπό μορφή ξεφυλλίσματος. Χρησιμοποιήθηκαν τρεις μέθοδοι ομαδοποίησης, δύο ανάλυσης συνδέσμων και μια κλασσική μέθοδος ομοιότητας περιεχομένου. Ο συνδυασμός των μεθόδων έχει σκοπό την χρησιμοποίηση τόσο της πληροφορίας περιεχομένου όσο και της πληροφορίας που δίνεται μέσω των παραπομπών. Μελετώντας τα αποτελέσματα ξεφυλλίσματος φάνηκε η ακρίβεια των αποτελεσμάτων. Η περεταίρω μελέτη των αποτελεσμάτων θα μπορέσει να αξιολογήσει και την ανάκληση του συστήματος.

Η χρήση της σημασιολογικής ομοιότητας για την επιλογή και ταξινόμηση των αποτελεσμάτων δίνει στο σύστημα μια υποδομή καθώς τα σημασιολογικά μέτρα αρχίζουν να χρησιμοποιούνται όλο και περισσότερο στα σημερινά συστήματα. Με την βελτίωση των μέτρων αυτών αυτόματα βελτιώνεται και το σύστημα που τα χρησιμοποιεί.

Ο όγκος πληροφορίας για τις υλοποιήσεις των αλγορίθμων ομαδοποίησης ανάλυσης συνδέσμων είναι σημαντικός. Λόγω της μεγάλης υπολογιστικής πολυπλοκότητας η

ταχύτητα υπολογισμού των κριτηρίων σε κάθε μέθοδο αυξάνει καθώς μεγαλώνει η συλλογή. Η προσθήκη καινούργιων άρθρων συνιστά τον υπολογισμό εκ νέου των ομάδων. Υπάρχουν αλγόριθμοι ανάλυσης συνδέσμων οι οποίοι βασίζονται σε διαφορετικές ιδέες αλλά η εφαρμογή τους στην παρούσα συλλογή δεν είχε αποτελέσματα και για το λόγω αυτό δεν χρησιμοποιήθηκαν.

5.2 Μελλοντική εργασία

Σαν μελλοντική δουλειά για αυτή την διπλωματική εργασία είναι κατά πρώτο λόγω η αξιολόγηση των αποτελεσμάτων. Μπορούν να κατασκευαστούν κάποιες ερωτήσεις και να ζητηθεί από μια ομάδα ατόμων να αξιολογήσουν τα αποτελέσματα. Η δυσκολία σε αυτό είναι να βρεθούν άτομα με τις απαιτούμενες γνώσεις, ούτως ώστε να μπορούν να κατανοήσουν τα κείμενα και να γίνει σωστή αξιολόγηση.

Βελτιστοποιήσεις όσων αφορά την υλοποίηση του αλγορίθμου Edge-Betweenness είναι εφικτές. Ο επαναληπτικός αλγόριθμος της μεθόδου υπολογίζει κάθε φορά όλα τα συντομότερα μονοπάτια μεταξύ ζευγών των κόμβων και αποφασίζει να αφαιρέσει μια συγκεκριμένη ακμή. Στη συνέχεια γίνεται από την αρχή υπολογισμός όλων των μονοπατιών. Αυτό μπορεί να αποφευχθεί με ένα έξυπνο τρόπο που θα επιλέγει ποια από τα μονοπάτια έχουν επηρεαστεί από την αφαίρεση της ακμής. Αυτό βέβαια προϋποθέτει την αποθήκευση όλων των μονοπατιών, κάτι που σε μεγάλους γράφους θα είναι πολύ απαιτητικό. Με αυτή την βελτιστοποίηση η πολυπλοκότητα του αλγορίθμου θα μειωθεί δραματικά.

Μια αλλαγή που μπορεί να επιφέρει αλλαγές στην απόδοση είναι η τιμή του κατωφλίου για την επιλογή των όρων στην ταυτοποίηση των ομάδων (Κεφάλαιο 4.3). Έχει αναφερθεί ότι το κατώφλι για το C-value έχει υπολογιστεί στο άρθρο [1]. Μεταβάλλοντας αυτή την τιμή μπορούμε να έχουμε περισσότερους ή λιγότερους όρους στις ταυτότητες των ομάδων, κάτι που θα επηρεάσει με τη σειρά του την σημασιολογική ομοιότητα στο τελικό βήμα της εργασίας.

Σαν τελευταία προτεινόμενη αλλαγή είναι η εύρεση μιας άλλης λογικής για τον συνδυασμό των τριών ομαδοποιήσεων. Στην παρούσα υλοποίηση χρησιμοποιείται το αλγεβρικό άθροισμα των ομοιοτήτων. Μπορεί να δοθεί ένα διαφορετικό βάρος σε κάθε μέθοδο κάτι που θα επηρεάσει την ταξινόμηση των τελικών αποτελεσμάτων.

Βιβλιογραφία

[1] Angelos Hliaoutakis, Kalliopi Zervanou, Euripides G.M. Petrakis, Evangelos E. Milios. Automatic document indexing in large medical collections. *ACM International Workshop on Health Information and Knowledge Management (HIKM 2006)*, Νοέμβριος 2006. Arlington, VA, USA.

[2] E. Milios, Y. Zhang, B. He and L. Dong. Automatic Term Extraction and Document Similarity in Special Text Corpora. σελίδες 22–25. Αύγουστος 2003.

[3] Elsevier. Scirus - White Paper : How Scirus Works, 2004.

[4] Joshua O'Madadhain, Danyel Fisher, Padhraic Smyth, Scott White and Yan-Biao Boey. Analysis and Visualization of Network Data using JUNG, 2006.

[5] S. Ananiadou K. Franzi. The C/NC Value Domain Independent Method for Multi-Word Term Extraction. 1999.

[6] Papimeni K. Why inverse document frequency? Στο *North American Chapter of The Association for Computational Linguistics*, σελίδες 25–32, 2001.

[7] Yuhua Li, Zuhair A. Bandar και David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, Ιούλιος/Αύγουστος 2003.

[8] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, σελίδες 321–330, Μάιος 2004.

[9] Martin Rosvall and Carl T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *National Academy of Sciences of United States of America*, Δεκέμβριος 2006.

[10] Matthew J Rattigan, Marc Maier and David Jensen. Graph Clustering with Network Structure Indices. Στο *24th International Conference On Machine Learning*, 2007.

[11] Nikolaos Hourdakis. Design and Evaluation of Clustering Approaches for Large Document Collections, the "BIC-MEANS" method, Οκτώμβριος 2006.

[12] Rada Mihalcea, Courtney Corley and Carlo Strapparava. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. Στο *American Association for Artificial Intelligence*, Ιούλιος 2006. Boston.

[13] S. Ananiadou, S. Albert, and D. Schuhmann. Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline. *Genome Informatics Series*, Νοέμβριος 2000.

[14] Salton Gerard, Wong A. and Yang C. S. A Vector Space Model for Automatic Indexing, Ιούλιος 1974.

[15] Tak W. Yan and Héctor García-Molina. Index structures for selective dissemination of information under the Boolean model, Ιούνιος 1994.

[16] Stijnvan Dongen. *Graph Clustering by Flow Simulation.* Διδακτορική Διατριβή , University of Utrecht, Μάϊος 2000.

Παράρτημα Α΄

Εργαλεία

Α΄.1 PubMed Central

Το PubMed Central (PMC) είναι ένα ανοικτό αρχείο από βιοιατρική βιβλιογραφία του U.S. National Institutes of Health (NIH). Την δημιουργία και την διαχείριση αυτού του αρχείου έχει το NIH's National Center for Biotechnology Information (NCBI) που ανήκει στην Εθνική Βιβλιοθήκη Ιατρικής (National Library of Medicine (NLM)). Η ύπαρξη του αρχείου δίνει στην NLM την πρωτιά στη διατήρηση και συντήρηση μιας τόσο μεγάλης, χωρίς περιορισμό βιοιατρικής βιβλιογραφίας. Σκοπός του PMC είναι να γεμίσει το κενό μιας παγκόσμιας ψηφιακής βιβλιοθήκης. Η Εθνική Βιβλιοθήκη Ιατρικής ευελπιστεί ότι δίνοντας σε όλους τους χρήστες δωρεάν και χωρίς περιορισμούς πρόσβαση στη υλικό του αρχείου είναι ο καλύτερος τρόπος να εγγυηθεί την αντοχή στο χρόνο και την χρησιμότητα του αρχείου καθώς η τεχνολογία αλλάζει με την πάροδο του χρόνου.

Η συμμετοχή των εκδοτών στο PubMed Central είναι εθελοντική, αν και η προσθήκη των άρθρων πρέπει να υπόκειται σε κάποιο συντακτικό πρότυπο. Τα άρθρα προτιμώνται να έχουν ολόκληρη την πληροφορία τους διαθέσιμη στο PMC ούτως ώστε το αρχείο να μπορεί να γίνει ένα πραγματικό ψηφιακό αντίστοιχο της εκτεταμένης συλλογής της NLM από εκτυπωμένα άρθρα. Σε παραλληλισμό με αυτή την ιδέα, η NLM ψηφιοποιεί άρθρα που υπάρχουν μόνο σε εκτυπωμένη μορφή. Παρόλο που είναι επιθυμητή η άμεση διάθεση των άρθρων, ένα άρθρο μπορεί να καθυστερήσει για μικρό χρονικό διάστημα μετά την δημοσίευσή του.

Α΄.2 Apache Lucene

Το Apache Lucene είναι ένα πακέτο ανοικτού λογισμικού γραμμένο αποκλειστικά σε Java. Έχει πολύ καλή απόδοση και πλούσια λειτουργικότητα όσων αφορά ευρετηρίαση και αναζήτηση σε μεγάλες συλλογές κειμένων. Χρησιμοποιήθηκε το πακέτο αυτό για την αποθήκευση της χρήσιμης για την εργασία πληροφορίας από τα κείμενα του PMC (Κεφάλαιο 4.1.2). Το πακέτο διατίθεται δωρεάν στη ιστοσελίδα *http://lucene.apache.org/java/docs/index.html*.

Α΄.3 MATLAB

Το MATLAB είναι ένα μαθηματικό πακέτο και γλώσσα προγραμματισμού. Αναπτύχθηκε από την MathWorks και βασικός σκοπός του είναι ο εύκολος χειρισμός πινάκων, η γραφική αναπαράσταση συναρτήσεων και δεδομένων και η εύκολη υλοποίηση αλγορίθμων. Το πακέτο αυτό είναι ιδιαίτερα χρήσιμο στην υλοποίηση του αλγορίθμου Markov για ομαδοποίηση (Κεφάλαιο 2.1.2). Μεγάλο πλεονέκτημα του συγκεκριμένου πακέτου είναι η πολυεπεξεργασία. Στην τελευταία έκδοση του πακέτου πολλές από τις έτοιμες συναρτήσεις μπορούν να χρησιμοποιήσουν όλους τους διαθέσιμους πυρήνες του συστήματος για πράξεις που υποστηρίζουν πολυεπεξεργασία όπως πολλαπλασιασμό, πρόσθεση, αφαίρεση, ύψωση πίνακα σε δύναμη και άλλες πράξεις.

Α΄.4 MySQL

Το MySQL είναι ένα σχεσιακό σύστημα διαχείρισης βάσης δεδομένων (Relational Database Management System (RDBMS)). Την κυριότητα και επιχορήγησή του έχει η Σουηδική εταιρία MySQL AB. Αυτό το πακέτο είναι επίσης μέρος του ανοικτού λογισμικού και είναι ευρέος χρησιμοποιημένο σε πολλά συστήματα. Ο λόγος που επιλέξαμε αυτό το πακέτο ήτανε κατά κύριο λόγω ότι διανέμεται δωρεάν και για το ότι αποτελεί ένα πολύ σταθερό περιβάλλον βάσης δεδομένων για σχεσιακό σύστημα. Έχει χρησιμο-

ποιηθεί σε αυτήν την εργασία για την αποθήκευση του γράφου παραπομπών (Κεφάλαιο 4.1.1). Με ερωτήσεις σε SQL γίνεται εύκολη η ανάκτηση όλων των παραπομπών ενός άρθρου αλλά και το αντίστροφο. Να ανακτηθούν τα άρθρα που χρησιμοποιήθηκαν από για την συγγραφή ενός συγκεκριμένου άρθρου.

Α΄.5 JUNG

Είναι μια βιβλιοθήκη που προσφέρει λειτουργικότητα για την αναπαράσταση, ανάλυση και γραφική απεικόνιση δεδομένων που μπορούν να περιγραφούν σαν γράφος ή δίκτυο. Το JUNG (Java Universal Network/Graph Framework) είναι γραμμένο σε Java και αυτό του δείνει την δυνατότητα να χρησιμοποιεί τις εκτεταμένες δυνατότητές της, και επίσης άλλες ήδη υπάρχουσες βιβλιοθήκες.

Η αρχιτεκτονική της βιβλιοθήκης είναι τέτοια που επιτρέπει την υποστήριξη μιας μεγάλης ποικιλίας από αναπαραστάσεις από οντότητες και τις συνδέσεις τους, όπως κατευθυνόμενους και μη κατευθυνόμενους γράφους, γράφους με παράλληλες ακμές και υπεργράφους. Παρέχει ένα μηχανισμό για εισαγωγή σχολίων στο γράφο και ακμές με μεταδεδομένα. Αυτό διευκολύνει την δημιουργία αναλυτικών εργαλείων για πολύπλοκα δεδομένα που εξετάζουν τις σχέσεις μεταξύ οντοτήτων και επίσης τα μεταδεδομένα των οντοτήτων και σχέσεων. Η βιβλιοθήκη είναι διαθέσιμη στην ιστοσελίδα
http://jung.sourceforge.net/

Α΄.6 Oracle BerkeleyDB Java Edition

Αυτό το πακέτο είναι όπως και τα πιο πάνω ανοικτού λογισμικού. Αποτελεί και αυτό ένα περιβάλλον για βάσεις δεδομένων. Είναι γραμμένο αποκλειστικά σε Java. Όπως και το αρχικό πακέτο Oracle BerkeleyDB εκτελείται στο χώρο διευθύνσεως του προγράμματος που το τρέχει, χωρίς να είναι αναγκαία η εγκατάσταση κάποιου κεντρικού εξυπηρετητή. Αυτό επιτρέπει στον προγραμματιστή να αποθηκέψει και να ανακτήσει πληροφορία πολύ εύκολα, γρήγορα και αξιόπιστα. Δεν είναι σύστημα που υλοποιεί σχεσιακές βάσεις

δεδομένων αλλά λεξικά. Δηλαδή υπάρχουνε ζευγάρια από λέξεις κλειδιά και δεδομένα. Το χαρακτηριστικό αυτό φάνηκε ιδιαίτερα χρήσιμο στην αποθήκευση των ομάδων που εξάχθηκαν από την ομαδοποίηση (Κεφάλαιο 4.2). Για την εδραίωση μιας βάσης πρέπει να δημιουργηθεί ένα περιβάλλον (Environment) μέσα στο οποίο μπορούν να υπάρχουν πολλές βάσεις δεδομένων.