ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΕΘΟΔΟΛΟΓΙΕΣ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΜΕ ΧΡΗΣΗ ΟΝΤΟΛΟΓΙΩΝ ΣΕ ΣΥΛΛΟΓΕΣ ΙΑΤΡΙΚΩΝ ΚΕΙΜΕΝΩΝ

ΣΙΣΚΟΣ ΔΗΜΗΤΡΙΟΣ

ΙΟΥΛΙΟΣ 2008

ΕΠΙΒΛΕΠΩΝ

ΑΝ. ΚΑΘΗΓΗΤΗΣ ΠΕΤΡΑΚΗΣ ΕΥΡΙΠΙΔΗΣ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

ΑΝ. ΚΑΘΗΓΗΤΗΣ ΠΕΤΡΑΚΗΣ ΕΥΡΙΠΙΔΗΣ
ΕΠ. ΚΑΘΗΓΗΤΗΣ ΔΕΛΗΓΙΑΝΝΑΚΗΣ ΑΝΤΩΝΙΟΣ
ΕΠ. ΚΑΘΗΓΗΤΗΣ ΛΑΓΟΥΔΑΚΗΣ ΜΙΧΑΗΛ Γ.

Περιεχόμενα

Περίληψη

Το πρόβλημα που μελετάται είναι μεθοδολογίες ανάκτησης πληροφορίας με χρήση οντολογιών σε συλλογές ιατρικών κειμένων. Πιο συγκεκριμένα θέλουμε να ερευνήσουμε τις συνέπειες που θα προκύψουν στο πεδίο της Ανάκτησης Πληροφορίας, αν προσαρμόσουμε μια οντολογία σε μία συγκεκριμένη συλλογή κειμένων. Στη δουλειά μας χρησιμοποιήσαμε ως οντολογία το MeSH[1] ενισχυμένο με σχέσεις που πήραμε από το Metathesaurus[2] και ως συλλογή το Ohsumed[3]. Τους πιο σημαντικούς όρους από το Ohsumed τους εξάγαμε με την μέθοδο Εξαγωγής Όρων AMTEx. Για να πάρουμε τη σημασιολογική ομοιότητα μεταξύ των όρων της οντολογίας μας δημιουργήσαμε μία μέθοδο σημασιολογικής ομοιότητας, την X-SimilarityMetathesaurus.

Τέλος η εκτίμηση των αποτελεσμάτων μας έγινε με χρήση της μεθόδου Ανάκτησης Πληροφορίας που πρότεινε η Rada Mihalcea πάνω στην ιατρική συλλογή κειμένων Ohsumed. Εκεί γίνεται η σύγκριση της περιορισμένης με την μεγαλύτερη οντολογία με διαγράμματα precision-recall.

1 http://www.nlm.nih.gov/mesh/meshhome.html
2 http://0-www.nlm.nih.gov.catalog.llu.edu/pubs/factsheets/umlsmeta.html
3 http://ir.ohsu.edu/ohsumed/ohsumed.html

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέπων Αναπληρωτή Καθηγητή μου κύριο Ευριπίδη Πετράκη για τις πολύτιμες συμβουλές του και τη στήριξή του κατά τη διάρκεια της συνεργασία μας. Ακόμα θέλω να ευχαριστήσω τα μέλη της εξεταστικής μου επιτροπής κυρίους Δεληγιαννάκη Αντώνιο και Λαγουδάκη Μιχαήλ Γ.

Επίσης θα ήθελα να ευχαριστήσω όλους τους ανθρώπους του εργαστηρίου που με βοήθησαν όποτε τους το ζήτησα και πιο συγκεκριμένα τους Α. Χλιαουτάκη, Ε. Δρυμώνα, Ε. Μπαράτη, Σ. Μπατσάκη, Σ. Πέτρου, Σ. Αργυρόπουλο και Κ. Ζερβάνου.

Κλείνοντας θα ήθελα να ευχαριστήσω όλους τους φίλους μου που με στήριξαν και με στηρίζουν, ενώ ιδιαιτέρως τους: Πέτρο, Τάσο, Νικηφόρο, Στέφανο, Πέγκυ. Τέλος ευχαριστώ με ιδιαίτερη θέρμη τον αδερφό μου Αλέξανδρο και τους γονείς μου για τη φροντίδα τους.

1. Εισαγωγή

Η Ανάκτηση Πληροφορίας, είναι ένα αντικείμενο το οποίο ερευνάται με ιδιαίτερη ένταση τα τελευταία είκοσι χρόνια[1]. 'Οταν μιλάμε για Ανάκτησης Πληροφορίας αναφερόμαστε στον εντοπισμό μιας συγκεκριμένης πληροφορίας που αναζητούμε. Η πληροφορία αυτή μπορεί να είναι ολόκληρα κείμενα, μέρος κειμένων, ή metadata τα οποία περιγράφουν κείμενα. Εφαρμογές Α.Π. βρίσκουμε από απλά συστήματα βάσεων δεδομένων έως και σε μηχανές αναζήτησης στο διαδίκτυο. Η βασική ιδέα λειτουργίας στις μηχανές αναζήτησης είναι η επιστροφή κειμένων στο χρήστη, που να περιέχουν όρους τους οποίους ο χρήστης έχει γράψει στην ερώτηση (query). Εδώ αξίζει να σημειωθεί ότι τα αποτελέσματα οφείλονται σε δύο παράγοντες: στην ακρίβεια του ερωτήματος και στο μοντέλο ανάκτησης πληροφορίας που χρησιμοποιείται. Για το πρώτο ευθύνεται ο χρήστης και δεν μπορούμε να επέμβουμε σε μεγάλο βαθμό, ενώ για το δεύτερο ο ερευνητής και είναι εύκολα κατανοητό ότι εκεί εστιάζονται οι προσπάθειες βελτίωσης. Η Ανάκτηση Πληροφορίας με τα κλασικά μοντέλα, δηλαδή Vector Space, Probabilistic και Boolean model, βασίζεται στην λεξικογραφική ομοιότητα και μόνο, των όρων του ερωτήματος με όρους του κειμένου. Σε αυτό το σημείο γίνεται αντιληπτό ότι αυτά τα απλά μοντέλα δεν επιστρέφουν κείμενα τα οποία μπορεί να περιέχουν συνώνυμους ή παραπλήσιους σημασιολογικά όρους με αυτούς του ερωτήματος. Για την αντιμετώπιση αυτής της αδυναμίας χρησιμοποιήσαμε ένα μοντέλο Ανάκτησης (Mihalcea) το οποίο χρησιμοποιεί και σημασιολογική ομοιότητα μεταξύ όρων. Αυτό βέβαια προϋποθέτει τη χρήση κάποιας μεθόδου σημασιολογικής ομοιότητας (similarity measure).

Σε αυτό το σημείο μας βοηθάει η οντολογία της οποίας η ύπαρξη, η χρήση και ο τρόπος κατασκευής της θα αναλυθεί στο κεφάλαιο 2. Αυτή τη στιγμή αρκεί να γνωρίζουμε ότι στα πεδία της επιστήμης υπολογιστών και της επιστήμης των πληροφοριών, οντολογία είναι μια περιγραφή ενός συνόλου εννοιών σε μία γνωστική περιοχή, καθώς επίσης και σχέσεις οι οποίες συνδέουν αυτές τις έννοιες[1]. Με άλλα λόγια είναι ένας τρόπος περιγραφής της πληροφορίας σε ένα συγκεκριμένο γνωστικό πεδίο.

Έτσι λοιπόν δημιουργήσαμε μία οντολογία την οποία φτιάξαμε εμπλουτίζοντας το MeSH thesaurus με σχέσεις από το Metathesaurus. Το MeSH περιέχει μόνο IS-A σχέσεις. Στηριζόμενοι λοιπόν σε αυτήν την οντολογία έχουμε πλέον τη δυνατότητα να χρησιμοποιήσουμε μεθόδους σημασιολογικής ομοιότητας, οι οποίες στηρίζονται στα υβριδικά σημασιολογικά κριτήρια και να βρούμε σημασιολογική ομοιότητα μεταξύ όλων των όρων που εμφανίζονται στην οντολογία μας. Μία τέτοια μέθοδο υλοποιήσαμε.

Οι μέθοδοι σημασιολογικής ομοιότητας χωρίζονται σε τέσσερις κύριες κατηγορίες: υπολογισμού ακμής (Edge counting),

1 http://en.wikipedia.org/wiki/Ontology_%28computer_science%29

περιεχομένου πληροφορίας (Information content), χαρακτηριστικών (feature based) και υβριδικών (hybrid).

Τέλος παίρνουμε τα ζευγάρια των όρων και τα εισάγουμε στην μέθοδο Ανάκτησης Πληροφορίας της Mihalcea, εκτιμώντας τα αποτελέσματα της με ένα precision-recall διάγραμμα μαζί με άλλες μεθόδους. Επίσης αναζητούμε να βρούμε τι γίνεται αν μειώσουμε την οντολογία μας και τι αποτελέσματα θα έχει αυτό στο precision-recall διάγραμμα.

2.Σχετική Δουλειά Και Υπόβαθρο

Σε αυτό το κεφάλαιο θα μιλήσουμε για το θεωρητικό κομμάτι που χρειάζεται κανείς να γνωρίζει για την κατανόηση αυτής της εργασίας, καθώς επίσης και για σχετικές δουλειές που έχουν υλοποιηθεί στο παρελθόν.

2.1 Οντολογίες

Αρχικά είναι σκόπιμο να δούμε από που προέρχεται ο όρος οντολογία .Στην φιλοσοφία είναι το βασικότερο παρακλάδι της μεταφυσικής, και χωρίζεται στις λέξεις ον και -λογία που παραπέμπει στην επιστήμη. Δηλαδή είναι η επιστήμη που ασχολείται με το ον. Αυτό μου μελετάει η επιστήμη αυτή εξετάζει την ύπαρξη, τις βασικές της κατηγορίες και σχέσεις που υπάρχουν σε αυτή. Μία οντότητα περιγράφεται από ουσιαστικά. Εάν το δούμε από την πλευρά της επιστήμης υπολογιστών, θα διαπιστώσουμε ότι δεν διαφέρει πάρα πολύ από την φιλοσοφική εξήγηση. Είναι ένα σύνολο από οντότητες(ουσιαστικά) που συνδέονται μεταξύ τους κυρίως με IS-A σχέσεις, αλλά και με άλλες, με σκοπό την αναπαράσταση ενός συγκεκριμένου ή ενός γενικού γνωστικού αντικειμένου.

Τα προβλήματα που υπάρχουν και καλούνται να λύσουν οι οντολογίες, είναι στον τομέα της επικοινωνίας μεταξύ ατόμων, συστημάτων και προγραμμάτων. Έπρεπε δηλαδή να βρεθεί ένας τρόπος ώστε όταν χρησιμοποιείται μία συγκεκριμένη ορολογία να εννοούνται συγκεκριμένα πράγματα, έτσι ώστε να διευκολυνθεί η επικοινωνία και να μπορέσει κατ' επέκταση να αυτοματοποιηθεί. Επίσης εάν υπάρχει κοινά αποδεκτός τρόπος αναπαράστασης της γνώσης, για παράδειγμα η XML, τότε θα είναι πιο εύκολο για οποιοδήποτε σύστημα ή πρόγραμμα να την αποκωδικοποιήσει και να τη χρησιμοποιήσει για δικούς του σκοπούς.

Οι οντολογίες χρησιμοποιούνται σε πάρα πολλές εφαρμογές. Αυτές μπορεί να είναι τεχνητή νοημοσύνη (Artificial Intelligence), σημασιολογικός ιστός (Semantic Web), τεχνητό λογισμικό (software engineering), βιοατρική τεχνολογία (biomedical infirmatics), library science και αρχιτεκτονική της πληροφορίας (information architecture).

Τώρα είναι καλό να αναφερθούμε σε οντολογίες που ήδη υπάρχουν. Επειδή οντολογίες σύμφωνα με τον ορισμό είναι δύσκολο να υπάρξουν θα αναφερθούμε και σε οντολογίες ελαφριάς μορφής όπως είναι οι θησαυροί οι βάσεις γνώσης.

Ας ξεκινήσουμε από τις <u>γενικές</u> οντολογίες:

(1) Wordnet[20]: Μοντελοποίηση του Αγγλικού λεξιλογίου. Μέσα βρίσκονται ουσιαστικά, ρήματα, επιρρήματα και επίθετα τα οποία είναι οργανωμένα σε ομάδες συνωνύμων. Επίσης χρησιμοποιείται και ως λεξικό και ως θησαυρός.

(2) SENSUS[21]: Είναι ένας θησαυρός 90.000 κόμβων, ο οποίος

είναι επέκταση και διαφοροποίηση του Wordnet. Κάθε κόμβος αναπαριστά μία έννοια ενώ οι κόμβοι συνδέονται μεταξύ τους με IS-A σχέσεις. Όσο πλησιάζουμε προς τη ρίζα του δέντρου συναντάμε πιο γενικούς όρους.

(3) Cyc[22,23]: Είναι μια βάση γνώσης(Knowledge Base), που περιέχει 200.000 όρους και σχέσεις οι οποίες συνδέουν τους όρους αυτούς. Περιέχει πλήθος από βασική ανθρώπινη πληροφορία.

Οντολογίες <u>συγκεκριμένου γνωστικού πεδίου</u>:

(1) Ιατρικές έννοιες: UMLS[24], SNOMED, MeSH, Metathesaurus. Το UMLS (Unified Medical Language System) περιέχει οντολογίες πολλών χρήσεων και πολλών γλωσσών, με αντικείμενο την ιατρική και την υγεία. Αυτές οι οντολογίες είναι το Metathesaurus, το SPECIALIST Lexicon and Lexical Tools και το Semantic Network. Συγκεκριμένα διαθέτει παραπάνω από 1 εκατομμύριο βιοϊατρικούς όρους και πάνω από 2,8 εκατομμύρια όρους από controled vocabularies και classifications, που βρίσκονται σε αρχεία ασθενών, έγκυρα δεδομένα υγείας, expert συστήματα, βιβλιογραφικές και ολόκληρων κειμένων βάσεις. Επίσης οι όροι συνδέονται με διάφορες σχέσεις.

(2) Έννοιες χώρου: SDTS[1]. Το SDTS περιγράφει μία οντολογία από αντικείμενα χώρου και σχέσεις που τα συνδέουν. Οι έννοιες της οντολογίας αυτής χρησιμοποιούνται για τοπογραφικούς

1 http://mcmcweb.er.usgs.gov/sdts/

και υδρογραφικούς χάρτες.

Γλώσσες Για Κατασκευή Οντολογιών

RDF (Resource Description Framwork): Η RDF[2] είναι μία γλώσσα η οποία χρησιμεύει για αναπαράσταση πληροφορίας στο διαδίκτυο. Συγκεκριμένα σώζει meta data πληροφορία του διαδικτύου, όπως είναι ο τίτλος, ο συγγραφέας και ο χρόνος κατασκευής ενός αρχείου. Χρησιμοποιείται επίσης και σε καταστάσεις όπου η πληροφορία πρέπει να προσπελαστεί από εφαρμογές, εφόσον προσφέρει ένα κοινό τρόπο αναπαράστασης της γνώσης. Επομένως δίνεται και η δυνατότητα ανταλλαγής πληροφορίας, χωρίς τον κίνδυνο της απώλειας. Η RDF για να περιγράψει πράγματα του διαδικτύου μέσω απλών λέξεων και αριθμών βασίζεται στους URIs (Uniform Resource Identifiers). Αυτό επιτρέπει την αναπαράσταση της πληροφορίας μέσω γράφου κόμβων και ακμών. Επιπρόσθετα η RDF μπορεί να συνταχθεί και με XML (RDF/XML) έτσι ώστε να μπορεί να αποθηκεύει και να ανταλλάζει αυτούς τους γράφους. Τέλος υπάρχει και το RDF Schema (RDF-S), το οποίο συνδέει τις κλάσεις μεταξύ τους.

---

2 http://www.w3.org/RDF

DAML+OIL: Η DAML+OIL[3] είναι μια σημασιολογική γλώσσα υλικού του διαδικτύου. Χτίστηκε πάνω στην RDF και RDF-S προσθέτοντας νέα πράγματα. Συγκεκριμένα όρισε RDF τριπλέτες(RDF triples : subject, predicate, object).

OWL: Η OWL[4] (Ontology Web Language) είναι η πιο διαδεδομένη γλώσσα κατασκευής οντολογίας, η οποία δημιουργήθηκε από την Web Ontology Working Group. Είναι μια γλώσσα κατασκευής και ανταλλαγής οντολογιών. Περιγράφει όρους και δείχνει με ποιες σχέσεις συνδέονται αυτές μεταξύ τους. Από ένα κώδικα της OWL γίνεται αντιληπτό ότι στηρίζεται στις RDF και DAML+OIL.

TOWL: Η TOWL[5] είναι συνέχεια της OWL. Η σημαντική καινοτομία σε σχέση με την OWL είναι η εισαγωγή του χρόνου, και η περιγραφή μιας ενδεχόμενης αλλαγής ενός αντικειμένου. Έτσι δίνεται η δυνατότητα καλύτερης αναπαράστασης του πραγματικού κόσμου. Πρέπει να πούμε ότι αυτή η γλώσσα είναι ακόμα σε ερευνητικό επίπεδο.

Τώρα θα περιγράψουμε το MeSH και το Metathesaurus, δύο εργαλεία τα οποία χρησιμοποιήθηκαν για την έρευνά μας.

---

3 http://www.daml.org/language/

4 http://www.w3.org/TR/owl-features

5 http://www.towl.org/

2.2 MeSH (Medical Subject Headings)

Το MeSH είναι ένας θησαυρός ο οποίος αποτελείται από ιατρικούς όρους. Όταν μιλάμε για θησαυρό, εννοούμε όρους οι οποίοι είναι συνδεδεμένοι μεταξύ τους μόνο με σχέσεις IS-A.

Αναλυτικότερα αποτελείται από ομάδες όρων, που ονομάζονται επικεφαλίδες (descriptors) και είναι ταξινομημένες σε ιεραρχική και αλφαβητική δομή. Στην κορυφή της ιεραρχίας βρίσκονται πολύ γενικοί όροι, όπως Anatomy και Mental Disorders ενώ όσο κατεβαίνουμε εξειδικεύονται, όπως Ancle Conduct Disorder . Στο MeSH βρίσκονται 24.767 επικεφαλίδες και 97.000 entry terms. Τα entry terms είναι όροι, από τους οποίους ο καθένας αντιστοιχεί σε μία επικεφαλίδα. Η αντιστοίχηση αυτή συνήθως δηλώνει συνωνυμία. Για παράδειγμα ο Vitamin C είναι entry term της επικεφαλίδας Ascorbic Acid .Επίσης το MeSHαποτελείται από 16δέντρα όπως φαίνεται στο παρακάτω σχήμα. Ο θησαυρός του MeSH είναι αρκετά φαρδύς και έχει μικρό βάθος (11 επίπεδα).

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Natural Sciences [H]
9. Anthropology, Education, Sociology and Social Phenomena [I]
10. Technology, Industry, Agriculture [J]
11. Humanities [K]
12. Information Science [L]
13. Named Groups [M]
14. Health Care [N]
15. Publication Characteristics [V]
16. Geographicals [Z]

Στο παρακάτω σχήμα φαίνονται τα παιδιά του Anatomy .

1. ⊟ Anatomy [A]
    - Body Regions [A01] +
    - Musculoskeletal System [A02] +
    - Digestive System [A03] +
    - Respiratory System [A04] +
    - Urogenital System [A05] +
    - Endocrine System [A06] +
    - Cardiovascular System [A07] +
    - Nervous System [A08] +
    - Sense Organs [A09] +
    - Tissues [A10] +
    - Cells [A11] +
    - Fluids and Secretions [A12] +
    - Animal Structures [A13] +
    - Stomatognathic System [A14] +
    - Hemic and Immune Systems [A15] +
    - Embryonic Structures [A16] +
    - Integumentary System [A17] +
2. ⊞ Organisms [B]

Οι όροι του MeSH χρησιμοποιούνται για δεικτοδότηση στο MEDLINE.

2.3 Metathesaurus

Το Metahesaurus δημιουργήθηκε από την Natural Library of Medicine(NLM), στα πλαίσια μίας μεγαλύτερης εργασίας, του Unified Medical Language System(UMLS). Ο σκοπός του UMLS, είναι η κατανόηση ιατρικών όρων από συστήματα υπολογιστών. Στο UMLS υπάρχουν τρεις πηγές γνώσης: Metathesaurus, Semantic Network, SPECIALIST Lexicon and Lexical Tools.

Εμείς λοιπόν θα εστιάσουμε στο Metathesaurus το οποίο χρησιμοποιήσαμε στην εργασία μας. Το Metathesaurus είναι μία βάση δεδομένων, πολλών σκοπών και πολυγλωσσική, η οποία περιέχει βιοϊατρικές και σχετιζόμενες με την υγεία έννοιες, τις διάφορες ονομασίες των εννοιών και σχέσεις οι οποίες τις συνδέουν. Το Metathesaurus χτίστηκε από ηλεκτρονικές αποδόσεις θησαυρών, classification, code sets και από ομάδες όρων που βρέθηκαν σε συλλογές από ιστορικά ασθενών, παραστατικά υγείας, δημόσιων ιατρικών στατιστικών δεδομένων, ιατρικής βιβλιογραφίας και ερευνητικών δεδομένων. Οι παραπάνω πηγές καλούνται ως λεξιλογικές πηγές (source vocabularies) του Metahesaurus. Παραδείγματα λεξιλογικών πηγών είναι τα MeSH, Alternative billing concepts 2008, Alcohol and other Drugs thesaurus, Medical entities Dictionary και άλλα πολλά.

Οι όροι και οι σχέσεις του Metathesaurus βγαίνουν από τον συνδυασμό των λεξιλογικών πηγών, καθώς επίσης και από έρευνα που έχει γίνει από την NLM στα πλαίσια του UMLS όλα αυτά τα χρόνια. Επίσης αξίζει να σημειωθεί ότι εάν ένας όρος υπάρχει σε περισσότερα από μία λεξιλογική πηγή με διαφορετική ερμηνεία, τότε περιλαμβάνονται όλες οι ερμηνείες διευκρινίζοντας βέβαια από ποια λεξιλογική πηγή προήλθαν. Οι σχέσεις από την πλευρά τους χωρίζονται σε σχέσεις μεταξύ όρων συγκεκριμένου λεξιλογίου(intra-source vocabulaty relationships) και από διαφορετικών λεξιλογίων(inter-source vocabulary relationships). Για την πρώτη περίπτωση σχέσεων εισάγονται και σχέσεις οι οποίες έχουν προέλθει από στατιστικά στοιχεία εμφάνισης όρων μεταξύ τους σε μία βάση κειμένων.

Οι σχέσεις που εμφανίζονται στο Metahesaurus είναι οι εξής 12:

AQ: *Αν ένας όρος εμφανιστεί με κάποιον άλλον του αλλάζει τη σημασία.*

CHD: *Σχέση παιδιού από μία λεξιλογική πηγή του Metathesaurus.*

DEL: *Έχει αφαιρεθεί.*

PAR: *Σχέση πατέρα από μία λεξιλογική πηγή του Metathesaurus.*

QB: *Το αντίστροφο της AQ.*

RB: *Ένας όρος με κάποιον άλλον έχουν μακρινή σημασιολογικά σχέση.*

RN: *Ένας όρος με κάποιον άλλον έχουν κοντινή σημασιολογικά σχέση.*

RO: *Ένας όρος με κάποιον άλλον έχουν σχέση η οποία δεν είναι συνώνυμη, κοντινή ή μακρινή.*

RQ: *Υπάρχει σχέση πιθανώς συνωνυμίας.*

SIB: *Αδερφική σχέση από μία λεξιλογική πηγή του Metathesaurus*

SY: *Σχέση συνωνυμίας.*

XR: *Δεν υπάρχει σχέση.*

2.4 Εξαγωγή Όρων (Information Extraction)

Η εξαγωγή όρων σχετίζεται με την επιλογή και εξαγωγή των πιο σημαντικών όρων ενός ή περισσότερων κειμένων. Μπορεί να χρησιμοποιηθεί για δεικτοδότηση και για βοήθεια σε συστήματα ανάκτησης πληροφορίας. Παραδείγματα μεθόδων εξαγωγής όρων είναι οι C/NC Value [25], MMTx[7] και AMTEx [26]. Οι μέθοδοι αυτοί επιστρέφουν πολυλεκτικούς όρους, το οποίο είναι πολύ σημαντικό αφού οι πολυλεκτικοί όροι περιέχουν περισσότερη και πιο συγκεκριμένη πληροφορία από τους μονολεκτικούς. Επίσης η προσπάθεια των μεθόδων αυτών είναι να επιστρέψουν όρους οι οποίοι να περιγράφουν το γνωστικό αντικείμενο των κειμένων.

Εμείς θα ασχοληθούμε με την μέθοδο AMTEx, η οποία δίνει τα καλύτερα αποτελέσματα και την χρησιμοποιήσαμε στην εργασία μας. Η AMTEx όμως στηρίζεται σε μεγάλο βαθμό στην C/NC Value και αξίζει να αναφέρουμε μερικά λόγια και για αυτήν.

C/NC Value

Το C/NC Value εξάγει πολυλεκτικούς και εμφωλεμένους όρους από οποιαδήποτε συλλογή κειμένων . Είναι μια υβριδική μέθοδος η οποία χρησιμοποιεί και γλωσσολογική αλλά και στατιστική πληροφορία. Μέσα από γλωσσολογικά φίλτρα που περιέχουν ουσιαστικά, επίθετα και προθέσεις, εξάγονται όροι από τα κείμενα .

7http://mmtx.nlm.nih.gov

Ύστερα έρχεται το στατιστικό κομμάτι της μεθόδου, που έχει να κάνει με τη συχνότητα εμφάνισης όρων και με τον αριθμό των όρων μία λέξη εμφανίζεται μαζί στα κείμενα. Αναλυτικότερα έχουμε:

C-Value:

$$C-value=\begin{cases} \log_2|a|f(a), \\ \log_2|a|(f(a)-\frac{1}{P(T_a)}\sum_{b\epsilon T_a} f(b)) \end{cases}$$

Η πρώτη σχέση είναι για μη εμφωλευμένους όρους και η δεύτερη για όρους που εμφανίζονται στα κείμενα εμφωλευμένοι. Όπου a είναι ο όρος, f(a) η συχνότητα εμφάνισής του, |a| το μήκος του όρου σε λέξεις, Ta το σύνολο των όρων που περιέχουν τον a και τέλος το P(Ta) είναι το πλήθος των όρων αυτού του συνόλου.

CF(a):

$$CF(a)=\Sigma_{w\epsilon C_a} f_a(w) * weight(w)$$

όπου Ca το σύνολο των λέξεων που εμφανίζονται μαζί με τον όρο a,

όπου fa(w) η συχνότητα εμφάνισης της λέξης w με τον όρο a,

όπου weight(w): $weight(w)=\frac{t(w)}{n}$

όπου t(w) είναι το πλήθος των όρων με τους οποίους εμφανίζεται η λέξη w, όπου n το πλήθος όλων των όρων.

NC- Value:

$$NC-Value(a)=0.8*CValue(a)+0.2*CF(a)$$

Βλέπουμε ότι το NC-Value τα ενώνει όλα τα γλωσσολογικά και στατιστικά στοιχεία μας.

2.5 AMTEx

Η AMTEx είναι μια μέθοδος εξαγωγής όρων, η οποία δεν εφαρμόζεται σε ανεξάρτητες συλλογές κειμένων αλλά σε ιατρικές, όπως είναι το MEDLINE και το OHSUMED. Αυτό έδωσε τη δυνατότητα στη μέθοδο να προσθέσει και άλλες πηγές πληροφορίας όπως είναι το MeSH. Αυτός είναι ο λόγος που έδωσε καλύτερα αποτελέσματα από το C-NC Value.

Τα βήματα που ακολουθεί το AMTEx φαίνονται παρακάτω:

(1) Χρήση του C/NC Value και εξαγωγή πολυλεκτικών όρων.

(2) Ταξινόμηση των όρων σύμφωνα με το NC-Value.

(3) Διατήρηση μόνο των όρων που βρίσκονται στο MeSH.

(4) Εισαγωγή από το MeSH και των μονολεκτικών όρων που

βρίσκονται εμφωλευμένοι στους πολυλεκτικούς όρους.

(5) Πρόσθεση stemmed όρων.

(6) Πρόσθεση όρων από το MeSH που είναι σημασιολογικά όμοιοι με τους ήδη υπάρχοντες. Αυτό γίνεται με την εισαγωγή όρων από το MeSH που βρίσκονται γύρω από τους ήδη εξαγώμενους όρους, με κριτήριο κάποιο κατώφλι (threshold).

Είναι φανερό λοιπόν γιατί το AMTEx δίνει καλύτερα αποτελέσματα από το C/NC-Value σε συλλογές ιατρικών κειμένων.

2.6 Ανάκτηση Πληροφορίας (Information Retrieval)

Το πεδίο της Ανάκτησης Πληροφορίας είναι πολύ σημαντικό και έχει απασχολήσει πολλούς ερευνητές και μεγάλες εταιρίες (eg. Google). Η επιλογή κειμένων μέσω της Ανάκτησης Πληροφορίας στηρίζεται στην ομοιότητα μεταξύ κειμένων. Όταν μιλάμε για κείμενο μπορεί να εννοούμε και ερώτημα. Τέτοια μοντέλα που υπολογίζουν ομοιότητα μεταξύ κειμένων έχουν προταθεί πολλά. Κυριότερο είναι το Vector Space Model το οποίο είναι και το ποιο γρήγορο αλλά δεν δίνει τα καλύτερα αποτελέσματα επειδή ψάχνει μόνο την λεξιλογική ομοιότητα μεταξύ των όρων των κειμένων και όχι την σημασιολογική. Αντίθετα το μοντέλο που πρότεινε η Rada Mihalcea στη εργασία της Corpus-based and Knowledge-based Measures of Text

Semantic Similarity [33] εισάγει και την σημασιολογική ομοιότητα.

Rada Mihalcea 's Model

Στην εργασία της η Rada Mihalcea μιλάει για λέξεις και όχι για όρους. Στο σημείο αυτό θα μιλήσουμε και εμείς για λέξεις γιατί περιγράφουμε την αυθεντική εργασία.

Για να λειτουργήσει το σημασιολογικό μοντέλο χρειάζεται ζευγάρια λέξεων και την ομοιότητά τους, την οποία παίρνουμε από κάποια μέθοδο σημασιολογικής ομοιότητας, όπως για παράδειγμα είναι η Lietal. Ύστερα μπορούμε να επεκτείνουμε (expand) το ερώτημά μας και να εισάγουμε συνώνυμες λέξεις με αυτές του ερωτήματος. Επίσης η μέθοδος παίρνει υπόψη της και το πόσο σημαντική είναι μία λέξη στο κείμενο, με κριτήριο το idf (inverse document frequency), που βρίσκεται με τη διαίρεση του συνολικού αριθμού των κειμένων, με τον αριθμό των κειμένων που περιέχουν τη λέξη. Η εξίσωση που περιγράφει την ομοιότητα μεταξύ μεταξύ των κειμένων είναι η εξής:

$$sim(T_{1,}T_2)=1/2\left(\frac{\Sigma_{w\epsilon T_1} maxSim(w,T_2)*idf(w)}{\Sigma_{w\epsilon T_1} idf(w)} + \frac{\Sigma_{w\epsilon T_2} maxSim(w,T_1)*idf(w)}{\Sigma_{w\epsilon T_2} idf(w)}\right) \quad (2.6.1)$$

Δηλαδή για κάθε λέξη του πρώτου κειμένου κοιτάει να βρει με ποια από όλες τις λέξεις έχει τη μεγαλύτερη ομοιότητα και τις προσθέτει. Ύστερα κάνει το ίδιο για το δεύτερο κείμενο ως προς το πρώτο. Δηλαδή από τη σχέση είναι φανερό ότι πρόθεση είναι να βρεθεί η ομοιότητα των κειμένων και όχι η ανομοιότητα.

2.7 Συλλογή κειμένων

Πολύ σημαντικό κομμάτι είναι η συλλογή κειμένων που επιλέγει κανείς να εφαρμόσει την δουλειά του. Εμάς η δουλειά μας θέλαμε να γίνει πάνω σε ιατρική συλλογή και για αυτόν το λόγο διαλέξαμε το Ohsumed.

Ohsumed

Το Ohsumed είναι μία συλλογή 348.566 αναφορών (περιλήψεων) από την διαδικτυακή βάση πληροφοριών του MEDLINE. Αποτελείται από περιλήψεις 270 ιατρικών περιοδικών της περιόδου 1987-1991. Τα πεδία που περιέχονται για κάθε ένα κείμενο είναι ο τίτλος (title), συγγραφέας (author), πηγή (source), όροι δεικτοδότησης (indexing terms) και τύπος δημοσίευσης (publication type). Αξίζει να σημειωθεί ότι το Ohsumed είναι ελεύθερο στο κοινό για ερευνητικές ανάγκες. Παρακάτω φαίνονται τα πεδία του Ohsumed που αναφέραμε πιο πάνω:

.I sequential identifier

(important note: documents should be processed in this order)

.U MEDLINE identifier (UI)

(<DOCNO> used for relevance judgements)

.M Human-assigned MeSH terms (MH)

.T Title (TI)

.P Publication type (PT)

.W Abstract (AB)

.A Author (AU)

.S Source (SO)

Επίσης μαζί με το Ohsumed υπάρχουν και 63 ερωτήσεις (queries), οι οποίες έχουν 2 πεδία : τίτλο (title) και περιγραφή (description). Για πειράματα Ανάκτησης Πληροφορίας δόθηκαν για κάθε ερώτηση οι σχετικές απαντήσεις από τις περιλήψεις που έχει το Ohsumed.

2.8 Σημασιολογική Ομοιότητα

Είναι πολύ σημαντικό να μπορούμε να υπολογίσουμε την τη σημασιολογική ομοιότητα (semantic similarity), μεταξύ όρων μιας οντολογίας ή και διαφορετικών. Το αποτέλεσμα αυτής της διεργασίας μπορεί να χρησιμοποιηθεί έτσι ώστε να έχουμε καλύτερα αποτελέσματα σε ένα Α.Π. σύστημα. Αυτό συμβαίνει λόγω του γεγονότος ότι δεν κοιτάει το Α.Π. μόνο λεξιλογική ομοιότητα μεταξύ των όρων, όπως δηλαδή κάνει ένα απλό σύστημα σαν το VSM (Vector Space Model), αλλά και σημασιολογική. Τα τέσσερα κύρια κριτήρια που χρησιμοποιούν οι μέθοδοι υπολογισμού σημασιολογικής ομοιότητας είναι τα εξής:

- Κριτήριο υπολογισμού ακμής
- Κριτήριο περιεχομένου πληροφορίας
- Κριτήριο βασιζόμενο στα χαρακτηριστικά
- Υβριδικό κριτήριο

2.8.1 Μέθοδοι Υπολογισμού ακμής (Edge Counting Measures)

Αυτό το κριτήριο στηρίζεται στη θέση που βρίσκονται οι δύο όροι μεταξύ τους στην ταξινομία. Οι μέθοδοι που θα εξηγηθούν παρακάτω βασίζονται στην spreading activation theory[4,5]. Η βασική ιδέα αυτής της θεωρίας είναι ότι, όσο πιο κοντά στην ταξινομία βρίσκονται δύο όροι και όσο περισσότερα μονοπάτια τους συνδέουν τόσο πιο όμοιοι είναι.

Shortest Path[6,7]: Αυτή η μέθοδος στηρίζεται στην ιδιότητα, ότι όσο πιο κοντά στην ταξινομία βρίσκονται δύο όροι μεταξύ τους τόσο πιο όμοιοι είναι. Η εξίσωση που περιγράφει την μέθοδο αυτή είναι η εξής:

$$sim = 2MAX - L \qquad (2.2.1),$$

όπου MAX: το μέγιστο μονοπάτι μεταξύ των όρων και

L : το μικρότερο μονοπάτι μεταξύ των όρων.

Weighted Links[8]: Η μέθοδος αυτή θα μπορούσε να χαρακτηριστεί επέκταση της sortest path, επειδή αυτό που κάνει είναι να δίνει διαφορετικά βάρη στις ακμές της ταξινομίας. Τα κριτήρια σύμφωνα με τα οποία υπολογίζονται τα βάρη είναι τα εξής: *πυκνότητα της ταξινομίας* σε αυτό το σημείο, *το βάθος* που βρίσκεται, καθώς επίσης και η *συσχέτιση μεταξύ των κόμβων πατέρα-παιδιού*.

Hirst and St-Onge[9]: Σε αυτή τη μέθοδο, όσο πιο κοντά είναι οι όροι και όσο λιγότερες είναι οι αλλαγές κατεύθυνσης στο μονοπάτι, τόσο πιο όμοιοι είναι οι όροι μεταξύ τους. Η σημασιολογική ομοιότητα περιγράφεται από την εξίσωση:

weight = C - *path length* - k* *number of changes of direction* ,

όπου C,k: σταθερές.

Li et Al[10]: Σε αυτή τη μέθοδο συνυπολογίζεται το συντομότερο μονοπάτι μεταξύ δύο όρων, L, καθώς επίσης και το βάθος που βρίσκεται ο ειδικότερος πατέρας τους, H. Η σημασιολογική ομοιότητα υπολογίζεται από τη παρακάτω μη γραμμική εξίσωση, η οποία παίρνει τιμές από 0(ανόμοιοι όροι) μέχρι 1(συνώνυμοι όροι):

$$sim(c1, c2) = e^{(-\alpha L)} * \frac{(e^{(\beta H)} - e^{(-\beta H)})}{(e^{(\beta H)} + e^{(-\beta H)})} \qquad (2.2.2)$$

όπου α=0.2 και β=0.6. Τα α και β είναι σταθερές οι οποίες δείχνουν τη συνεισφορά του συντομότερου μονοπατιού και του βάθους αντίστοιχα.

Οι παραπάνω μέθοδοι στηρίζονται μονάχα στην ιεραρχία των όρων υποθέτοντας ότι οι ακμές δηλώνουν απόσταση. Αξίζει να σημειωθεί όμως ότι οι όροι που βρίσκονται πιο κοντά στα φύλλα των δέντρων είναι πολύ περισσότεροι και πιο ειδικοί από αυτούς που βρίσκονται πιο κοντά στη ρίζα. Επομένως δύο όροι που έχουν μια συγκεκριμένη απόσταση και βρίσκονται ψηλά στην ιεραρχία, είναι λιγότερο όμοιοι από δύο άλλους όρους που έχουν την ίδια απόσταση αλλά βρίσκονται πιο κάτω στη ιεραρχία. Αυτό το σημείο δεν το παρατηρούμε στην shortest path μέθοδο, ενώ αντίθετα το παρατηρούμε στην Weighted Link και στη Lietal, όπου έχουμε και καλύτερα αποτελέσματα. Η καλύτερη από τις παραπάνω μεθόδους είναι η Lietal, όπως φαίνεται και στη [10].

2.8.2 Μέθοδοι Περιεχόμενου Πληροφορίας(Information Content Measures)

Σε αυτή την κατηγορία μεθόδων η σημασιολογική ομοιότητα μεταξύ των όρων στηρίζεται στην πληροφορία που περικλείει ένας όρος (Information Content). Η πληροφορία ενός όρου βασίζεται στην συχνότητα εμφάνισής του σε μία μεγάλη συλλογή κειμένων. Όπως θα φανεί αργότερα όσο πιο πολλές φορές εμφανίζεται ένας όρος τόσο λιγότερη πληροφορία περιέχει. Επίσης γνωρίζουμε ότι όσο πιο κοντά σε φύλλο βρίσκεται ένας όρος σε μία ταξινομία τόσο περισσότερη πληροφορία περιέχει. Έτσι λοιπόν για ταχύτητα δεν μετράται η συχνότητα εμφάνισης για να υπολογιστεί το information content, αλλά υπολογίζεται απευθείας από την ταξινομία. Περνώντας τώρα στο βασικό κομμάτι, που δεν είναι άλλο από την ομοιότητα δύο όρων, η ιδέα είναι να βρεθεί ποιος είναι ο *ειδικότερος* κοινός πατέρας των δύο όρων και να υπολογιστεί το information content του. Όσο μεγαλύτερο είναι τόσο περισσότερη σχέση έχουν οι δύο όροι μεταξύ τους.

Μιλώντας ειδικότερα επειδή μέσω της ταξινομίας ουσιαστικά υποθέτουμε την συχνότητα, υπολογίζουμε την πιθανότητα εμφάνισης του όρου στο κείμενο. Η πιθανότητα βρίσκεται από την εξίσωση:

$$p(c)=freq(c)/N \qquad (2.2.3)\ ,$$

όπου N ο συνολικός αριθμός των όρων της ταξινομίας,

και $freq(c)=\Sigma_{n\epsilon words(c)} n$ όπου words(c) οι όροι που βρίσκονται κάτω από τον c.

Η παραπάνω εξίσωση αποδεικνύει ότι όσο πιο πάνω στην ταξινομία βρίσκεται ένας όρος τόσο πιο μεγάλη συχνότητα εμφάνισης έχει. Ακόμα το information content υπολογίζεται από το λογάριθμο $-\ln(p(c))$, ο οποίος δείχνει ότι όσο πιο γενικός είναι ένας όρος τόσο λιγότερη πληροφορία περιέχει.

Μία ακόμα καλύτερη μέθοδος υπολογισμού του information content προτάθηκε από τον Nuno Seco[11] και υπολογίζεται από την εξίσωση:

$$ic(c)=\frac{\log\frac{hypo(c)+1}{max_c}}{\log\frac{1}{max_c}} \qquad (2.2.4),$$

όπου hypo(c) ο αριθμός των υπώνυμων (hyponyms) του c,
maxc ο αριθμός των όρων της ταξινομίας.
Επίσης η εξίσωση έχει σύνολο τιμών το[0,1].

Όλες οι μέθοδοι που ακολουθούν στηρίζονται στο information content, δηλαδή της πιθανότητας εμφάνισής στα κείμενα, του πιο κοντινού πατέρα των δύο όρων. Άρα:

$$p_{mis}(c_{1,}c_2)=min_{c\,\epsilon\,S(c_{1,}c_2)}\{p(c)\} \qquad (2.2.5)$$

Lord et al[12]: Αυτή είναι μία πάρα πολύ απλή μέθοδος υπολογισμού της σημασιολογικής ομοιότητας δύο όρων και στηρίζεται στον κοινό πατέρα των όρων με το μεγαλύτερο information content. Ειδικότερα στηρίζεται στην πιθανότητα εμφάνισής του. Δηλαδή:

$$sim_{Lord}(c_{1,}c_2)=1-p_{mis} \qquad (2.6)$$

Το σύνολο τιμών του είναι από 0(ανόμοιοι όροι) μέχρι 1(συνώνυμοι).

Resnik[13]: Αυτή η μέθοδος χρησιμοποιεί το information content του κοινού πατέρα των δύο όρων, με το μεγαλύτερο information content:

$$sim_{Resnik}(c_{1,}c_2)=-\ln p_{mis} \qquad (2.2.7)$$

Επομένως όσο περισσότερη πληροφορία έχουν κοινή δύο όροι, τόσο πιο όμοιοι είναι. Το σύνολο τιμών του θεωρητικά είναι από 0 μέχρι άπειρο εφόσον το σύνολο τιμών του pmis είναι [0,1]. Πρακτικά αν Ν είναι οι όροι στην ταξινομία το pmis φτάνει μέχρι 1/Ν άρα το μέγιστο στη σημασιολογική μέθοδο είναι το ln(N).

Lin[14]: Αυτή η μέθοδος χρησιμοποιεί το information content και του κοινού τους πατέρα, αλλά και το δικό τους. Η εξίσωση που το περιγράφει είναι η εξής:

$$sim_{Lin}(c_{1,}c_2)=\frac{2\ln \mathrm{p}_{mis}(c_{1,}c_2)}{lnp(c_1)+lnp(c2)} \qquad (2.2.8)$$

Ισχύει ότι $p_{mis}\geq p(c_1)$ και $p_{mis}\geq p(c_2)$ άρα το σύνολο τιμών είναι το [0,1]. Με αυτή τη μέθοδο δεν υπάρχει η αδυναμία του Rasnik, στη σύγκριση ίδιου όρου, αφού το αποτέλεσμα που βγαίνει είναι καθαρό 1.

Jiang et al[15]: Αυτή η μέθοδος σε αντίθεση με τις άλλες μετράει τη σημασιολογική απόσταση των δύο όρων (semantic distance):

$$dist_{Jiang}(c_{1,}c_2) = -2\ln p_{mis}(c_{1,}c_2) - (\ln p_{mis}(c_1) + \ln p_{mis}(c_2)) \quad (2.2.9)$$

Ενώ η σημασιολογική ομοιότητα υπολογίζεται από την εξίσωση:

$$sim_{Jiang}(c_{1,}c_2) = 1 - dist_{Jiang} \qquad (2.2.10)$$

Όπως και στην μέθοδο του Resnik, έτσι και σε αυτή τα αποτελέσματα μπορεί να βγουν πολύ μεγάλα, αλλά στην πράξη φτάνουν μέχρι το 2ln(N), όπου N ο αριθμός των όρων της ταξινομίας. Βλέπουμε ότι όπως και στην Lin, έτσι και εδώ γίνεται χρήση του information content του κοινού πατέρα αλλά και των όρων που θέλουμε την ομοιότητα τους.

2.8.3 Μέθοδοι Βασισμένοι Σε Χαρακτηριστικά (Feature-Based Measures)

Μέχρι τώρα τα σημασιολογικά κριτήρια δεν παίρναν υπόψη τους τα χαρακτηριστικά των όρων, ενώ αυτά περιέχουν πολύ σημαντική πληροφορία. Η παρακάτω μέθοδος χρησιμοποιεί τα χαρακτηριστικά ενώ δε συμπεριλαμβάνει την θέση του όρου στη ταξινομία (edge countin measure) και το information content των όρων.

Tversky[16]: Αυτή η μέθοδος βασίζεται στις χαρακτηριστικές ομάδες λέξεων που περιγράφουν ένα συγκεκριμένο όρο. Όσο πιο πολλά κοινά χαρακτηριστικά έχουν δύο όροι και όσο το δυνατόν λιγότερα μη κοινά τόσο πιο όμοιοι είναι. Η εξίσωση που περιγράφει την μέθοδο είναι η εξής:

$$sim_{Tversky}(c_{1,}c_2)=\frac{|C_1 \cap C_2|}{|C_1 \cap C_2|+k|C_1 \setminus C_2|+(k-1)|C_2 \setminus C_1|} \qquad (2.2.11)$$

όπου C1 και C2 είναι οι χαρακτηριστικές ομάδες όρων των c1 και c2 και κ ϵ [0,1] περιγράφει το βάρος που έχουν τα μη κοινά χαρακτηριστικά των δύο όρων. Η σημασιολογική ομοιότητα της μεθόδου αυτής παίρνει τιμές από 0 μέχρι 1.

2.8.4 Υβριδικές μέθοδοι (Hybrid Measures)

Υβριδικές μέθοδοι, είναι οι μέθοδοι οι οποίες συνδυάζουν κάποια από τα παραπάνω σημασιολογικά κριτήρια. Άρα συνδυάζουν την απόσταση μεταξύ των δύο όρων, το information content και τα χαρακτηριστικά των όρων.

Rodriguez et al[17]: Αυτή η μέθοδος μπορεί να χρησιμοποιηθεί και πάνω σε μία οντολογία (single ontology), αλλά και πάνω σε διαφορετικές (cross ontology similarity). Η σημασιολογική ομοιότητα καθορίζεται από σύγκριση των entity κλάσεων πάνω στις ομάδες συνωνύμων, σημασιολογικών γειτόνων και διαφορετικών χαρακτηριστικών και περιγράφεται από την παρακάτω εξίσωση:

$$S(a^p, b^q) = w_w * S_w(a^p, b^q) + w_u * S_u(a^p, b^q) + w_n * S_n(a^p, b^q) \quad (2.2.12)$$

όπου $w_w, w_u, w_n \geq 0$ και $w_w + w_u + w_n = 1$, τα βάρη των Sw, Su και Sn. Τα Sw, Su και Sn, απεικονίζουν την ομοιότητα των ομάδων συνωνύμων, των ομάδων χαρακτηριστικών και των γειτονικών ομάδων αντίστοιχα των: entity class α της οντολογίας p και entity class b της οντολογίας q. Τα Sw, Su και Sn υπολογίζονται ως εξής:

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + a|A \setminus B| + (a-1)|A \setminus B|} \quad (2.2.13),$$

Όπως μπορεί εύκολα να παρατηρηθεί η εξίσωση αυτή μοιάζει πάρα πολύ με την Tversky. Η μόνη τους διαφορά είναι στον υπολογισμό του α, το οποίο υπολογίζεται ως συνάρτηση του βάθους που βρίσκονται οι δύο όροι.

$$a(c_{1,} c_2) = \begin{cases} d\dfrac{(c_{1,} c_{mis})}{d(c_{1,} c_2)}, d(c_{1,} c_{mis}) \leq d(c_{2,} c_{mis}) \\ 1 - d\dfrac{(c_{1,} c_{mis})}{d(c_{1,} c_2)}, d(c_{1,} c_{mis}) > d(c_{2,} c_{mis}) \end{cases}$$

όπου $d(c_{1,}c_2)=d(c_{1,}c_{mis})+d(c_{2,}c_{mis})$

X-Similarity[18]: Η X-Similarity μέθοδος στηρίζεται πάνω στην Rodriguez, επειδή είναι η καλύτερη μέθοδος θεωρητικά και πρακτικά και μπορεί να εφαρμοστεί πάνω σε διαφορετικές οντολογίες. Η εφαρμογή της έγινε πάνω στο MeSH και στο Wordnet. Αναλυτικά οι αλλαγές που έγιναν σε σχέση με την Rodriguez είναι οι εξής:

- Αντικατάσταση του Su με το Sgloss. Δηλαδή αντικαθιστάμε το Feature Matching με το Glossary Matching. Στο glossary matching συγκρίνονται με ένα απλό VSM οι ορισμοί των δύο όρων. Η σύγκριση γίνεται με την λογική της παρακάτω εξίσωσης:

$$S_{gloss}(a^p,b^q)=\frac{|A\cap B|}{|A\cup B|} \qquad (2.2.14)$$

  όπου Α και Β περιέχουν τους όρους των ορισμών του a και b.

- Αντικατάσταση του Sn με το Snm. Στην κλασική περίπτωση του semantic neighborhood matching ομαδοποιούνται οι όροι σε ακτίνα ρ από τον a και τον b αντίστοιχα και συγκρίνονται μεταξύ τους με την εξίσωση (2.2.13). Στην περίπτωση του X-Similarity γίνεται το ακόλουθο:

$$S_{nm}(a^p,b^q)=max_i\frac{|A_i\cap B_i|}{|A_i\cup B_i|} \qquad (2.2.15)$$

  όπου i δηλώνει μία σχέση (relation) μεταξύ των όρων a και b

  και Α και Β είναι οι ορισμοί (definitions) των όρων που συνδέουν τους όρους a και b αντίστοιχα με την σχέση i.

Συνδυάζοντας τα παραπάνω η εξίσωση του X-Similarity έχει ως εξής:

$$S(a^p,b^q)=w_w*S_w(a^p,b^q)+w_{gloss}*S_{gloss}(a^p,b^q)+w_{nm}*S_{nm}(a^p,b^q)$$

(2.2.16)

Την τελευταία μέθοδο αξίζει να την συγκρατήσουμε, γιατί είναι αυτή στην οποία στηριχθήκαμε για να προτείνουμε την δική μας.

2.9 *Κατασκευή Οντολογιών Από Κείμενα*

Η αναπαράσταση των εννοιών (concepts) σε μια ταξινομία ή οντολογία, από μία συγκεκριμένη γνωστική περιοχή είναι πάρα πολύ σημαντική και πολύ χρήσιμη. Για αυτό το λόγο βρίσκεται στο ερευνητικό επίκεντρο, όπου έχουν γίνει σημαντικές προσπάθειες αυτοματοποίησης της κατασκευής οντολογιών. Οι μέθοδοι που έχουν προταθεί είναι αυτόματες ή ημιαυτόματες, αναλόγως αν έχει χρησιμοποιηθεί και άλλη πηγή πληροφορίας εκτός από μία δοσμένη συλλογή κειμένων. Συνήθως οι μέθοδοι που υπάρχουν στη βιβλιογραφία χωρίζονται σε δύο στάδια:

(1) Επεξεργασία κειμένων η οποία έχει ως αποτέλεσμα την εξαγωγή των σημαντικότερων όρων.

(2) Δημιουργία ταξινομίας μέσω ομαδοποίησης (clustering) ή κάποιας άλλης διαδικασίας, όπως η χρήση ανθρώπινου παράγοντα (expert).

Παρακάτω φαίνονται δουλειές σχετικές με την κατασκευή οντολογιών από κείμενα.

2.9.1 Automatic Ontology Construction From The Literature

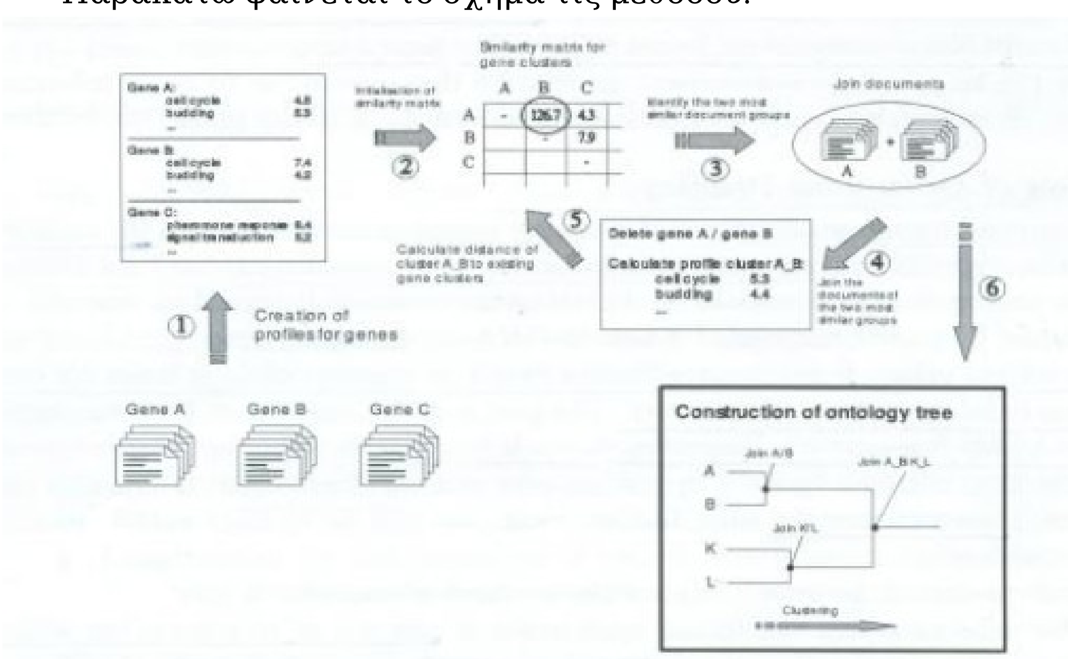
Η μέθοδος αυτή είναι αυτόματη και ομαδοποιεί (classification) γονίδια (genes), με βοήθεια από την βιβλιογραφία[27]. Η συλλογή κειμένων που χρησιμοποιήθηκε είναι το MEDLINE. Το αποτέλεσμα είναι μία ελαφριά οντολογία (light-weighted ontology), εφόσον επιστρέφεται μια ιεραρχία των όρων. Τα βήματα της συγκεκριμένης μεθόδου έχουν ως εξής:

(1) Αρχικά ομαδοποιούνται τα κείμενα που αναφέρονται σε γονίδια, ανάλογα με το γονίδιο στο οποίο αναφέρονται. Από την κάθε ομάδα κειμένων (γονιδίων) εξάγονται οι πιο σημαντικοί όροι, οι οποίοι είναι το σήμα κατατεθέν της κάθε ομάδας.

(2) Σε αυτό το βήμα ομαδοποιούνται οι ομάδες γονιδίων που είναι πιο όμοιες. Δηλαδή έχουν περισσότερους κοινούς όρους.

(3) Τέλος επαναλαμβάνονται τα βήματα 1 και 2 μέχρι να μην υπάρχουν κοινοί σημαντικοί όροι μεταξύ των clusters.

Παρακάτω φαίνεται το σχήμα τις μεθόδου.



2.9.2 ASIUM

Το ASIUM[28] (Acquisition of Semantic knowledge Using Machine learning methods), είναι μια ημιαυτόματη μέθοδος, αφού ύστερα από κάθε επίπεδο ομαδοποίησης χρειάζεται κάποιος να βάλει μία ετικέτα στη νέα ομάδα.

Το ASIUM κάνει χρήση του συντακτικού αναλυτή SYLEX, ο οποίος επιστρέφει αναλυμένες τις προτάσεις σε sub-categorization frames. Το sub-categorization frame έχει την παρακάτω μορφή:

<verb><syntactic role | preposition : headword>*

Σε περίπτωση που μία πρόταση είναι διφορούμενη ο SYLEX επιστρέφει και τις δύο εκδοχές.

Η λογική που λειτουργεί το ASIUM είναι η εξής. Κυρίως όροι (Headwords) που εμφανίζονται στα frames με το ίδιο ρήμα (verb) και την ίδια πρόθεση (preposotion) ή συντακτικό ρόλο (syntacti role), ανήκουν στο ίδιο synthetic frame και δηλώνουν την ίδια έννοια (concept). Για παράδειγμα τα frames :

<to travel><subject: father> <by : car>

<to travel><subject: neighbor> <by : train>

<to drive><subject: friend> <object : car>

<to drive><subject: collegue> <object : motorbike>

<to drive><subject: friend> <by : motorbike>

Ανήκουν στα παρακάτω synthetic frames:

<to travel><subject: [father(1),neighbor(1)]> <by : [car(1),train(1)]>

<to drive><subject: [friend(2),collegue(1)]> <by : [car(1),motorbike(2)]>

Ύστερα χτίζεται η οντολογία από κάτω προς τα πάνω, σύμφωνα με την Conceptual Clustering μέθοδο. Δύο ομάδες ενώνονται όταν βρίσκονται πολύ κοντά μεταξύ τους. Για παράδειγμα δύο ομάδες που έχουν τους ίδιους κυρίως όρους με την ίδια συχνότητα έχουν απόσταση 0. Τέλος μετά από κάθε ένωση απαιτείται η μετονομασία του από κάποιον.

2.9.3 Clustering Concept Hierarchy From Text

Η μέθοδος που εξετάζεται βασίζεται στη υπόθεση ότι ουσιαστικά (nouns) που εμφανίζονται με ίδια συμφραζόμενα (context) είναι όμοια. Η Clustering Concept Hierarchy From Text[29], είναι αυτόματη μέθοδος που επιστρέφει οντολογία (heavy weighted ontology), και την ιεραρχία της την χτίζει μέσω της FCA (Formal Concept Analysis). Η μέθοδος χωρίζεται σε δύο βήματα.

Στο πρώτο βήμα γίνεται συντακτική ανάλυση μέσω του αναλυτή LoPar (trainable and statistical left corner parser). Τα αποτελέσματα που επιστρέφει τα εξής ζευγάρια: verb/object, verb/subject, verb/PP-complement, όπου τα nouns είναι heads και τα verbs είναι atributes. Τέλος από αυτά τα αποτελέσματα κρατάμε μόνο αυτά που βρίσκονται πάνω από ένα συγκεκριμένο κατώφλι t.

Στο δεύτερο και τελευταίο βήμα εισάγουμε τα αποτελέσματα στην FCA[30], ο οποίος δημιουργεί μόνος του μία ιεραρχία, η οποία μπορεί να συνδέει έναν πατέρα με παραπάνω από δυο παιδιά.

Η FCA έχει συγκριθεί με άλλες μεθόδους ομαδοποίησης όπως ο Agglomerative και ο Bi-Section-Kmeans, από τους οποίους υπερτερεί.

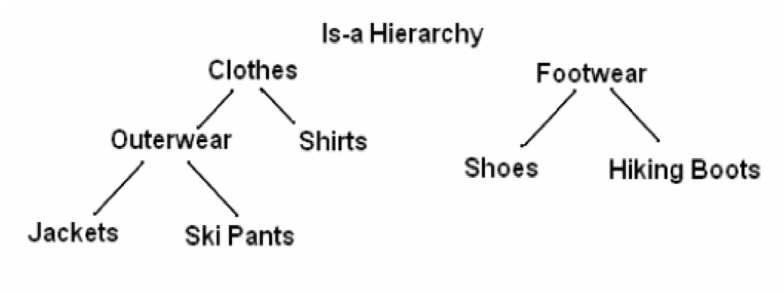
2.9.4 Text-To-Onto

Η Text-to-Onto[31] είναι μία ημιαυτόματη μέθοδος, η οποία πέρα από τις σχέσεις ιεραρχίας, δημιουργεί και άλλες σχέσεις μεταξύ των οντοτήτων. Η μέθοδος χωρίζεται σε τρία μέρη: στην επεξεργασία δεδομένων και εξαγωγή όρων, στην ιεράρχηση των όρων και τέλος στην εξαγωγή σχέσεων με τη βοήθεια των General Association Rules.

Για το πρώτο βήμα, δηλαδή την ανάλυση φυσικής γλώσσας χρησιμοποιείται ο επεξεργαστής κειμένων SMES. O SMES περιέχει έναν tokenizer βασισμένο σε κανονικές εκφράσεις και ένα λεκτικό αναλυτή που χρησιμοποιεί λεξικά.

Στο δεύτερο βήμα χρησιμοποιείται ένας αλγόριθμος ομαδοποίσης.

Στο τρίτο και τελευταίο βήμα έχοντας τα ομάδες αντικειμένων (transactions) που έχουν εξαχθεί στο πρώτο βήμα, την ιεραρχία που έχει εξαχθεί στο δεύτερο και βασιζόμενοι στα General Association Rules, δημιουργούνται οι νέες σχέσεις μας. Παρακάτω φαίνεται σχηματικά πως βγαίνουν οι σχέσεις:



Database D	
Transaction	Items Bought
1	Shirt
2	Jacket, Hiking Boots
3	Ski Pants, Hiking Boots
4	Shoes
5	Shoes
6	Jacket

Rules		
Rule	Support	Conf.
Outerwear=>Hiking Boots	33%	66.6%
Outerwear => Footwear	33%	66.6%
Hiking Boots=> Outerwear	33%	100%
Hiking Boots=>Clothes	33%	100%

2.9.5 Building Large Scale Ontology Networks

Η μέθοδος[32] αυτή διαφοροποιείται από τις προηγούμενες, από τη στιγμή που δεν δημιουργεί την οντολογία από την αρχή αλλά χρησιμοποιεί το DMOZ (Open directory project), το Wordnet και μία συλλογή κειμένων συγκεκριμένου γνωστικού πεδίου. Είναι ημιαυτόματη μέθοδος αφού χρειάζεται ένας ειδικός να την παρακολουθεί, και χωρίζεται σε δύο φάσεις.

Στην πρώτη φάση παίρνουμε ένα κλαδί του DMOZ, το οποίο ανήκει στο γνωστικό πεδίο της συλλογής κειμένων που διαλέξαμε και το ενισχύουμε με synsets από το Wordnet.

Στην δεύτερη φάση παίρνουμε την συλλογή κειμένων που έχουμε διαλέξει και εξάγουμε τους σημαντικότερους όρους. Ύστερα συγκρίνουμε τους όρους που πήραμε με τους όρους του κλαδιού του DMOZ ώστε να δούμε ποιοι λείπουν. Εδώ είναι που έρχεται η σειρά των experts, οι οποίοι κοιτάνε ποιοι όροι έχουν νόημα ώστε να απορριφθούν ή όχι. Αν επιλεγούν τότε κοιτάμε αν υπάρχουν στο Wordnet και ακολούθως αν υπάρχει σε κάποιο άλλο σημείο του DMOZ. Αυτή η διαδικασία συνεχίζεται μέχρι να ολοκληρωθεί η οντολογία και να εγκριθεί από τους experts.

Όπως μπορεί να παρατηρήσει κανείς κάθε μία μέθοδος έχει τα δυνατά της και τα αδύνατά της σημεία. Αυτός είναι ο λόγος για τον οποίο δεν μπορεί να επιλεγεί κάποια μέθοδο ως καλύτερη. Αυτό που μπορεί να γίνει, είναι ύστερα από σοβαρή μελέτη να επιλεγεί η καλύτερη μέθοδος για την περίσταση. Επίσης είναι θεμιτός και οποιοσδήποτε συνδυασμός των παραπάνω αν έχει διαπιστωθεί ότι θα δώσει καλύτερα αποτελέσματα.

3.Περιγραφή Του Προβλήματος

Κάνοντας χρήση της θεωρίας που αναφέρθηκε παραπάνω, θα περιγράψουμε και θα αναλύσουμε τα βήματα της εργασίας μας καθώς και τα αποτελέσματά της. Το πρόβλημα που μελετάται είναι μεθοδολογίες ανάκτησης πληροφορίας με χρήση οντολογιών σε συλλογές ιατρικών κειμένων. Πιο συγκεκριμένα θέλουμε να ερευνήσουμε τις συνέπειες που θα προκύψουν αν προσαρμόσουμε μια οντολογία σε μία συγκεκριμένη συλλογή κειμένων. Θέλουμε δηλαδή να δούμε τι πρόκειται να γίνει αν μικρύνουμε μία ταξινομία όπως το MeSH, αφαιρώντας τους όρους που δεν εμφανίζονται στη συλλογή.

Παλαιότερα είχαν γίνει διάφορα πειράματα Ανάκτησης Πληροφορίας κάνοντας χρήση έτοιμων οντολογιών χωρίς να γίνει κάποια προσαρμογή ως προς την ιεραρχία των όρων. Έτσι λοιπόν σε αυτήν την εργασία καλούμαστε να ερευνήσουμε αν επηρεάζονται τα αποτελέσματα που θα δώσει το σύστημα Ανάκτησης Πληροφορίας εάν προσαρμόσουμε την οντολογία μας στη συλλογή κειμένων. Με άλλα λόγια θέλουμε να δούμε εάν είχαμε τα βέλτιστα αποτελέσματα Ανάκτησης, χωρίς να κάνουμε χρήση συγκεκριμένων οντολογιών. Αξίζει όμως να σημειωθεί ότι στην εργασία αυτή χρησιμοποιήσαμε μια συγκεκριμένη μεθοδολογία, με συγκεκριμένα εργαλεία για την κατασκευή της οντολογίας, που σημαίνει ότι μία διαφορετική προσέγγιση θα μπορούσε να οδηγήσει σε διαφορετικά αποτελέσματα.

Σε αυτό το σημείο αξίζει να σημειωθεί πως σχετίζεται η οντολογία με την Ανάκτηση Πληροφορίας. Οι οντολογίες όπως είπαμε στο κεφάλαιο 2.8, έχουν σχέση με την σημασιολογική ομοιότητα, μεταξύ όρων. Ύστερα η σημασιολογική ομοιότητα των όρων χρησιμοποιείται από μία μέθοδο Ανάκτησης Πληροφορίας, όπως είναι της Rada Mihalcea[33], η οποία πλέον δεν κοιτάει μόνο λεξιλογική ομοιότητα μεταξύ των όρων (VSM), αλλά και σημασιολογική. Δηλαδή το σύστημα δεν ψάχνει να βρει κείμενα τα οποία να περιέχουν μόνο τους όρους του ερωτήματος, αλλά και συνώνυμους και όρους που μοιάζουν σε κάποιο βαθμό.

Επομένως στο κείμενο που θα ακολουθήσει θα περιγράψουμε με ποια λογική κατασκευάσαμε την οντολογία μας, και πάνω σε ποια σημασιολογικά κριτήρια στηριχθήκαμε έτσι ώστε να πάρουμε την σημασιολογική ομοιότητα μεταξύ των όρων της οντολογίας.

3.1 Κατασκευή Οντολογίας

Η οντολογία μας είναι ένας συνδυασμός του MeSH thesaurus με σχέσεις από το Metathesaurus. Μία ιεραρχία όρων χαρακτηρίζεται ως οντολογία όταν μεταξύ τους οι όροι συνδέονται και με άλλες σχέσεις πέρα από τις IS-A. Οντολογία θα μπορούσαμε να πούμε ότι είναι η τέλεια αναπαράσταση ενός γενικού ή ενός συγκεκριμένου γνωστικού πεδίου. Αυτό όπως είναι κατανοητό είναι εξαιρετικά δύσκολο να γίνει αυτόματα και πάρα πολύ χρονοβόρο για να γίνει με το χέρι. Για να επιστρέψουμε στο θέμα μας, θεωρούμε ότι προσθέτοντας στο MeSH σχέσεις από το Metathesaurus, προσεγγίζουμε καλύτερα τον ορισμό της οντολογίας.

Αξίζει λοιπόν να αναφερθούμε επιγραμματικά ξανά στο MeSH και στο Metathesaurus για να γίνουν κατανοητά τα βήματα της εργασίας μας. Το MeSH thesaurus αποτελείται λοιπόν από 16 δέντρα με ιατρικούς όρους (και τους ορισμούς τους) οι οποίοι συνδέονται μεταξύ τους με σχέσεις IS-A. Επίσης αναφέρουμε ότι ο ίδιος όρος μπορεί να βρίσκεται σε διαφορετικά σημεία στην ιεραρχία του MeSH(senses) , πράγμα το οποίο συμβαίνει για τον λόγο ότι ο όρος μπορεί να έχει διαφορετική σημασία αναλόγως σε ποια γνωστική περιοχή και σε ποια συμφραζόμενα χρησιμοποιείται. Επίσης το MeSH μπορεί να θεωρηθεί ιδιαίτερα ακριβές και έγκυρο, αφού είναι αποτέλεσμα εργασίας επιστημόνων του ιατρικού κλάδου.

Το Metathesaurus όπως είπαμε στο κεφάλαιο 2.3 αποτελείται από πολλούς όρους οι οποίοι ενώνονται μεταξύ τους με σχέσεις οι οποίες έχουν βρεθεί από διάφορες λεξιλογικές πηγές και από στατιστικά στοιχεία από συγκεκριμένες συλλογές κειμένων. Διαλέξαμε να εμπλουτίσουμε το MeSH με σχέσεις από το Metathesaurus επειδή περιέχει σχέσεις μεταξύ συγκεκριμένων όρων και όχι μεταξύ κατηγοριών όπως συμβαίνει στην περίπτωση του Semantic Network. Οι σχέσεις που βρέθηκαν για το Metathesaurus μέσα στην βάση του UMLS είναι οι εξής:

AQ: *Αν ένας όρος εμφανιστεί με κάποιον άλλον του αλλάζει τη σημασία.*

QB: *Το αντίστροφο της AQ.*

RB: *Ένας όρος με κάποιον άλλον έχουν μακρινή σημασιολογικά σχέση.*

RN: *Ένας όρος με κάποιον άλλον έχουν κοντινή σημασιολογικά σχέση.*

RO: *Ένας όρος με κάποιον άλλον έχουν σχέση η οποία δεν είναι συνώνυμη, κοντινή ή μακρινή.*

Επομένως η ομοιότητα των όρων πλέον δεν εξαρτάται μόνο από την ιεραρχία της οντολογίας, εφόσον υπάρχουν σχέσεις που συνδέουν όρους, οι οποίοι μπορεί να βρίσκονται μακρυά στην ταξινομία του MeSH.

Τώρα μένει να περιγράψουμε τα βήματα που ακολουθήσαμε για να κατασκευάσουμε την οντολογία μας:

(1) Από το Metathesaurus κρατάμε μόνο τους όρους που ανήκουν και στο MeSH.

(2) Βρίσκουμε τις σχέσεις που συνδέουν τους όρους μεταξύ τους και τους ενώνουμε.

(3) Κατασκευή οντολογίας.

Ακολουθεί ανάλυση των βημάτων.

3.1.1 Διατήρηση MeSH όρων

Τους MeSH όρους τους έχουμε σε μορφή XML. Σε κάθε αρχείο XML περιέχονται : ο ορισμός του εκάστοτε όρου, τα συνώνυμά του, τα υπερώνυμα (hypernyms), τα υπώνυμα (hyponyms), καθώς επίσης και οι διαφορετικές του έννοιες (senses). Με άλλα λόγια στα XML αρχεία περιγράφεται σε ποια θέση στην ιεραρχία του MeSH βρίσκεται κάθε όρος, καθώς επίσης και ο ορισμός του. Παράδειγμα:


```
- <word>
    1,2-Dimethylhydrazine
    <key>1,2-Dimethylhydrazine</key>
  - <pos name="noun">
    - <sense number="1">
      - <synset number="9">
          <item>1,2-Dimethylhydrazine</item>
          <item>1,2 Dimethylhydrazine</item>
          <item>N,N'-Dimethylhydrazine</item>
          <item>N,N' Dimethylhydrazine</item>
          <item>1,2-Dimethyl-hydrazine</item>
          <item>1,2 Dimethyl hydrazine</item>
          <item>sym-Dimethylhydrazine</item>
          <item>sym Dimethylhydrazine</item>
          <item>SDMH</item>
        - <defn>
            A DNA alkylating agent that has been shown to be a potent carcinogen and is widely used to induce colon tumors in experimental
            animals.
          </defn>
        </synset>
      - <links>
        - <hypernym>
          - <synset number="1">
              <item>Dimethylhydrazines</item>
            - <defn>
                Hydrazines substituted with two methyl groups in any position.
              </defn>
            </synset>
          - <hypernym>
            - <synset number="1">
                <item>Methylhydrazines</item>
              - <defn>
                  Hydrazines substituted by one or more methyl groups in any position.
                </defn>
              </synset>
```

```
          </synset>
        - <hypernym>
          - <synset number="1">
              <item>Hydrazines</item>
              <defn>null</defn>
            </synset>
          - <hypernym>
            - <synset number="2">
                <item>Organic Chemicals</item>
                <item>Chemicals, Organic</item>
              - <defn>
                  A broad class of substances containing carbon and its derivatives. Many of these chemicals will frequently contain hydrogen
                  with or without oxygen, nitrogen, sulfur, phosphorus, and other elements. They exist in either carbon chain or carbon ring for
                </defn>
              </synset>
            - <hypernym>
              - <synset number="1">
                  <item>chemicals and drugs</item>
                  <defn>medical higher concept</defn>
                </synset>
              </hypernym>
            </hypernym>
          </hypernym>
        </hypernym>
      </hypernym>
    </links>
  </sense>
 </pos>
</word>
```


σχήμα 3.1

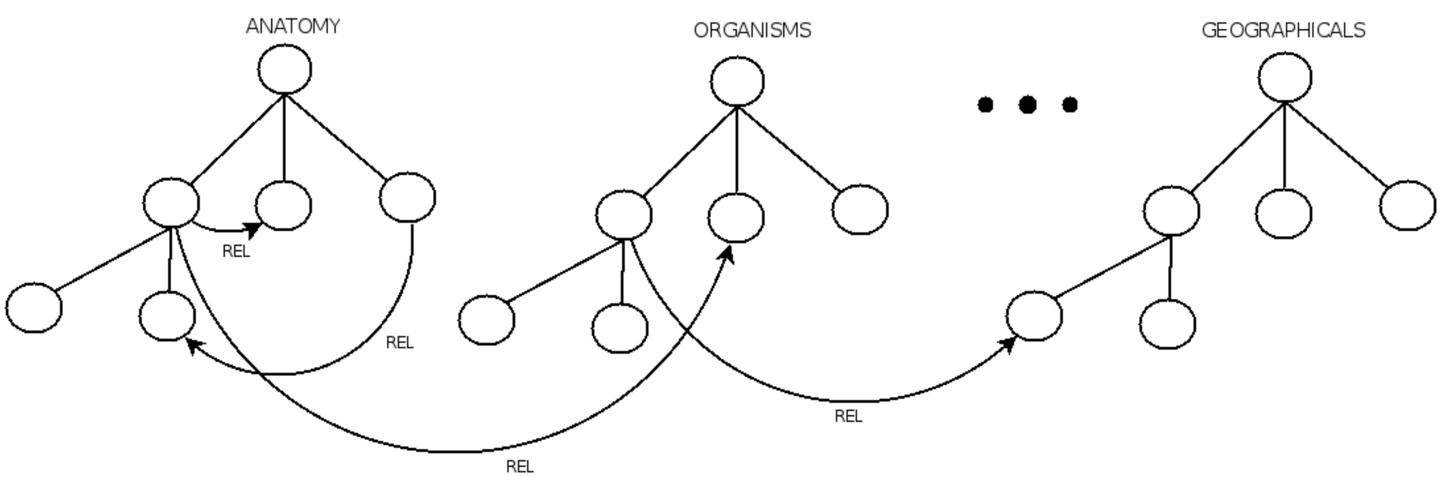
Έτσι για να μπορέσουμε να προχωρήσουμε γρηγορότερα αποθηκεύσαμε όλες τις επικεφαλίδες του MeSH, με τους ορισμούς τους σε ένα hash table, με κλειδί (key) την επικεφαλίδα και με τιμή (value) τον ορισμό (definition). Οι εισχωρήσεις είναι 24.767 αφού τόσες είναι οι επικεφαλίδες. Οι υπόλοιποι όροι που δεν φαίνονται και είναι πολλοί ως προς το πλήθος ανήκουν στα entry terms, τα οποία είναι συνήθως συνώνυμα και υπάγονται στις επικεφαλίδες.

Άρα μέχρι στιγμής έχουμε τους όρους, τα ορισμούς τους και την ιεραρχία των όρων από το MeSH.

3.1.2 Σύνδεση όρων μέσω των σχέσεων του Metathesaurus

Σε αυτό το κομμάτι της εργασίας μας βρίσκουμε με ποιες σχέσεις συνδέονται οι όροι μεταξύ τους. Η λογική αυτού του βήματος είναι απλή. Εφόσον ξέρουμε σε ποια κατηγορία ανήκει κάθε MeSH όρος και τις σχέσεις που συνδέουν τις κατηγορίες μεταξύ τους είναι πολύ εύκολο να συσχετίσουμε κάθε όρο με κάποιον άλλον. Επομένως δημιουργήσαμε ένα hashtable, το οποίο έχει σαν κλειδί έναν όρο του MeSH με μία σχέση και σαν τιμή όλους τους όρους που συνδέονται με αυτά τα δύο. Οι σχέσεις που συνδέουν τους όρους φαίνονται στο κεφάλαιο 3.1(σελίδα 46).

Η οντολογία μας λοιπόν θα έχει την παρακάτω μορφή. Δηλαδή θα έχει τα 16 δέντρα του MeSH και θα ενώνονται με τις παραπάνω σχέσεις:



Σε αυτό το βήμα έχουμε ολοκληρώσει την κατασκευή της οντολογίας μας. Για χάρη ταχύτητας μεταφέρουμε την ιεραρχία του MeSH που βρίσκεται σε XML αρχεία σε hashtable, με κλειδί έναν όρο και με τιμή τα συνώνυμά του, τα υπερώνυμά του και τα υπώνυμά του. Όπως είχαμε πει όμως πρωτύτερα, ένας όρος μπορεί να βρίσκεται σε διαφορετικά σημεία στο MeSH, δηλαδή να έχει διαφορετικές έννοιες (senses). Εμείς λοιπόν κρατάμε μόνο την έννοια , η οποία παρουσιάζεται πιο συχνά.

3.2 Κατασκευή Οντολογίας Από Συλλογή Κειμένων

Όπως είδαμε στο κεφάλαιο 2.9 η κατασκευή οντολογίας από μια συγκεκριμένη συλλογή κειμένων χωρίζεται σε δύο στάδια:

(1) Επεξεργασία κειμένων η οποία έχει ως αποτέλεσμα την εξαγωγή των σημαντικότερων όρων.
(2) Δημιουργία ταξινομίας μέσω ομαδοποίησης ή κάποιας άλλης διαδικασίας, όπως η χρήση ανθρώπινου παράγοντα (expert).

Στο πρώτο βήμα καλούμαστε να εξάγουμε τους σημαντικότερους όρους της συλλογής κειμένων. Αυτό μπορεί να γίνει με κάποια μέθοδο εξαγωγής όρων, όπως εξηγείται στο κεφάλαιο 2.4. Παραδείγματα τέτοιων μεθόδων είναι το C/NC Value και το AMTEx. Το C/NC Value δεν έχει σημασία πάνω σε τι είδους συλλογή κειμένων, δηλαδή γνωστική περιοχή, την εφαρμόζουμε. Άρα οι όροι που εξάγονται είναι κάθε γνωστικού αντικειμένου, από τη στιγμή που θα θεωρηθούν σημαντικές. Αντίθετα η AMTEx[26] είναι για το συγκεκριμένο γνωστικό πεδίο της ιατρικής και όπως αναφέραμε στο κεφάλαιο 2.5 δίνει καλύτερα αποτελέσματα σε ιατρικές συλλογές κειμένων. Επιγραμματικά να θυμίσουμε ότι αρχικά η AMTEx εξάγει όρους με το C/NC Value[25], και όποιος από αυτούς τους όρους δεν βρίσκεται στο MeSH απορρίπτεται. Ύστερα με κάποιο κατώφλι εάν θέλουμε κάνουμε επέκταση (expansion) και παίρνουμε όρους του MeSH, οι οποίοι βρίσκονται γύρω από ήδη εξαγόμενους όρους.

Εμείς χρησιμοποιήσαμε τη συλλογή κειμένων του Ohsumed, το οποίο όπως αναφέραμε στο κεφάλαιο 2.7, αποτελείται από 348.566 περολήψεις κειμένων, που πηγάζουν από 270 ιατρικά περιοδικά. Επομένως έχουμε ένα σύνολο κειμένων ιατρικού περιεχομένου, το οποίο μας δίνει τη δυνατότητα να χρησιμοποιήσουμε την μέθοδο AMTEx για να εξάγουμε τους κυριότερους όρους των κειμένων. Οι MeSH όροι που εξάγαμε με την μέθοδο εξαγωγής όρων της AMTEx είναι 4.329. Αξίζει να σημειώσουμε ότι δεν κάναμε επέκταση στους όρους που πήραμε με την AMTEx. Φτάνοντας λοιπόν στο τέλος του πρώτου βήματος, πήραμε από το Ohsumed τους MeSH όρους που το περιγράφουν όσο το δυνατόν καλύτερα.

Στο δεύτερο βήμα πρέπει να φτιάξουμε την ιεραρχία των όρων που εξάγαμε. Για αυτόν το σκοπό εμείς θα βοηθηθούμε από το MeSH και τις σχέσεις του Metathesaurus. Σε αυτό το σημείο λοιπόν θα χρησιμοποιήσουμε την οντολογία που περιγράψαμε στο κεφάλαιο 3.1, δηλαδή της ταξινομίας του MeSH που εμπλουτίσαμε με σχέσεις από το Metathesaurus.

Ο τρόπος που δημιουργήσαμε την οντολογία μας είναι ο παρακάτω:

(1) Κρατάμε από το MeSH όλες τις ρίζες των 16 δέντρων του, είτε υπάρχουν είτε δεν υπάρχουν στο Ohsumed, δηλαδή κρατάμε τους εξής όρους: 'Anatomy', 'Organisms', 'Diseases', 'Chemical and Drugs', 'Analytical, Diagnostic and Therapeutic Techniques and Equipment', 'Psychiatry and Psychology', 'Biological Sciences', 'Anthropology, Education, Sociology and Social Phenomena', ''Technology, Industry, Agriculture', 'Humanities', 'Information Science', 'Named Groups', 'Health Care', 'Publication Characteristics', 'Geographicals'.

(2) Ύστερα κοιτάμε έναν έναν τους κόμβους της οντολογίας όλου του MeSH. Εάν ο κόμβος περιέχει όρο που έχει εξαχθεί από το Ohsumed τον κρατάμε, διαφορετικά τον αφαιρούμε.

Επομένως η νέα μας οντολογία θα περιέχει 4345 κόμβους (όροι AMTEx + ρίζες).

3.3 Μέθοδος σημασιολογικής ομοιότητας για την Οντολογία MeSH-Metathesaurus.

Γίνεται εύκολα κατανοητό ότι πρέπει να βρούμε ένα τρόπο να ελέγξουμε την ποιότητα της οντολογίας μας. Επομένως αυτό που κάναμε ήταν να φτιάξουμε μία μέθοδο σημασιολογικής ομοιότητας, την X-SimilarityMetathesaurus.

Όπως γνωρίζουμε η περισσότερη πληροφορία σε κείμενα και ερωτήματα υπάρχει μέσα στα ουσιαστικά. Επομένως οι σημαντικότεροι όροι ουσιαστικών εξάγονται από κείμενα και κατ' επέκταση συγκρίνονται μεταξύ τους για να βρούμε την ομοιότητα κειμένου με κείμενο ή ερώτημα. Το MeSH όπως γνωρίζουμε είναι μία ταξινομία από όρους ουσιαστικών, άρα μπορεί κάλλιστα να χρησιμοποιηθεί για σημασιολογική ομοιότητα. Επομένως μπορεί να χρησιμοποιηθεί και για συστήματα Ανάκτησης Πληροφορίας, τα οποία εκτός από την λεξιλογική ομοιότητα εντοπίζουν και την σημασιολογική μεταξύ των όρων.

Η X-SimilairtyMetathesaurus είναι μια μέθοδος που στηρίζεται πολύ στην X-Similarity[18]. Είναι δηλαδή μία υβριδική cross-ontology μέθοδος. Η κύρια διαφορά τους είναι ότι η X-Similarity δημιουργήθηκε για οντολογίες με σχέσεις IS-A και του Wordnet, ενώ η X-SimilarityMetathesaurus δημιουργήθηκε για οντολογίες με σχέσεις IS-A και οντολογίες με σχέσεις του Metathesaurus. Ο τύπος που την περιγράφει είναι ο ίδιος με του X-Similarity, δηλαδή:

$$S(a^p,b^q)=w_w*S_w(a^p,b^q)+w_{gloss}*S_{gloss}(a^p,b^q)+w_{nm}*S_{nm}(a^p,b^q)$$

Τα w όπως είπαμε είναι τα βάρη κάθε κριτηρίου, και ισχύει ότι Ww=Wgloss=Wnm.

Το Sw είναι η ομοιότητα μεταξύ των ομάδων συνωνύμων των a και b. Είναι ευνόητο ότι εάν κοιτάμε δύο όρους που βρίσκονται στην ίδια οντολογία δεν θα υπάρξει ομοιότητα. Όταν αναφερόμαστε στην ίδια οντολογία δεν υπολογίζουμε το Sw.

Το Sgloss είναι ομοιότητα μεταξύ των ορισμων των δύο όρων, η οποία βρίσκεται με ένα απλό Vector Space Model (τύπος 2.2.14).

Το Snm είναι η ομοιότητα των γειτονικών όρων (semantic neighborhood matching) η οποία βρίσκεται με τον τύπο 2.2.15.

$$S_{nm}(a^p,b^q)=max_i\frac{|A_i\cap B_i|}{|A_i\cup B_i|}$$

Σε αυτό το σημείο επομένως παίρνουμε τις 5 σχέσεις από το Metathesaurus(AQ, QB, RB, RN, RO) και τις δύο σχέσεις του MeSH(hyponyms, hypernyms). Για τις σχέσεις του Metathesaurus χρησιμοποιούμε VSM ενώ για τα τα hypernyms και hyponyms την μέθοδο Tverky. Ακολουθεί παράδειγμα:

Έστω δύο όροι a και b και θέλουμε να υπολογίσουμε την ομοιότητα τους ως προς τη σχέση AQ. Αρχικά βρίσκουμε με ποιους όρους συνδέεται ο a με τη σχέση AQ(π.χ. k,l,m) και μετά κάνουμε το ίδιο και για τον b(π.χ. r,s,t). Ύστερα ενώνουμε τους ορισμούς της κάθε ομάδας όρων(δηλαδή def(k,l,m) και def(r,s,t)) και μετά τα συγκρίνουμε με το VSM(vsm[def(k,l,m) def(r,s,t)]). Αυτό όπως καταλαβαίνετε το κάνουμε για όλες τις σχέσεις και κρατάμε το αποτέλεσμα με το μεγαλύτερο βαθμό ομοιότητας.

Όταν αναφερόμαστε σε σύγκριση όρων μίας οντολογίας, και ιδιαιτέρως της ταξινομίας του MeSH, και όχι διαφορετικών αφαιρούμε το Sw. Η διαδικασία αυτή ακολουθείται και στην X-Similarity. Αυτό το κάνουμε γιατί στο MeSH όπως έχουμε πει όλα τα συνώνυμα βρίσκονται σαν entry terms στη συγκεκριμένη επικεφαλίδα στην οποία ανήκουν. Με άλλα λόγια δεν υπάρχει περίπτωση δύο επικεφαλίδες του MeSH να έχουν κοινά συνώνυμα. Αυτό θα είχε ως αποτέλεσμα να δίνει διαρκώς Sw=0, το οποίο δεν θα έδειχνε την πραγματική ομοιότητα των όρων. Αυτό που κάνουμε λοιπόν είναι να ξεχωρίσουμε τη σχέση που αναφέρεται στα υπερώνυμα (hypernyms) και να αντικαταστήσουμε το Sw. Ο διαχωρισμός αυτός γίνεται προφανώς επειδή τα υπερόνυμα στην οντολογία περιέχουν πολύ σημαντική πληροφορία, αφού εισάγουν στοιχεία της ταξινομίας. Μπορεί να κατανοήσει κανείς ότι δύο όροι που έχουν κοινούς πατέρες περιέχουν και κοινή πληροφορία. Δηλαδή η νέα σχέση είναι η εξής:

$$S(a,b)=w_{gloss}*S_{gloss}(a,b)+w_{nm}*S_{nm}(a,b)+w_{hyp\,nm}*S_{hyp\,nm}(a,b) \quad .$$

Το επόμενο στάδιο λοιπόν είναι να ελέγξουμε την αποτελεσματικότητα της μεθόδου μας πάνω στη νέα οντολογία, συγκρίνοντάς την με άλλες μεθόδους σημασιολογικής ομοιότητας.

3.4 Πειραματικά Αποτελέσματα

Ένας τρόπος για μια αρχική εκτίμηση των μεθόδων σημασιολογικής ομοιότητας, είναι για κάποια ζευγάρια όρων του MeSH να συγκρίνουμε τους βαθμούς ομοιότητας που δίνουν οι διάφορες μέθοδοι, με τους βαθμούς ομοιότητας που έχουν υπολογιστεί από ανθρώπους. Η εκτίμηση λοιπόν έγινε χρησιμοποιώντας την Pearson correlation συνάρτηση(εξίσωση 3.4.1). Υποθέτουμε πως έχουμε δύο μεταβλητές X και Y, με μέσες τιμές $\bar{X}$ και $\bar{Y}$, και διασπορά σ_x και σ_y. Η εξίσωση που περιγράφει τη συνάρτηση είναι η εξής:

$$r=\frac{\sum_{i-1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)\sigma_x\sigma_y} \quad (3.4.1)$$

Όπου X και Y στην περίπτωσή μας είναι οι τιμές της εκάστοτε σημασιολογικής μεθόδου και οι ανθρώπινες αντίστοιχα.

Η συσχέτιση όπως φαίνεται στους παρακάτω πίνακες γίνεται σε 36 ζευγάρια όρων που έχουν επιλεγεί από το MeSH. Οι πίνακες είναι χωρισμένοι σε μεθόδους με βάση το σημασιολογικό κριτήριο σύμφωνα με το οποίο έχουν υπολογιστεί. Αναλυτικότερα στον πίνακα 3.4.1 παρουσιάζονται οι μέθοδοι Shortest Path, Weighted Links, Lietal που έχουν δημιουργηθεί σύμφωνα με το κριτήριο υπολογισμού ακμής(κεφάλαιο 2.8.1), στον πίνακα 3.4.2 παρουσιάζονται οι μέθοδοι Lin, Lord et al, Jiang et al και Resnik που έχουν δημιουργηθεί με το κριτήριο περιεχομένου πληροφορίας(κεφάλαιο 2..8.2) ενώ τέλος στον πίνακα 3.4.3 παρουσιάζονται οι μέθοδοι Tversky η οποία είναι μέθοδος που βασίζεται στα χαρακτηριστικά(κεφάλαιο 2.8.3) και οι μέθοδοι Rodriguez, X-Similarity, X-SimilarityMetathesaurus που έχουν δημιουργηθεί με υβριδικά κριτήρια(κεφάλαιο 2.8.4).

Επίσης παρόλο που οι υβριδικές μέθοδοι είναι και Cross Ontology Methods εμείς θα παρουσιάσουμε αποτελέσματα βασισμένα σε μία οντολογία, γιατί στην εργασία αυτή δε μας απασχολεί να κρίνουμε την απόδοση των μεθόδων σε όλες τις πτυχές τους, αλλά να κρίνουμε αν μία οντολογία η οποία είναι προσαρμοσμένη σε μία συλλογή κειμένων είναι καλύτερη ή χειρότερη από μία η οποία δεν είναι.

Από τις μεθόδους υπολογισμού ακμής βλέπουμε ότι η καλύτερη μέθοδος είναι η Lietal με ένα πάρα πολύ καλό αποτέλεσμα της τάξης του 0.71. Όπως έχουμε πει, οι μέθοδοι αυτοί στηρίζονται στο βάθος που βρίσκονται δύο όροι στο δέντρο και στην απόσταση μεταξύ τους. Αξίζει να σημειωθεί ότι η Lietal είναι μία πάρα πολύ γρήγορη μέθοδος, κάτι που βοηθάει στην έρευνα αλλά και στις εφαρμογές.

Από τις μεθόδους περιεχομένου πληροφορίας καλύτερη απόδοση έχει η μέθοδος του Lin με ένα αποτέλεσμα της τάξης του 0.72. Αυτή η μέθοδος, όπως είπαμε, στηρίζεται στην κοινή πληροφορία, που έχουν δύο όροι, η οποία βγαίνει με τον πρακτικό τρόπο του κοινού πατέρα.

Από τις υβριδικές μεθόδους τα καλύτερα αποτελέσματα τα παίρνουμε από την Rodriguez και από την X-similarity. Η X-similarityMetathesaurus, παρουσιάζει χειρότερα αποτελέσματα από τις άλλες δύο. Υπάρχει όμως ένα σημείο που μας επιτρέπει να πιστεύουμε ότι σε ένα σύστημα Ανάκτησης Πληροφορίας μπορεί να δουλέψει αρκετά καλά. Το σημείο αυτό είναι ότι η X-SimilarityMetathesaurus δίνει καλά αποτελέσματα για τα ζευγάρια με μεγάλη ομοιότητα και εμάς αυτό που μας ενδιαφέρει είναι η ομοιότητα μεταξύ των όρων και όχι η ανομοιότητα. Επίσης οι υβριδικές μέθοδοι δεν στηρίζονται πολύ στην ιεραρχία της οντολογίας. Το μόνο σημείο που μπαίνει η ιεραρχία είναι ότι συγκρίνονται τα hypernyms και τα hyponyms των δύο όρων.

	Human	shortestPat	Weighted Links	Li et al
Anemia - Appendicitis	0.03125	0.65	0.61	0.13
Dementia Atopic Dermatitis	0.0625	0.6	0.58	0.1
Bacterial Pneumonia - Malaria	0.15625	0.65	0.61	0.13
Osteoporosis - Patent Ductus Arteriosus	0.15625	0.6	0.58	0.1
Amino Acid Sequence - Anti-Bacterial Agents	0.15625	0	0.31	0
Acquired Immunodeficiency Syndrome - Congenital Heart Defects	0.0625	0.7	0.65	0.16
Otitis Media - Infantile Colic	0.15625	0.55	0.55	0.08
Meningitis - Tricuspid Atresia	0.03125	0.65	0.61	0.13
Sinusitis - Mental Retardation	0.03125	0.65	0.61	0.13
Hypertension - Kidney Failure	0.5	0.65	0.61	0.13
Hyperlipidemia - Hyperkalemia	0.15625	0.85	0.88	0.51
Hypothyroidism - Hyperthyroidism	0.40625	0.9	0.92	0.63
Sarcoidosis - Tuberculosis	0.40625	0.5	0.53	0.07
Vaccines - Immunity	0.59375	0	0.43	0
Asthma - Pneumonia	0.375	0.85	0.88	0.52
Diabetic Nephropathy - Diabetes Mellitus	0.5	0.95	0.95	0.77
Lactose Intolerance - Irritable Bowel Syndrome	0.46875	0.75	0.85	0.36
Urinary Tract Infection - Pyelonephritis	0.65625	0.8	0.86	0.42
Neonatal Jaundice - Sepsis	0.1875	0.7	0.65	0.16
Sickle Cell Anemia - Iron Deficiency Anemia	0.4375	0.75	0.85	0.36
Psychology - Cognitive Science	0.59375	0.972	0.97	0.8
Adenovirus - Rotavirus	0.4375	0.75	0.83	0.35
Migraine - Headache	0.71875	0.6	0.7	0.17
Myocardial Ischemia - Myocardial Infarction	0.75	0.95	0.97	0.8
Hepatitis B - Hepatitis C	0.5625	0.9	0.94	0.66
Carcinoma - Neoplasm	0.75	0.85	0.87	0.45
Pulmonary Valve Stenosis - Aortic Valve Stenosis	0.53125	0.9	0.93	0.66
Failure to Thrive - Malnutrition	0.625	0.65	0.61	0.13
Breast Feeding - Lactation	0.84375	0.75	0.6	0.08
Antibiotics - Antibacterial Agents	0.9375	1	1	0.99
Seizures - Convulsions	0.84375	0.95	0.97	0.81
Pain - Ache	0.875	1	1	0.99
Malnutrition Nutritional Deficiency	0.875	1	1	0.98
Measles - Rubeola	0.90625	1	1	0.99
Chicken Pox - Varicella	0.96875	1	1	0.99
Down Syndrome Trisomy 21	0.875	1	1	0.99
	Human	**shortestPath**	**Weighted Links**	**Li et al**
Human correlation	1	0.509952134	0.640352544	0.705976039

πίνακας 4.3.1

	Human	Lin	Lord et al	Jiang et al	Resnik
Anemia - Appendicitis	0.03125	0	0	0.19	0
Dementia Atopic Dermatitis	0.0625	0	0	0.16	0
Bacterial Pneumonia - Malaria	0.15625	0	0	0.29	0
Osteoporosis - Patent Ductus Arteriosus	0.15625	0	0	0.03	0
Amino Acid Sequence - Anti-Bacterial Agents	0.15625	0	0	0.15	0
Acquired Immunodeficiency Syndrome - Congenital Heart Defects	0.0625	0	0	0.27	0
Otitis Media - Infantile Colic	0.15625	0	0	0.07	0
Meningitis - Tricuspid Atresia	0.03125	0	0	0.19	0
Sinusitis - Mental Retardation	0.03125	0	0	0.36	0
Hypertension - Kidney Failure	0.5	0	0	0.21	0
Hyperlipidemia - Hyperkalemia	0.15625	0.39	0.286	0.47	0.33
Hypothyroidism - Hyperthyroidism	0.40625	0.72	0.48	0.75	0.65
Sarcoidosis - Tuberculosis	0.40625	0	0	0.25	0
Vaccines - Immunity	0.59375	0	0	0.52	0
Asthma - Pneumonia	0.375	0.8	0.4	0.87	0.52
Diabetic Nephropathy - Diabetes Mellitus	0.5	0.74	0.44	0.79	0.58
Lactose Intolerance - Irritable Bowel Syndrome	0.46875	0.47	0.37	0.47	0.47
Urinary Tract Infection - Pyelonephritis	0.65625	0.6	0.37	0.67	0.47
Neonatal Jaundice - Sepsis	0.1875	0	0	0.19	0
Sickle Cell Anemia - Iron Deficiency Anemia	0.4375	0.72	0.45	0.76	0.6
Psychology - Cognitive Science	0.59375	0.77	0.46	0.81	0.62
Adenovirus - Rotavirus	0.4375	0.32	0.23	0.45	0.26
Migraine - Headache	0.71875	0.26	0.2	0.37	0.23
Myocardial Ischemia - Myocardial Infarction	0.75	0.84	0.43	0.89	0.57
Hepatitis B - Hepatitis C	0.5625	0.82	0.47	0.86	0.64
Carcinoma - Neoplasm	0.75	0.62	0.21	0.85	0.24
Pulmonary Valve Stenosis - Aortic Valve Stenosis	0.53125	0.78	0.48	0.81	0.65
Failure to Thrive - Malnutrition	0.625	0	0	0.18	0
Breast Feeding - Lactation	0.84375	0	0	0.04	0
Antibiotics - Antibacterial Agents	0.9375	1	0.63	1	1
Seizures - Convulsions	0.84375	0.89	0.55	0.9	0.8
Pain - Ache	0.875	1	0.57	1	0.86
Malnutrition Nutritional Deficiency	0.875	1	0.46	1	0.62
Measles - Rubeola	0.90625	1	0.6	1	0.92
Chicken Pox - Varicella	0.96875	1	0.63	1	1
Down Syndrome Trisomy 21	0.875	1	0.63	1	1
	Human	**Lin**	**Lord et al**	**Jiang et al**	**Resnik**
Human correlation	1	0.723524427	0.701574406	0.710080769	0.718107379

πίνακας 4.3.2

	Human	Tversky	Rodriguez	X-similarity	X-similarityMet
Anemia - Appendicitis	0.03125	0.2	0.1	0.113	0.4
Dementia Atopic Dermatitis	0.0625	0.2	0.1	0.092	0.4
Bacterial Pneumonia - Malaria	0.15625	0.2	0.1	0.091	0.107
Osteoporosis - Patent Ductus Arteriosus	0.15625	0.2	0.1	0.099	0.4
Amino Acid Sequence - Anti-Bacterial Agents	0.15625	0	0	0.02	0.096
Acquired Immunodeficiency Syndrome - Congenital Heart Defects	0.0625	0.25	0.125	0.099	0.42
Otitis Media - Infantile Colic	0.15625	0.16	0.083	0.077	0.39
Meningitis - Tricuspid Atresia	0.03125	0.2	0.1	0.081	0.4
Sinusitis - Mental Retardation	0.03125	0.2	0.1	0.078	0.4
Hypertension - Kidney Failure	0.5	0.2	0.1	0.077	0.44
Hyperlipidemia - Hyperkalemia	0.15625	0.6	0.3	0.21	0.2
Hypothyroidism - Hyperthyroidism	0.40625	0.75	0.375	0.27	0.77
Sarcoidosis - Tuberculosis	0.40625	0.14	0.07	0.069	0.38
Vaccines - Immunity	0.59375	0	0	0.0458	0.2
Asthma - Pneumonia	0.375	0.6	0.3	0.225	0.53
Diabetic Nephropathy - Diabetes Mellitus	0.5	0.75	0.375	0.176	0.47
Lactose Intolerance - Irritable Bowel Syndrome	0.46875	0.57	0.78	0.558	0.55
Urinary Tract Infection - Pyelonephritis	0.65625	0.5	0.25	0.165	0.5
Neonatal Jaundice - Sepsis	0.1875	0.25	0.125	0.102	0.4
Sickle Cell Anemia - Iron Deficiency Anemia	0.4375	0.57	0.285	0.231	0.57
Psychology - Cognitive Science	0.59375	0.8	0.4	0.258	0.55
Adenovirus - Rotavirus	0.4375	0.5	0.25	0.195	0.55
Migraine - Headache	0.71875	0.28	0.143	0.124	0.49
Myocardial Ischemia - Myocardial Infarction	0.75	0.83	0.58	0.384	0.65
Hepatitis B - Hepatitis C	0.5625	0.83	0.416	0.348	0.78
Carcinoma - Neoplasm	0.75	0.4	0.2	0.105	0.42
Pulmonary Valve Stenosis - Aortic Valve Stenosis	0.53125	0.8	0.4	0.264	0.27
Failure to Thrive - Malnutrition	0.625	0.2	0.1	0.066	0.4
Breast Feeding - Lactation	0.84375	0.16	0.083	0.055	0.23
Antibiotics - Antibacterial Agents	0.9375	1	1	1	1
Seizures - Convulsions	0.84375	0.83	0.75	0.48	1
Pain - Ache	0.875	1	1	1	1
Malnutrition Nutritional Deficiency	0.875	1	1	1	1
Measles - Rubeola	0.90625	1	1	1	1
Chicken Pox - Varicella	0.96875	1	1	1	1
Down Syndrome Trisomy 21	0.875	1	1	1	1
	Human	**Tversky**	**Rodriguez**	**X-similarity**	**X-similarityMet**
Human correlation	1	0.6711480	0.69325930	0.690016722	0.677583275

πίνακας 4.3.3

Τέλος πρέπει να τονίσουμε ότι η παραπάνω συσχέτιση δίνει απλά μία πρώτη άποψη για το πόσο καλή είναι μία μέθοδος, χωρίς να είναι απόλυτα έγκυρα τα αποτελέσματα. Μια καλύτερη άποψη για τα αποτελέσματα των μεθόδων είναι μέσα από μία εφαρμογή Ανάκτησης Πληροφορίας. Αυτό είναι και το θέμα του επόμενου κεφαλαίου.

4.Εφαρμογή Ανάκτησης Πληροφορίας στο Ohsumed

Αυτό το κεφάλαιο είναι πολύ σημαντικό, γιατί με αυτό κλείνει η έρευνά μας. Δηλαδή θα πάρουμε απάντηση στο εάν μία οντολογία προσαρμοσμένη σε μία συλλογή κειμένων είναι καλύτερη από μία οντολογία μεγαλύτερη που δεν είναι προσαρμοσμένη. Στην περίπτωσή μας η συλλογή είναι το Ohsumed και η οντολογία μας είναι το MeSH εμπλουτισμένο με σχέσεις από το Metathesaurus. Η προσαρμοσμένη οντολογία προήλθε από την παραπάνω και περιέχει τους σημαντικότερους όρους που εξάχθηκαν με την AMTEx μέθοδο. Επίσης θα εκτιμήσουμε την απόδοση της σημασιολογικής μεθόδου X-SimilarityMetathesaurus που προτείναμε στο κεφάλαιο 3. Στο κεφάλαιο 3 είδαμε ένα τρόπο για μία αρχική εκτίμηση των σημασιολογικών μεθόδων με τη χρήση της συσχέτισης με τα 36 ζευγάρια όρων, αλλά όπως είπαμε εάν δεν εφαρμόσουμε τα αποτελέσματα των σημασιολογικών μεθόδων σε ένα σύστημα Ανάκτησης Πληροφορίας δεν μπορούμε να είμαστε σίγουροι για την αποτελεσματικότητα τους. Έτσι λοιπόν σε αυτό το κεφάλαιο αναφερόμαστε στο πως δημιουργήθηκε το σύστημα Ανάκτησης Πληροφορίας, το οποίο στηρίζεται σε ένα διαφορετικό μοντέλο από το απλό VSM. Το μοντέλο στο οποίο στηριχθήκαμε είναι το μοντέλο της Rada Michalcea(κεφάλαιο 2.6). Επίσης σε αυτό το κεφάλαιο θα αναλύσουμε τα precision-recall διαγράμματα που θα εμφανιστούν.

Η συλλογή κειμένων στην οποία έγινε η Ανάκτηση Πληροφορίας είναι το Ohsumed(κεφάλαιο 2.7).

4.1 Μέθοδος που Χρησιμοποιήσαμε

Αρχικά δεικτοδοτήσαμε το Ohsumed παίρνοντας τους MeSH όρους από τον τίτλο και το κείμενο. Ύστερα πήραμε ένα σύνολο 63 ερωτήματα μαζί με 100 απαντήσεις για το κάθε ερώτημα, δοσμένες από άνθρωπο. Αυτό το σύνολο δημιουργήθηκε για να υπάρχει εκτίμηση των αποτελεσμάτων Ανάκτησης πάνω στο Ohsumed.

Το απλούστερο και πιο συνηθισμένο μοντέλο Ανάκτησης είναι το VSM το οποίο για να βρει την ομοιότητα μεταξύ δύο κειμένων ψάχνει τους κοινούς μόνο όρους, για παράδειγμα το pain και το ache παρόλο που είναι συνώνυμα θα τα δει σαν δύο διαφορετικές λέξεις. Η ομοιότητα στο VSM υπολογίζεται από την παρακάτω εξίσωση.

$$Sim(d_i, d_j) = \frac{\sum w_{it} w_{js}}{\sum_t w_{it}^2 \sum_s w_{js}^2}$$

όπου w_{it} και w_{js} τα βάρη των όρων των δύο κειμένων.

Εμείς στην εργασία μας προτιμήσαμε τη χρήση του μοντέλου που πρότεινε η Rada Mihalcea, το οποίο χρησιμοποιεί και τη σημασιολογική ομοιότητα των όρων που έχει εξαχθεί από κάποιο similarity measure. Δηλαδή δε θα καταλάβει μόνο τους συνώνυμους όρους όπως pain και ache, αλλά και αρκετά όμοιους όπως carcinoma και neoplasma. Η εξίσωση που περιγράφει το μοντέλο αυτό είναι η 2.6.1. Το μοντέλο αυτό δουλεύει με όρους. Αυτό λοιπόν που χρειαζόμαστε για το μοντέλο αυτό είναι το idf του κάθε όρου.

Αξίζει ακόμα να αναφέρουμε ότι δεν ήταν δυνατόν να πάρουμε όλα τα ζευγάρια όρων του MeSH και να υπολογίσουμε την ομοιότητά τους με κάποια μέθοδο σημασιολογικής ομοιότητας. Για αυτόν το λόγο λοιπόν διαλέγουμε εμείς τα ζευγάρια που θα πάρουμε. Η διαδικασία λοιπόν που ακολουθήσαμε ήταν να βρούμε τους σημαντικότερους όρους του Ohsumed με την μέθοδο της AMTEx. Η μέθοδος αυτή όπως είπαμε διαλέγει τους σημαντικότερους όρους από μία συλλογή κειμένων με το C/NC Value, και από αυτούς κρατάει μόνο τους MeSH όρους. Είναι επίσης κατανοητό ότι εφαρμόζεται πάνω σε ιατρικές συλλογές κειμένων. Οι όροι που βρέθηκαν είναι 4.329.

4.2 Διαγράμματα Precision-Recall

Το precision-recall διάγραμμα εκτιμάει την απόδοση των μεθόδων ανάκτησης πληροφορίας. Εδώ οφείλουμε να αναφέρουμε πως υπολογίζονται τα precision και recall.

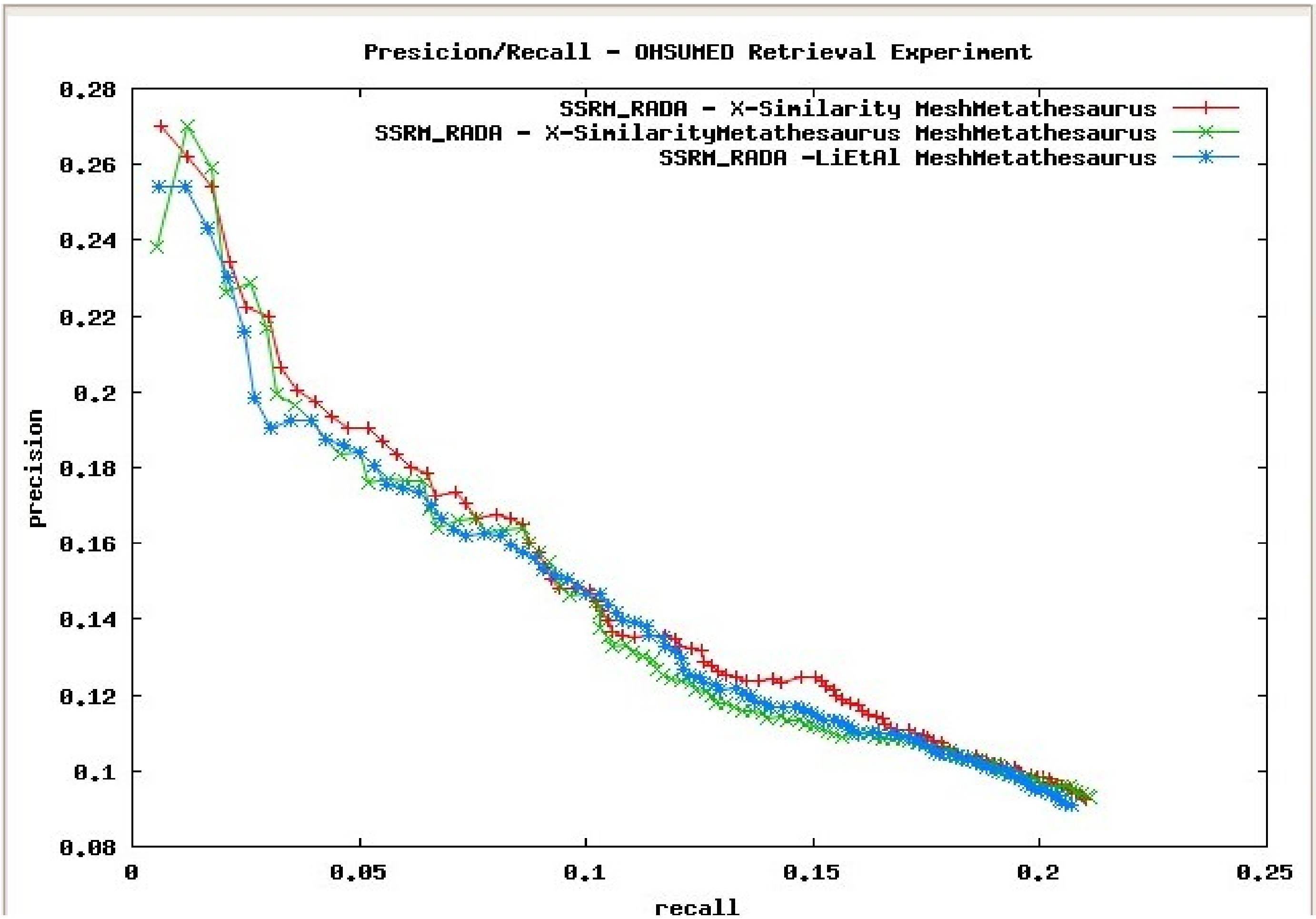
$$precision = \frac{|\{relevant\ documents\} \cap \{documents\ retrieved\}|}{|\{documents\ retrieved\}|}$$

$$recall=\frac{|\{relevant\,documents\}\cap\{documents\,retrieved\}|}{|\{relevant\,documents\}|}$$

Με αυτόν τον τρόπο υπολογίζουμε κάθε σημείο των διαγραμμάτων precision-recall, όταν αναφερόμαστε σε μία ερώτηση. Τα παρακάτω διαγράμματα έχουν προέλθει από ένα σύνολο 63 ερωτήσεων και 100 απαντήσεων. Ο αριθμός των σημείων είναι όσος και ο αριθμός των απαντήσεων, δηλαδή 100. Για να έχουμε ένα πιο αντιπροσωπευτικό precision-recall λαμβάνουμε υπόψη μας όλες τις ερωτήσεις. Δηλαδή για κάθε απάντηση παίρνουμε τον μέσο όρο όλων των precision και recall. Για παράδειγμα αν για την πρώτη απάντηση έχουμε για την ερώτηση1 precision=p1 recall=r1, ερώτηση2 precsion=p2 recall=r2, ..., ερώτηση63 precision=p63 recall=r63, τότε για την πρώτη απάντηση θα έχουμε precision=(p1+p2+...+p63)/63 και recall=(r1+r2+...+r63)/63.

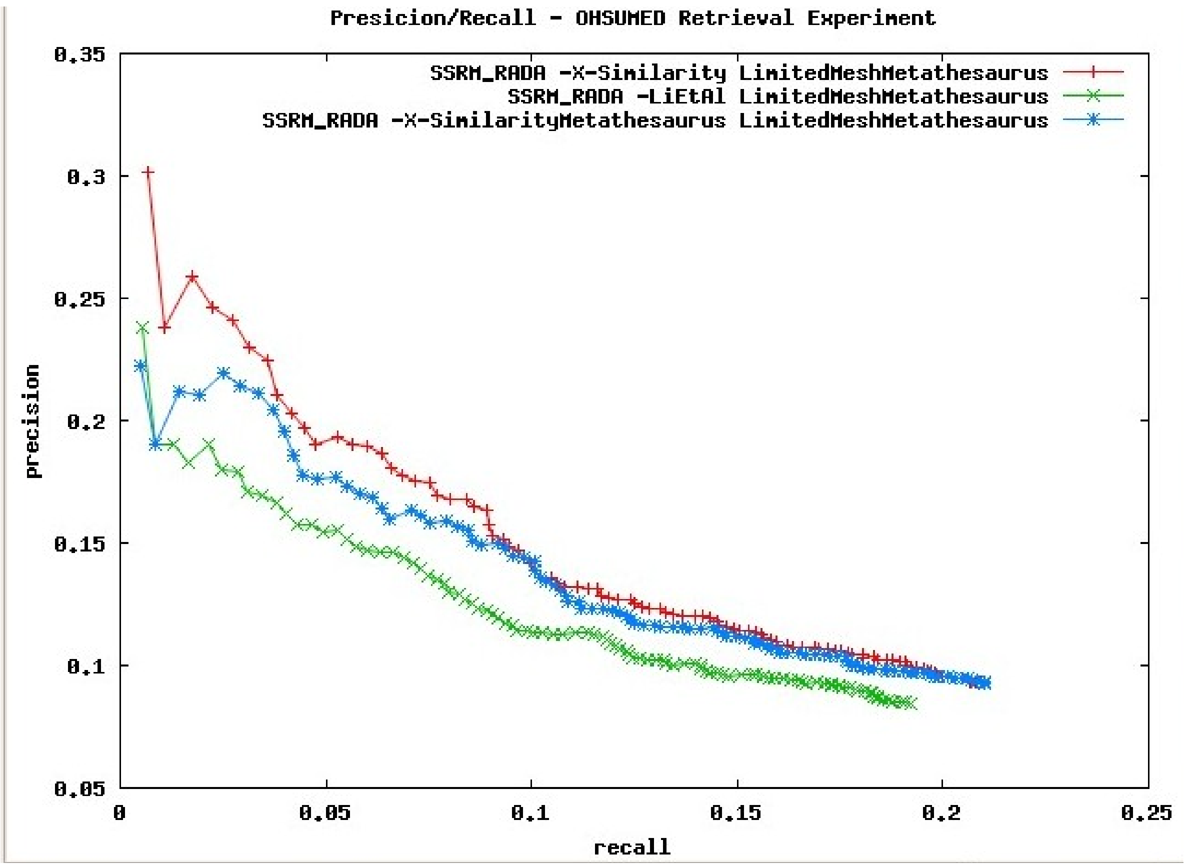
Τα διαγράμματα που παρουσιάζονται πιο κάτω έχουν όλα γίνει σύμφωνα με το μοντέλο της Rada Mihalcea. Σε αυτό στο οποίο διαφέρουν είναι ο τρόπος με τον οποίο πήραμε την ομοιότητα μεταξύ των όρων, δηλαδή η μέθοδος σημασιολογικής ομοιότητας. Οι μέθοδοι σημασιολογικής ομοιότητας που συγκρίναμε είναι οι εξής: η Lietal λόγω των καλών αποτελεσμάτων της στη συσχέτιση (correlation) του κεφαλαίου 3.4 και της ταχύτητάς της, το X-similarity και το X-similarityMetathesaurus.

Στο πρώτο διάγραμμα φαίνονται τα precision-recall των τριών μεθόδων σημασιολογικής ομοιότητας πάνω σε ολόκληρη την οντολογία του MeshMetathesaurus:



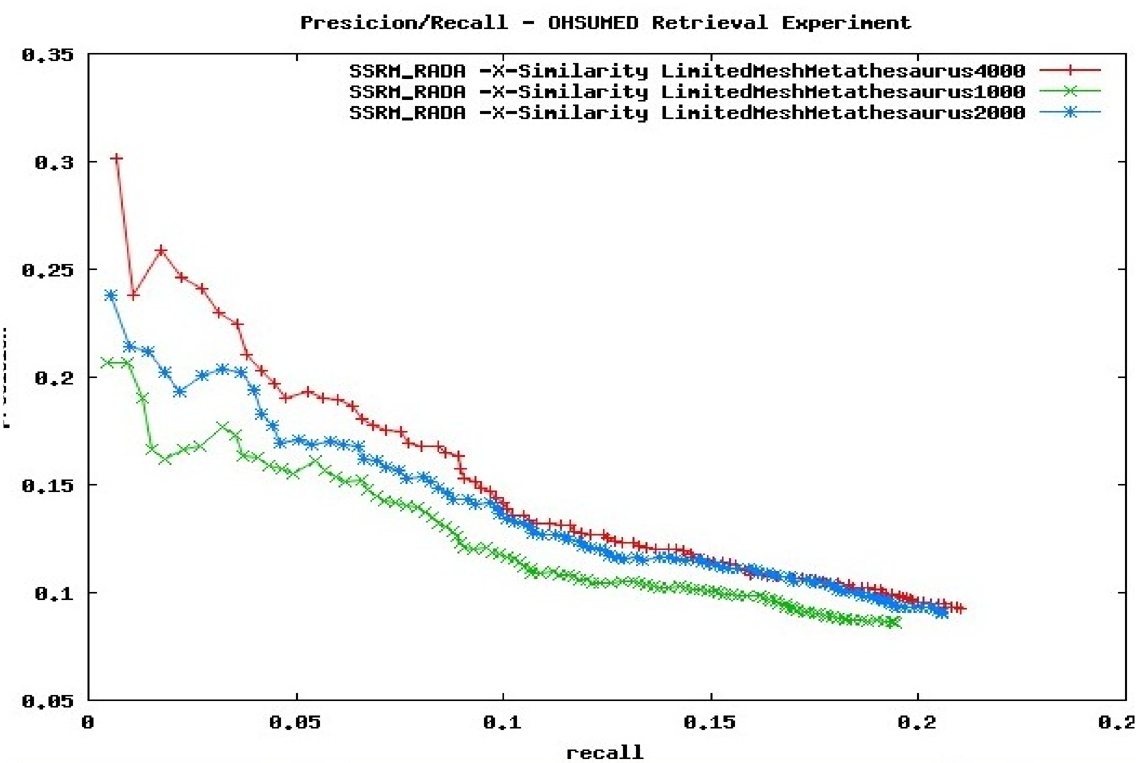
Παρατηρούμε ότι δεν υπάρχουν σημαντικές διαφορές στις μεθόδους μεταξύ τους σε ολόκληρο το MeshMetathesaurus.

Στο επόμενο διάγραμμα φαίνονται τα precision-recall των X-Similarity, X-SimilarityMetathesaurus και Lietal για την περιορισμένη οντολογία με τους 4.329 όρους η οποία φτιάχτηκε όπως φαίνεται στο κεφάλαιο 3.2:



Σε αυτό το διάγραμμα παρατηρούμε ότι η καλύτερη μέθοδος από όλες για την περιορισμένη οντολογία εμφανίζεται να είναι η X-Similarity με δεύτερη την X-SimilarityMetathesaurus και τελευταία την Lietal. Εάν συγκρίνουμε τα δύο παραπάνω διαγράμματα μεταξύ τους θα παρατηρήσουμε ότι η απόδοση της X-Similarity δεν μειώνεται με τον περιορισμό της οντολογίας στη συλλογή των κειμένων, της X-SimilarityMetathesaurus μειώνεται πολύ λίγο, ενώ στην περίπτωση της Lietal μειώνεται σε μεγάλο βαθμό. Αυτό οφείλεται στο ότι χάνουμε πληροφορία στην περίπτωση του Lietal, κάτι που είναι πάρα πολύ λογικό να συμβαίνει αφού η Lietal στηρίζεται μόνο στη θέση των όρων στην ιεραρχία όπως φαίνεται στο κεφάλαιο 2.8.1, δηλαδή στην απόσταση των όρων μεταξύ τους και στο βάθος που βρίσκονται. Επομένως αλλάζοντας την ιεραρχία μας αλλάζει και η ομοιότητα των όρων μεταξύ τους. Αντίθετα η X-Similarity και η X-SimilarityMetahesurus δεν στηρίζονται στην ιεραρχία τόσο πολύ και γι' αυτό δεν παρουσιάζουν μεγάλες διαφορές. Η μόνη ενασχόληση των μεθόδων αυτών με την ιεραρχία είναι η σύγκριση των hypernyms και hyponyms των δύο όρων με την fearure based μέθοδο, Tversky.

Τέλος στο επόμενο διάγραμμα έχουμε τα precision-recall του μοντέλου της Rada Mihalcea, με το X-Similartiy ως μεθόδων σημασιολογικής ομοιότητας. Η διαφορά των καμπυλών μεταξύ τους είναι ότι στηρίχθηκαν πάνω σε οντολογίες με τους 4.329, 2.000, 1000, σημαντικότερους όρους του Ohsumed, οι οποίοι εξάχθηκαν με την μέθοδο AMTEx. Οι καμπύλες λοιπόν φαίνονται παρακάτω:



Είναι φανερό από αυτό το διάγραμμα ότι όσο περισσότερο περιορίζουμε την οντολογία μας τόση περισσότερη πληροφορία χάνουμε.

Επομένως από τα παραπάνω αποτελέσματα βλέπουμε ότι με το μοντέλο της Rada Mihalcea και με χρήση της X-Similarity ως μέθοδο σημασιολογικής ομοιότητας δεν χάνουμε πληροφορία, αλλά ούτε κερδίζουμε εάν φτιάξουμε μια οντολογία μόνο με τους σημαντικότερους όρους που εξάγαμε. Εάν όμως μειώσουμε κι άλλο τους όρους, τότε έχουμε μείωση πληροφορίας πράγμα το οποίο φαίνεται στα διαγράμματα.

5.Συμπεράσματα

Όπως είπαμε στην αρχή ο σκοπός μας ήταν να δούμε κατά πόσο επηρεάζεται μία οντολογία εάν περιορίσουμε τους όρους της. Με άλλα λόγια εάν η νέα οντολογία περιγράφει καλύτερα ή χειρότερα μία γνωστική περιοχή εάν περιοριστεί σε αυτήν. Αυτός ο προβληματισμός υπάρχει καιρό και δεν είχε ερευνηθεί μέχρι τώρα. Σε αυτήν την εργασία με την μεθοδολογία που ακολουθήσαμε φάνηκε ότι δεν κερδίζουμε αλλά ούτε χάνουμε πληροφορία. Στο συμπέρασμα αυτό οδηγηθήκαμε κάνοντας μία εφαρμογή Ανάκτησης Πληροφορίας που στηρίχθηκε πάνω στο μοντέλο της Rada Mihalcea και στη σημασιολογική μέθοδο της X-Similarity.

Επίσης φτιάξαμε μία μέθοδο σημασιολογικής ομοιότητας για την συγκεκριμένη οντολογία MeshMetathesaurus περιμένοντας να βγάλουμε καλύτερα αποτελέσματα, εφόσον στηριχθήκαμε και στις σχέσεις του Metathesaurus. Ο λόγος που προτιμήσαμε αυτές τις σχέσεις είναι ότι ενώνουν όρους του MeSH ξεχωριστά και όχι ομάδες όρων όπως γίνεται στο Semantic Network. Ίσως μία έρευνα με σχέσεις από το Semantic Network θα μπορούσε να γίνει στο μέλλον. Τα αποτελέσματα πιθανότατα θα ήταν καλύτερα εάν οι σχέσεις ήταν λιγότερο γενικές από αυτές του Metathesaurus. Θα ήταν προτιμότερο να βρίσκαμε σχέσεις οι οποίες προέρχονται απευθείας από τη συγκεκριμένη συλλογή κειμένων που θέλουμε να ασχοληθούμε.

Βιβλιογραφία

[1]Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley Longman, 1999.

[2]S.J. Nelson, D. Johnston, and B.L. Humphreys. Relationships in Medical Subject Headings. In C.A. Bean and R. Green, editors, Relationships in the Organization of Knowledge, pages 171 184. Kluwer Academic Publishers, New York, 2001.

[3]W. Douglas Johnston Stuart J. Nelson and Betsy L. Humphreys. Relationships in Medical Subject Headings (MeSH). In National Library of Medicine, Bethesda, MD, USA, 2002.

[4]P.R. Cohen and R. Kjeldsen. Information Retrieval by Constrained Spreading Activation in Semantic Networks. Information Processing and Management, 23(4):255 268, 1987.

[5]G. Salton and C. Buckley. On the Use of Spreading Activation Methods in Automatic Information. In Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 147 160. ACM Press, 1988.

[6]H. Bulskov, R. Knappe, and T. Andreasen. On Measuring Similarity for Con-

ceptual Querying. In T. Andreasen, A. Motro, H. Christiansen, and H.L. Larsen, editors, Proceedings of the 5th International Conference on Flexible Query Answering Systems (FQAS 02), volume 2522 of LNAI, pages 100 111, Copenhagen, Denmark, 27-29 October 2002.

[7]R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17 30, January/February 1989.

[8]R. Richardson, A. Smeaton, and J. Murphy. Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words. Technical Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland, 1994.

[9]Graeme Hirst and David St-Onge. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Proceedings of Fellbaum, pages 305 332, 1998.

[10]Yuhua Li, Zuhair A. Bandar, and David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE 50 Transactions on Knowledge and Data Engineering, 15(4):871 882, July/August 2003.

[11]Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric

for semantic similarity in wordnet. 2004

[12]P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation. Bioinformatics, 19(10):1275 83, 2003.

[13]O. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. Journal of Artificial Intelligence Research, 11:95 130, 1999.

[14]D. Lin. Principle-Based Parsing Without Overgeneration. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 93), pages 112 120, Columbus, Ohio, 1993.

[15]J.J. Jiang and D.W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of the International Conference on Research in Computational Linguistic, Taiwan, 1998.

[16]A. Tversky. Features of Similarity. Psycological Review, 84(4):327 352, 1977.

[17]M.A. Rodriguez and M.J. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Tramsactions on Knowledge and Data Engineering, 15(2):442 456, March/April 2003.

[18]X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies Euripides G.M. Petrakis, Giannis Varelas,Angelos Hliaoutakis, Paraskevi Raftopoulou

Department of Electronic and Computer EngineeringTechnical University of Crete (TUC).

[19]Mike Uschold & Michael Gruinger Ontologies: Principles, Methods and Applications. February 1996. The University of Edinburgh, 1996.

[20]R. Beckwith and G. A. Miller. Implementing a Lexical Network. Technical Report 43, Princeton University, 1992.

[21]K. Knight and S. Luk. Building a Large-Scale Knowledge Base for Machine Translation. In Proceedings of thr National Conference on Artificial Intelligence (AAAI 94), Seattle, WA, 1994.

[22]T. O Hara, N. Salay, M. Witbrock, D. Schneider, B. Aldag, S. Bertolo, K. Panton, F. Lehmann, and et al. Inducing Criteria for Mass Noun Lexical Mappings 51 using the Cyc KB, and its Extension to WordNet. In Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5), Tilburg, The Netherlands, 15-17 January 2003.

[23]S. Reed and D. Lenat. Mapping Ontologies into Cyc. In Proceedings of the AAAI 02 Conference Workshop on Ontologies For The Semantic Web, Edmon-

ton, Canada, July 2002.

[24]S.J. Nelson, T. Powell, and B.L. Humphreys. The Unified Medical Language System (UMLS) Project. In A. Kent and C.M. Hall, editors, Encyclopedia of Library and Information Science, pages 369   378. Marcel Dekker, Inc., New York, 2002.

[25]K.Frantzi and S.Ananiadou. The C/NC Value Domain Independent Method for Multi-Word Term Extraction.
Journal of Natural Language Processing, 6(3):145-180, 1999.

[26]Angelos Hliaoutakis, Kalliopi Zervanou, Euripides G.M. Petrakis, Evangelos E. Milios. Automatic Document Indexing in Large Medical Collections. Technical University of Crete, Dalhousie Univerity of Canada.

[27]Christian Blaschke, A.V., Automatic Ontology Construction from the Literature.

[28]David Faure, C.N., ASIUM: learning subcategorization frames and restrictions of selection.

[29]Philipp Cimiano, A.H., Stefen Staab, Clustering Concept Hierarchies fromText.

[30]Wolff, K.E., A First Course in Formal Concept Analysis

[31]Alexander Maedche, R.V., The Ontology Extraction & Maintenance Framework Text-To-Onto.

[32]Varma, V., Building Large Scale Ontology Networks.

[33]Rada Mihalcea, Courtney Corley, Carlo Strapparava, Corpus-based and Knowledge-based Measures of Text Semantic Similarity, in Proceedings of the American Association for Artificial Intelligence (AAAI 2006), Boston, July 2006.