



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ  
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ  
ΒΙΒΛΙΟΘΗΚΗ  
εισ. 81303  
Αρ.  
παξ.

Διπλωματική Εργασία  
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«Αυτόματη Επέκταση Επερώτησης με χρήση Ιεραρχικού Θησαυρού»

Βασιλείου Παναγιώτα

Επιβλέπων: κ. Μ. Βαζιργιάννης

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2007

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ  
ΚΑΤΑΛΟΓΟΣ



0 000000 606059



# ΠΕΡΙΕΧΟΜΕΝΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Συνοπτική Περιγραφή του Θέματος της Διπλωματικής Εργασίας.....	3
1. Εισαγωγή.....	4
2. Αναζήτησης/Ανάκτησης Πληροφορίας.....	7
2.1 Εισαγωγή.....	7
2.2 Σύντομη παρουσίαση της διαδικασίας.....	8
2.3 Περιγραφή των Βασικών Μοντέλων της.....	10
2.4 Αξιολόγηση των Συστημάτων Ανάκτησης Πληροφορίας.....	13
3. Επερωτήσεις.....	17
3.1 Περιγραφή των τρόπων σχηματισμού των επερωτήσεων.....	17
3.2 Παρουσίαση πράξεων & λειτουργιών των επερωτήσεων.....	18
4. Word Sense Disambiguation.....	22
4.1 Εισαγωγή.....	22
4.2 Χρησιμότητα του Word Sense Disambiguation.....	23
4.3 Περιγραφή Προσεγγίσεων του Word Sense Disambiguation.....	25
5. WordNet.....	30
5.1 Εισαγωγή.....	30
5.2 Περιγραφή της λεξιλογικής βάσης του WordNet.....	31
5.3 Εφαρμογές του WordNet.....	40
6. Περιγραφή του συστήματος ανάκτησης που υλοποιήθηκε.....	46
7. Πειράματα.....	48
7.1 Συλλογές.....	48
7.2 Περιγραφή Πειραμάτων - Εμφάνιση Αποτελεσμάτων.....	48
7.3 Σχολιασμός Αποτελεσμάτων.....	65
8. Γενικά Συμπεράσματα - Μελλοντική Εργασία.....	67
9. Βιβλιογραφία.....	69



## **Συνοπτική Περιγραφή του Θέματος της Διπλωματικής Εργασίας**

Ένα από τα βασικότερα προβλήματα της διαδικασίας της Ανάκτησης Πληροφορίας (Information Retrieval, IR) με χρήση μικρού μήκους επερωτήσεων, δηλαδή ερωτήσεων με λίγες λέξεις, είναι ότι χωρίς περαιτέρω επεξεργασία, δεν μπορούν να ανακτηθούν όλα τα κείμενα που έχουν κοινό θέμα, παρά μόνο τα κείμενα που περιέχουν κοινές λέξεις με την επερώτηση.

Η μελέτη των προβλημάτων στο χώρο της Ανάκτησης της Πληροφορίας έδωσε το έναυσμα για την δημιουργία αυτής της διπλωματικής εργασίας. Πιο συγκεκριμένα, το κίνητρο που υποκίνησε την προσπάθεια αυτή είναι να αναπτυχθεί μια προσέγγιση για την επίλυση του παραπάνω προβλήματος, που αφορά την βελτίωση της απόδοσης της ανάκτησης της πληροφορίας, μέσω της εκμετάλλευσης της σημασιολογική πληροφορίας που υπάρχει στους ιεραρχικούς θησαυρούς και της διαδικασίας του Word Sense Disambiguation, που στοχεύει στην αντιμετωπίσει τη πολυνησημίας των λέξεων της φυσικής γλώσσας.

Η προσέγγιση που φαίνεται να οδηγεί στην λύση είναι η χρήση της μεθόδου επέκτασης επερώτησης (query expansion), όπου στην επερώτηση προστίθενται λέξεις που είναι σημασιολογικά όμοιες με τις λέξεις, που έχει αρχικά. Η εύρεση τέτοιων λέξεων είναι δυνατόν να πραγματοποιηθεί με τη χρήση ιεραρχικών θησαυρών ή ηλεκτρονικών λεξικών. Στην προκειμένη περίπτωση, αυτή γίνεται με τη χρήση του WordNet, που δεν είναι απλά ένας θησαυρός, αλλά ένα ολοκληρωμένο λεξικογραφικό σύστημα, που στοχεύει στην εύρεση των λέξεων, που είναι εννοιολογικά συνώνυμες μεταξύ τους. Η βασική του δομή είναι ένα σύνολο συνωνύμων, που λέγεται synset. Όμως, το WordNet, εκτός από το να ταξινομεί τις διαφορετικές ερμηνείες των λέξεων στα διαφορετικά synsets, παρέχει και κάποιες ισχύουσες σχέσεις, μεταξύ των διαφορετικών synsets, όπως είναι της συνεπαγωγής, των υπονύμων και των υπερνύμων.

Όπως είναι φανερό κι από τα παραπάνω, σκοπός της διπλωματικής είναι η ανάπτυξη ενός συστήματος που να κάνει επέκταση επερωτήσεων, με διαφορετικούς τρόπους, εκμεταλλευόμενοι τις σχέσεις του WordNet με ή χωρίς την ανάδραση του χρήστη (relevance feedback) αυτόματα. Στην περίπτωση, που η διαδικασία γίνεται αυτόματα, απαραίτητη είναι η εύρεση μιας μεθόδου που να αναθέτει βάρη στα υπέρνυμα των όρων της επερώτησης, ώστε να βοηθούν στην ανάκτηση κείμενων που είναι περισσότερο όμοια με την επερώτηση.

Στα πλαίσια της εργασίας υλοποιήθηκαν αρκετά πειράματα σε μικρές συλλογές, ώστε να μπορεί να γίνει σύγκριση των αποτελεσμάτων που προκύπτουν ανάμεσα στην αρχική επερώτηση και την επερώτηση που προκύπτει από την επέκταση. Η σύγκριση των διαφορετικών τρόπων επέκτασης της επερώτησης, δηλαδή η χρήση μόνο ορισμών των λέξεων ή ορισμών και υπερνύμων των ορισμών, έχει ως αποτέλεσμα να γίνει αντιληπτό εάν βελτιώνεται ή όχι η απόδοση της ανάκτησης.

## 1. Εισαγωγή

Οι έρευνες κατά την δεκαετία του '90 έδειχναν ότι οι περισσότεροι άνθρωποι προτιμούσαν να λαμβάνουν πληροφορίες ερχόμενοι σε προσωπική επαφή με άτομα που είχαν εξειδικευμένες γνώσεις στο θέμα που τους ενδιέφερε, παρά από αυτόματα συστήματα ανάκτησης πληροφορίας. Χαρακτηριστικό ήταν το παράδειγμα της κράτησης εισιτηρίων μέσω ταξιδιωτικών πρακτόρων. Παρόλο που και οι ακαδημαϊκές συζητήσεις στο πεδίο αυτό ήταν περιορισμένες την προηγούμενη δεκαετία, οι σημερινές επίμονες προσπάθειες βελτίωσης των μέτρων απόδοσης της ανάκτησης οδήγησαν τα συστήματα ανάκτησης πληροφοριών και τις μηχανές αναζήτησης του Web σε επίπεδα απόδοσης που να ικανοποιούν τις ανάγκες όλο και περισσότερων χρηστών. Έτσι, η εύρεση πληροφορίας από τέτοιου είδους συστήματα έχει μετατραπεί στην πιο δημοφιλή και αξιόπιστη επιλογή.

Πιο συγκεκριμένα, η Ανάκτηση της Πληροφορίας (Information Retrieval, IR) στοχεύει στην εύρεση υλικού (συνήθως εγγράφων) από μια μη δομημένη φύση (unstructured nature), συνήθως κείμενα (text), που ικανοποιεί την ανάγκη για πληροφόρηση, μέσα από μεγάλες συλλογές, όπως είναι συνήθως οι τοπικοί εξυπηρετητές υπολογιστών (computer servers) ή το Internet.

Αυτή η ανάγκη για πληροφόρηση του χρήστη, χρειάζεται να συνοψιστεί σε μια ερώτηση, συνήθως με τη χρήση λέξεων-κλειδιών (επερώτηση, query), ώστε οι μηχανές αναζήτησης, ή γενικότερα οποιοδήποτε IR σύστημα, να μπορούν να την επεξεργαστούν. Δυστυχώς, η επεξεργασία μιας άμεσης και πλήρης περιγραφής της επιθυμητής πληροφορίας του χρήστη από τα συστήματα IR είναι αδύνατη.

Η εύρεση, λοιπόν, της κατάλληλης επερώτησης αποτελεί και το μεγαλύτερο πρόβλημα στο χώρο της Ανάκτησης της Πληροφορίας, αφού αυτή καθορίζει την πληροφορία που είναι σχετική σημασιολογικά με αυτήν που αναζητά ο χρήστης, και ένας λάθος σχηματισμός της επιφέρει πολύ αρνητικά αποτελέσματα στην απόδοση της ανάκτησης. Τα συστήματα προσπαθούν να ανακαλύψουν την ομοιότητα των κειμένων με την επερώτηση, γι' αυτό και κρίνουν σημαντικό να κατατάσσουν ιεραρχικά τα αποτελέσματα, από το περισσότερο σχετικό κείμενο στο λιγότερο, ώστε ο χρήστης να βρίσκει την επιθυμητή πληροφορία στις πρώτες επιλογές, χωρίς να χάνει χρόνο στην εξέταση μιας μεγάλης λίστας αποτελεσμάτων. Το παραπάνω, αν και πολύ σημαντικό, είναι πολύ δύσκολο να επιτευχθεί, μιας και το σύστημα δεν επιστρέφει μόνο τα κείμενα που είναι σχετικά με την ερώτηση του χρήστη.

Το γεγονός που κάνει τα συστήματα ανάκτησης, να μην επιστρέφουν μόνο τα επιθυμητά αποτελέσματα, εξαιρώντας προς το παρόν την ευθύνη μιας ακατάλληλης επερώτησης, είναι τα προβλήματα στην ανάλυση της φυσικής γλώσσας. Η πολυσημία των λέξεων, δηλαδή οι περισσότερες από μια ερμηνεία που διαθέτει η κάθε λέξη, αποτελεί και το μεγαλύτερο πρόβλημα, αφού είναι δυνατόν να ανακτηθούν κείμενα που περιέχουν μεν τις λέξεις της επερώτησης, αλλά τις αποδίδουν διαφορετική ερμηνεία. Απαραίτητη, για τον παραπάνω λόγο κρίνεται η χρήση της διαδικασίας Word Sense Disambiguation (WSD), που έχει ως στόχο την αποσαφήνιση της ερμηνείας των όρων. Η διαδικασία αυτή είναι μείζονος σημασίας, αφού χωρίς αυτή δεν μπορεί να βρεθεί η ακριβής ερμηνεία των όρων, όπως αυτή καθορίζεται από την επερώτηση, με άμεσο αντίκτυπο τα αποτελέσματα της ανάκτησης να μην είναι τα επιθυμητά.

Επανερχόμενοι, λοιπόν, στα προβλήματα που παρατηρούνται με τις επερωτήσεις χωρίς την περαιτέρω επεξεργασία τους, είναι απαραίτητο να διευκρινιστεί ότι είναι φυσικό να χρησιμοποιούνται διαφορετικές λέξεις (συνώνυμες) για την περιγραφή της σημασιολογικής ερμηνείας των επερωτήσεων και των κειμένων, με άμεσο αποτέλεσμα να μην ανακτώνται



κείμενα που είναι σχετικά με την επερώτηση και να ανακτώνται μόνο τα κείμενα που περιέχουν τις λέξεις της επερώτησης.

Μια από τις λύσεις που αρχικά προτάθηκαν και που υιοθετήθηκε στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας για την βελτίωση της απόδοσης της ανάκτησης ήταν η επέκταση της αρχικής επερώτησης με νέους όρους, καθώς και η τροποποίηση του βάρους των όρων της επερώτησης.

Πιο συγκεκριμένα, έγινε προσπάθεια από τα συστήματα να εκμεταλλεύονται τη γνώση του χρήστη για την πληροφορία που αναζητά. Στόχος ήταν να εξεταστούν από το χρήστη τα αποτελέσματα της ανάκτησης και να τα κατηγοριοποιούσε σε σχετικά και μη σχετικά. Χρησιμοποιώντας, έπειτα, το σύστημα αυτή την πληροφορία, μπορούσε να επιλέγει περαιτέρω όρους από τα σχετικά κείμενα, για να πραγματοποιηθεί η επέκταση της ερώτησης με καινούριους όρους.

Εντούτοις, αυτή η προσέγγιση είχε και πάλι να αντιμετωπίσει κάποια προβλήματα, με σημαντικότερο την αδυναμία του χρήστη να εξετάσει τα αποτελέσματα, όταν αυτά είναι πάρα πολλά. Όπως, λοιπόν, είναι φανερό, προέκυψε η ανάγκη για μια πιο αυτόματη διαδικασία εύρεσης των σχετικών κειμένων. Η επόμενη σκέψη αφορούσε τη χρησιμοποίηση όρων που προέρχονται είτε από τους όρους του κείμενου που είναι κοντά στους όρους της επερώτησης (διπλανές λέξεις), είτε από τα συνώνυμα των όρων της επερώτησης, είτε από την εφαρμογή της διαδικασίας Stemming στους όρους της επερώτησης.

Είναι πολύ σημαντικό να σημειωθεί ότι η εύρεση των συνωνύμων και γενικά των ορισμών κάθε όρου της επερώτησης είναι δυνατόν να βρεθούν από ηλεκτρονικά λεξικά ή ιεραρχικούς θησαυρούς λέξεων. Ένα παράδειγμα τέτοιων δομών είναι και το WordNet, που δεν αποτελεί απλά ένα θησαυρό, αλλά ένα ολοκληρωμένο λεξικογραφικό σύστημα που περιέχει τόσο όλους τους δυνατούς ορισμούς κάθε λέξης, όσο και τις εκάστοτε σχέσεις που μπορεί να υπάρχουν ανάμεσα τους, όπως για παράδειγμα σχέσεις υπερνύμων και υπονύμων. Με τη χρήση, λοιπόν, των παραπάνω δομών είναι δυνατόν να βρεθούν οι ερμηνείες του κάθε όρου της επερώτησης και στη συνέχεια να επιλεγούν οι πιο κατάλληλοι ορισμοί για την επέκταση αυτών.

Σκοπός, λοιπόν, αυτής της διπλωματικής εργασίας είναι η έρευνα τρόπων επέκτασης της επερώτησης, καθώς και η τροποποίηση των βαρών των όρων της επεκταμένης επερώτησης, για τη βελτίωση της Ανάκτησης της Πληροφορίας. Γι' αυτό το λόγο, εκτελέστηκαν αρκετά πειράματα σε τέσσερις μικρές συλλογές, τις cactm, medline, cran και cis, που διαθέτουν επερωτήσεις και τις σχετικές απαντήσεις τους, για να είναι εφικτή η επαλήθευση των αποτελεσμάτων των διαφορετικών προσεγγίσεων.

Η βελτίωση ή η μείωση της απόδοσης της Ανάκτησης αξιολογείται, σύμφωνα με τα μέτρα της ακρίβειας (precision) και της ανάκλησης (recall). Πιο συγκεκριμένα, γίνεται βάση της καμπύλης precision-recall στα 11 (αντί για 10) συγκεκριμένα επίπεδα του recall, που είναι το 0%, 10%, 20%, ..., 100%. Για το επίπεδο του recall 0%, η τιμή του precision καθορίζεται με βάση της διαδικασίας της παραβολής.

Στο σύστημα ανάκτησης που υλοποιήθηκε για την εκτέλεση των πειραμάτων, υιοθετήθηκε το διανυσματικό (Vector Space) μοντέλο IR, όπου η αναπαράσταση κείμενων και επερωτήσεων γίνεται μέσω ενός διανύσματος από βάρη, διάστασης όσο το πλήθος των keywords. Τα βάρη αυτά καθορίζονται από την συχνότητα των όρων (TF\*IDF). Οι όροι των κειμένων και των επερωτήσεων επεξεργάζονται, ώστε να αφαιρεθούν λέξεις που δεν είναι χρήσιμες για την ανάκτηση (stopwords), όπως a, and, the κ.τ.λ. και κάθε λέξη μετασχηματίζεται βάσει της διαδικασίας του Stemming, στην ρίζα της, σύμφωνα με τον αλγόριθμο του Porter.

Παραδείγματος χάρη, οι λέξεις connecting, connection, connections έχουν την ίδια λέξη "stem", το connect. Ο υπολογισμός της ομοιότητας γίνεται βάση του συνημίτονου της γωνίας που δημιουργούν τα διανύσματα των κειμένων με την επερώτηση. Τέλος, γίνεται η κατάταξη των αποτελεσμάτων, σύμφωνα με το βαθμό ομοιότητα τους με την επερώτηση.

Αναλυτικά, τα πειράματα που εκτελέστηκαν στο παραπάνω σύστημα είναι τα ακόλουθα:

Η πρώτη σειρά πειραμάτων αφορά την αναπαράσταση των κειμένων και των επερωτήσεων σύμφωνα με τις έννοιες των όρων τους (senses), όπως καθορίζονται από το WordNet και τη διαδικασία του disambiguation[3].

Εν συντομίᾳ, η διαδικασία του WSD που υιοθετήθηκε απεικονίζει τις γειτονικές λέξεις των κειμένων και των επερωτήσεων σε senses που είναι συμπαγής στον ιεραρχικό θησαυρό. Οι λέξεις που αποσαφηνίζονται είναι μόνο ουσιαστικά, εξαιτίας του ότι αυτά έχουν από μόνα τους μια ερμηνεία, σε αντίθεση με τα ρήματα που εκφράζουν κυρίως σχέσεις μεταξύ των λέξεων.

Όσο αφορά τη δεύτερη σειρά πειραμάτων, η αναπαράσταση των κειμένων και των επερωτήσεων γίνεται σύμφωνα με τους όρους τους και όλα τα υπέρνυμα αυτών, που παρέχονται από τις σχέσεις που διαθέτει το WordNet.

Η τρίτη σειρά πειραμάτων ακολουθεί την ίδια φιλοσοφία με την δεύτερη, με τη μόνη διαφορά ότι αντί να χρησιμοποιούνται όλα τα υπέρνυμα στην αναπαράσταση των κειμένων και των επερωτήσεων, χρησιμοποιούνται κάθε φορά μόνο δύο, τέσσερα, έξι ή οκτώ, στην περίπτωση φυσικά που ο όρος διαθέτει τόσα υπέρνυμα.

Στην συνέχεια, επανεκτελούνται οι παραπάνω σειρές πειραμάτων με τη μόνη διαφορά ότι δεν συμμετέχουν στην αναπαράσταση των κειμένων και των επερωτήσεων οι όροι τους, αλλά μόνο τα senses και τα υπέρνυμα τους, ανάλογα με το πείραμα που μελετάμε.

Η τελευταία σειρά πειραμάτων αφορά μια διαφορετική προσέγγιση, που στοχεύει όχι στην τυχαία επιλογή του πλήθος των υπερνύμων, αλλά στην ανακάλυψη της σημαντικότητας του κάθε υπέρνυμου μέσω της ανάθεσης του ενός βάρους. Το βάρος κάθε υπέρνυμου προκύπτει σύμφωνα με τη μελέτη [5].

Εν συντομίᾳ, αυτή η μελέτη προτάσσει έναν αλγόριθμο που χρησιμοποιώντας περιορισμούς (τύπου must-link, cannot-link) για σημεία στο  $\mathbb{R}^n$  είναι δυνατόν να κατασκευάσει μια μετρική απόσταση στο  $\mathbb{R}^n$ , που θα αντικατοπτρίζει αυτούς τους περιορισμούς. Υιοθετώντας τον αλγόριθμο αυτό, και για την περίπτωση της ανάκτησης κειμένων, χρησιμοποιούνται σαν σημεία του  $\mathbb{R}^n$  τα διανύσματα των κείμενων των συλλογών που δημιουργούνται σύμφωνα με τις τιμές συχνότητας εμφάνισης των όρων των κειμένων της συλλογής (TF\*IDF) που περιέχει το εκάστοτε κείμενο, οπότε προκύπτει ένας τετραγωνικός πίνακας διάστασης, όσο είναι οι όροι της συλλογής, όπου σε κάθε όρο αντιστοιχεί ένα βάρος. Αυτό το βάρος προκύπτει σύμφωνα με τους περιορισμούς, τύπου must-link και cannot-link των κειμένων που προέρχονται από τα αρχικά αποτελέσματα του συστήματος και τα αποτελέσματα ύστερα από την ανάδραση του χρήστη.



## 2. Αναζήτησης/Ανάκτησης Πληροφορίας

### 2.1 Εισαγωγή

Η Ανάκτηση Πληροφορίας (Information Retrieval, IR) διαπραγματεύεται την παρουσίαση, την αποθήκευση, την οργάνωση της πληροφορίας, καθώς και τον τρόπο πρόσβασης σε αυτήν. Πιο συγκεκριμένα, οι ενέργειες που φαίνονται να είναι άμεσα χρήσιμες στους χρήστες είναι η παρουσίαση και η οργάνωση της πληροφορίας, αφού χάρις αυτών τους παρέχεται εύκολη πρόσβαση στην πληροφορία που τους ενδιαφέρει.

Πρέπει, όμως, σε αυτό το σημείο να επισημανθεί ότι η εύρεση της πληροφορίας που χρειάζεται ο χρήστης δεν είναι ένα απλό πρόβλημα, εξαιτίας της αδυναμίας άμεσης χρησιμοποίησης μιας πλήρους περιγραφής της επιθυμητής πληροφορίας στις τρέχοντες διεπαφές των συστημάτων ανάκτησης πληροφορίας, καθώς και των μηχανών αναζήτησης του Web, που στην συγκεκριμένη περίπτωση θεωρούνται για απλούστευση ένα παράδειγμα του προβλήματος της ανάκτησης πληροφορίας. Ο χρήστης δυστυχώς, θα πρέπει να συνοψίζει και να προσαρμόζει αυτή την πληροφορία σε μια ερώτηση με χρήση λέξεων-κλειδιών, ώστε να μπορεί να επεξεργαστεί από τις μηχανές αναζήτησης ή από οποιοδήποτε IR σύστημα.

Ο στόχος των IR συστημάτων, δοθέντος της επερώτησης του χρήστη, είναι να ανακτήσουν πληροφορία που θα είναι χρήσιμη και σχετική με την σημασιολογική ερμηνεία της πληροφορίας που αναζητά ο χρήστης. Σε αυτό το σημείο είναι σημαντικό να τονιστεί ότι σε αυτήν την διαδικασία δίνεται έμφαση στην ανακτηθέντα πληροφορία και όχι στα ανακτηθέντα δεδομένα.

Με τον όρο ανακτηθέντα δεδομένα στο περιβάλλον των IR συστημάτων, αποκαλούνται τα κείμενα της συλλογής που περιέχουν τις λέξεις-κλειδιά της επερώτησης του χρήστη. Όπως είναι κατανοητό, τις περισσότερες φορές, δεν είναι δυνατόν από μόνα τους, αυτά τα δεδομένα να παρέχουν ολοκληρωμένα την πληροφορία που επιζητά ο χρήστης. Έτσι αποδεικνύεται ότι ο χρήστης ενδιαφέρεται περισσότερο για την ανακτηθείσα πληροφορία, δηλαδή την σημασιολογική πληροφορία των κειμένων, παρά για τα ανακτηθέντα δεδομένα που ικανοποιούν την δοθείσα επερώτηση. Για παράδειγμα, σκοπός της γλώσσας των ανακτηθέντων δεδομένων είναι να ανακτήσουν όλα τα αντικείμενα που ικανοποιούν πλήρως καθορισμένες ιδιότητες, όπως τις εκφράσεις της σχεσιακής άλγεβρας. Η ανάκτηση ενός λάθους αντικειμένου ανάμεσα σε εκατοντάδες ανακτηθέντα αντικείμενα, σε ένα σύστημα ανακτηθέντων δεδομένων θα αποτελούσε ολοκληρωτική αποτυχία, αντίθετα σε ένα σύστημα ανακτηθείσας πληροφορίας είναι δυνατόν ανακτηθέντα αντικείμενα που ίσως να μην είναι ακριβή ή να περιέχουν μικρά λάθη, να μην επισημανθούν. Ο κύριος λόγος γι' αυτήν την διαφορά έγκειται στο ότι η ανακτηθείσα πληροφορία, συνήθως, διαχειρίζεται κείμενα στην φυσική γλώσσα, που ποτέ δεν είναι καλά δομημένα ή/και μπορούν να έχουν διφορούμενη σημασιολογία, αντίθετα τα συστήματα ανάκτησης δεδομένων, όπως οι σχεσιακές βάσεις, διαχειρίζονται δεδομένα που έχουν συγκεκριμένη δομή και σημασιολογία.

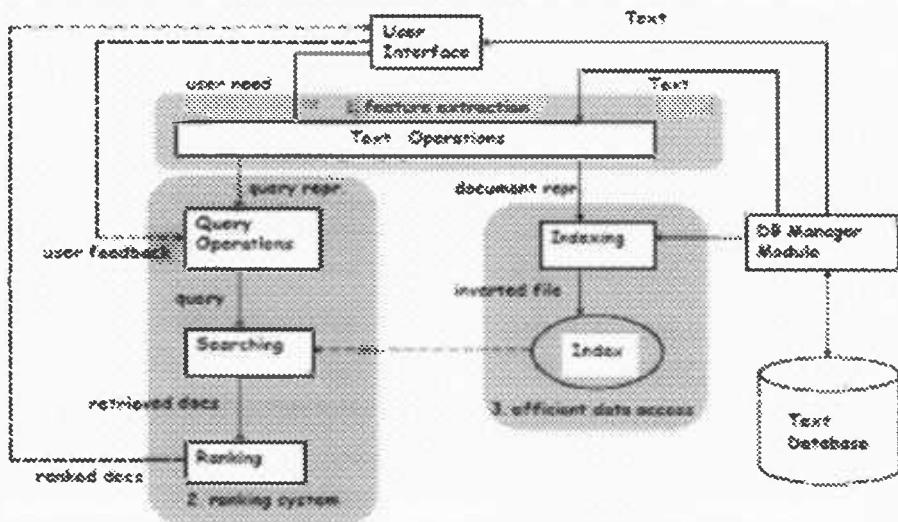
Ενώ τα ανακτηθέντα δεδομένα παρέχουν την λύση στα προβλήματα των συστημάτων βάσης, είναι αδύνατο να επιλύσουν το πρόβλημα της ανακτηθείσας πληροφορίας για ένα ζήτημα ή θέμα. Οπότε, για να μπορούν τα συστήματα IR να είναι αποτελεσματικά και να ανακτούν την πληροφορία που επιθυμεί ο χρήστης, θα πρέπει, με κάποιον τρόπο, να ερμηνεύουν το περιεχόμενο των αντικειμένων πληροφορίας (κείμενα) της συλλογής και να τα κατατάσσουν σύμφωνα με τον βαθμό ομοιότητας τους με την επερώτηση του χρήστη. Αυτή η ερμηνεία του περιεχομένου του κειμένου περιλαμβάνει την εξαγωγή, τόσο της συντακτικής, όσο και της σημασιολογικής πληροφορίας από το κείμενο. Η χρήση των παραπάνω πληροφοριών κατευθύνει την εύρεση των κειμένων, που είναι όμοια με την επιθυμητή

πληροφορία του χρήστη. Όμως, είναι εξίσου δύσκολο να γίνει τόσο η εξαγωγή αυτής της πληροφορίας, όσο και η χρησιμοποίηση της για την απόφαση της ομοιότητας.

Τέλος, κρίνεται απαραίτητο να επισημανθεί ότι η ιδέα της ομοιότητας, καθώς και η σημασία της σωστής κατάταξης των αποτελεσμάτων αποτελούν τα σημαντικότερα στοιχεία στην ανάκτηση της πληροφορίας. Είναι σημαντικό, τα συστήματα, αφού βρουν τρόπους που να αποκαλύπτουν την ομοιότητα των κειμένων με την επερώτηση, να κατατάσσουν iεραρχικά τα αποτελέσματα από το περισσότερο σχετικό κείμενο στο λιγότερο, ώστε ο χρήστης να βρίσκει την επιθυμητή πληροφορία στις πρώτες επιλογές, χωρίς να χάνει χρόνο στην εξέταση μιας μεγάλης λίστας αποτελεσμάτων. Είναι φυσικό ο πρωταρχικός ρόλος κάθε IR συστήματος να είναι η ανάκτηση όλων των κειμένων που είναι σχετικά με την ερώτηση του χρήστη, χωρίς να ανακτήσει έως και καθόλου μη-σχετικά κείμενα. Παρόλο αυτά, επειδή αυτό δεν είναι πάντα εφικτό είναι σημαντικό η πιο σχετική πληροφορία να είναι ψηλά στην κατάταξη, ώστε να εξυπηρετεί καλύτερα και γρηγορότερα τις ανάγκες των χρηστών.

## 2.2 Σύντομη παρουσίαση της διαδικασίας της Ανάκτησης Πληροφορίας

Για τον σκοπό της περιγραφής της διαδικασίας της ανάκτησης, χρησιμοποιείται η παρακάτω απλή προγραμματιστική αρχιτεκτονική.



Καταρχήν, πριν αρχίσει η διαδικασία της Ανάκτησης είναι απαραίτητο να οριστεί η βάση των κειμένων. Οι ενέργειες που πρέπει να προσδιοριστούν είναι ο καθορισμός των κειμένων που θα χρησιμοποιηθούν στην ανάκτηση, οι λειτουργίες (operations) που θα εκτελεστούν στα κείμενα και το μοντέλο των κειμένων, που διαπραγματεύεται ποια θα είναι η δομή των κειμένων και ποια τα στοιχεία που θα ανακτηθούν.

Οι λειτουργίες των κειμένων μετασχηματίζουν τα αρχικά κείμενα και προσδιορίζουν την λογική τους αναπαράσταση (logical view), δηλαδή την απεικόνισή τους με λέξεις-κλειδιά. Για την μείωση αυτής της απεικόνισης, εκτελούνται κάποιες ενέργειες, όπως η απαλοιφή των τετριμένων λέξεων (stopwords), όπως είναι τα άρθρα και οι σύνδεσμοι, η χρήση stemming, που αποκόπτει τις καταλήξεις των λέξεων, με στόχο την αναπαράσταση τους με τις γραμματικές τους ρίζες, καθώς και η αναγνώριση των ομάδων των ουσιαστικών, με σκοπό την εξάλειψη των επιθέτων, των επιφρημάτων και των ρημάτων. Για να γίνει πιο κατανοητή η διαδικασία του Stemming παρακάτω παρουσιάζεται ένα παράδειγμα κειμένου, στο οποίο εκτελούνται τρεις διαφορετικοί αλγόριθμοι Stemming, του Porter, που είναι ο πιο δημοφιλής, του Lovins, που είναι ο παλαιότερος και του Paice, που ανήκει στην κατηγορία των καινούριων αλγορίθμων.

*Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation*

*Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpre*

*Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret*

*Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret*

Είναι φανερό από τα παραπάνω, ότι η χρήση αυτών των πράξεων μειώνουν κατά πολύ την πολυπλοκότητα στην αναπαράσταση των κειμένων και επιτρέπουν την μετατροπή της λογικής τους αναπαράσταση από ολόκληρο κείμενο σε ένα σύνολο λέξεων.

Στην συνέχεια, αφού δημιουργηθεί η λογική αναπαράσταση των κειμένων, υλοποιείται το ευρετήριο τους που επιτρέπει την γρήγορη αναζήτηση σε μεγάλο όγκο δεδομένων. Είναι σημαντικό να επισημανθεί ότι οι πόροι που καταναλώνονται για τον ορισμό της βάσης και την δημιουργία του ευρετηρίου αποσβένονται κατά την διαδικασία της επεξεργασίας των επερωτήσεων των συστημάτων ανάκτησης.

Αφού εκτελεστούν όλα τα παραπάνω, η διαδικασία της ανάκτησης μπορεί να αρχίσει. Αφού, ο χρήστης προσδιορίσει την επιθυμητή πληροφορία, στην συνέχεια αυτή πρέπει να μετασχηματιστεί σύμφωνα με τις λειτουργίες που χρησιμοποιήθηκαν και στα κείμενα. Έπειτα, επεξεργάζεται η επερώτηση, ώστε να ανακτηθούν τα κείμενα. Σε αυτό το σημείο σημαντικός είναι ο ρόλος του ευρετηρίου που συντελεί στην γρήγορη επεξεργασία της ερώτησης.

Όμως, πριν επιστραφούν στον χρήστη τα κείμενα, κατατάσσονται σύμφωνα με την ομοιότητα τους με την επερώτηση. Το πιο δημοφιλές μέτρο ομοιότητας είναι το cosine similarity. Στην συνέχεια, ο χρήστης είναι δυνατόν να εξετάζει αυτό το σύνολο κειμένων για να αναζητήσει την χρήσιμη πληροφορία. Σε αυτό το σημείο, είναι πιθανόν ο χρήστης να προσδιορίσει ένα υποσύνολο από κείμενα που έχει εξετάσει, ως σχετικά και αρχίζει ο κύκλος της ανάδρασης με τον χρήστη. Σε αυτό τον κύκλο, το σύστημα χρησιμοποιεί τα επιλεγμένα κείμενα του χρήστη για να τροποποιήσει το σχηματισμό της επερώτησης. Η τροποποίηση αυτή βοηθάει στην απεικόνιση της αληθινής ανάγκης του χρήστη.

Τέλος, θα πρέπει να επισημανθεί, σκεπτόμενοι τις διεπαφές των χρηστών που είναι διαθέσιμες στα τρέχοντα συστήματα ανάκτησης πληροφοριών, συμπεριλαμβανομένου των μηχανών αναζήτησης του Web και των Web browsers, ότι ο χρήστης σχεδόν ποτέ δεν δηλώνει την πληροφορία που επιθυμεί, αλλά αντιθέτως, απαιτεί να του παραχθεί μια άμεση απεικόνιση της ερώτησης που να μπορεί να εκτελέσει το σύστημα. Παράλληλα η επερώτηση συχνά είναι ακατάλληλη, αφού οι χρήστες δεν έχουν γνώση των λειτουργιών του κειμένου και της επερώτησης. Τα παραπάνω δεδομένα οδηγούν στην παρατήρηση ότι ο σχηματισμός ανεπαρκών ερωτήσεων οδηγεί σε ανεπαρκή ανάκτηση.

Αναλυτική αναφορά για την διαδικασία της Ανάκτησης μπορεί να βρεθεί στις παραπομπές της βιβλιογραφίας [1] και [32].



## 2.3 Περιγραφή των Βασικών Μοντέλων της Ανάκτησης Πληροφορίας

Όπως έχει επισημανθεί και παραπάνω, το κύριο πρόβλημα στα συστήματα ανάκτησης πληροφορίας είναι το θέμα της πρόβλεψης των κειμένων που είναι σχετικά με την επερώτηση. Αυτή η απόφαση συνήθως εξαρτάται από τους αλγόριθμους κατάταξης, που έχουν ως στόχο να τοποθετήσουν τα κείμενα που ανακτήθηκαν σε μια σειρά που να προσδίδει το βαθμό ομοιότητας με την επερώτηση. Πιο συγκεκριμένα, όσο πιο ψηλά στην κατάταξη είναι ένα κείμενο, τόσο πιο μεγάλη πιθανότητα υπάρχει να είναι σχετικό με την πληροφορία που αναζητά ο χρήστης.

Ένας αλγόριθμος κατάταξης λειτουργεί σύμφωνα με βασικές προϋποθέσεις, που σχετίζονται με την ιδέα της ομοιότητας των κειμένων. Διαφορετικά σύνολα προϋποθέσεων (σύμφωνα με την ομοιότητα των κειμένων) αποφέρουν διαφορετικά μοντέλα ανάκτησης πληροφορίας. Τα μοντέλα του IR, που υιοθετούνται, καθορίζουν τις προβλέψεις για το τι είναι σχετικό με την επερώτηση και τι δεν είναι.

Κρίνεται απαραίτητο, να τονιστεί ότι υπάρχουν δύο τύποι ανάκτησης: το ad hoc και το filtering. Στα συμβατικά συστήματα ανάκτησης πληροφορίας, τα κείμενα της συλλογής παραμένουν σχετικά στατικά, ενώ μπορούν να υποβάλλονται καινούργιες επερωτήσεις στο σύστημα. Η παραπάνω λειτουργική μέθοδος ορίζεται με τον όρο ad hoc και αποτελεί την πιο δημοφιλής μορφή του user task. Μια παρόμοια, αλλά διαφορετική εργασία είναι αυτή, στην οποία οι επερωτήσεις παραμένουν σχετικά στατικές, ενώ εισέρχονται νέα κείμενα στο σύστημα. Η παραπάνω περιγραφή ορίζει το filtering. Αναλυτική αναφορά για αυτούς τους δύο τύπους περιέχεται στο [1].

Είναι σημαντικό, πριν την περιγραφή των κλασσικών μοντέλων αναπαράστασης των μοντέλων IR, να οριστεί η ακριβής τους έννοια. Τα μοντέλα IR αποτελούνται από τέσσερα συστατικά: το D που είναι το σύνολο που αποτελεί την λογική αναπαράσταση των κειμένων της συλλογής, το Q που αντιστοιχεί στο σύνολο που αποτελεί την λογική αναπαράσταση των πληροφοριών που χρειάζεται ο χρήστης (επερωτήσεις), το F που προσδιορίζει το πλαίσιο εργασίας του μοντέλου της αναπαράστασης των κειμένων, των επερωτήσεων και των σχέσεων τους και το R(q,d) που αποτελεί τη συνάρτηση κατάταξης και συνδέει έναν πραγματικό αριθμό με μια επερώτηση και ένα κείμενο από τα αντίστοιχα σύνολα.

Τα βασικά μοντέλα αναπαράστασης κειμένων θεωρούν ότι κάθε κείμενο αναπαρίσταται από ένα σύνολο λέξεων-κλειδιών, που λέγονται index terms. Ένα index term αποτελεί μια λέξη κειμένου, που η σημασιολογία του βοηθάει στην συγκράτηση του κύριου θέματος του κειμένου. Αυτοί οι όροι χρησιμοποιούνται στο ευρετήριο και περιέχουν μια περιληπτική παρουσίαση των περιεχομένων του κειμένου. Είναι φυσικό, οι διαφορετικοί όροι να μην μπορούν όλοι να είναι το ίδιο χρησιμή για την περιγραφή του περιεχομένου των κειμένων, γι' αυτό χρησιμοποιούνται κάποιες από τις ιδιότητες τους, που είναι εύκολο να μετρηθούν και αξιολογούν το ενδεχόμενο του όρου. Χαρακτηριστικό είναι το παράδειγμα μια συλλογής με εκατοντάδες χιλιάδες κείμενα, όταν κάποιος όρος εμφανίζεται σε όλα τα εκατοντάδες χιλιάδες κείμενα της συλλογής είναι εντελώς ανούσιος ως όρος, αφού δεν δίνει καμιά πληροφορία για το εάν το κείμενο έχει ενδιαφέρον για τον χρήστη. Η παραπάνω ιδιότητα για το πόσο «καλός» είναι ένας όρος κερδίζεται μέσω της εκχώρησης ενός αριθμητικού βάρους σε κάθε όρο του κειμένου. Τα βάρη των όρων είναι αμοιβαία ανεξάρτητα, δηλαδή γνωρίζοντας το βάρος ενός όρου σε ένα κείμενο δεν υπάρχει καμιά πληροφορία για το βάρος κάποιου άλλου όρου στο κείμενο.

Τα βασικά μοντέλα αναπαράστασης κειμένων είναι το Boolean μοντέλο, το Vector-Based μοντέλο και το Probabilistic Retrieval μοντέλο [1] [32]. Το Boolean μοντέλο είναι

παλαιότερο και το πιο απλούστερο μοντέλο, που βασίζεται στο σύνολο της θεωρίας και την Boolean άλγεβρα. Το πλαίσιο εργασίας που παρέχει είναι κατανοητό από οποιονδήποτε χρήστη του IR συστήματος. Οι επερωτήσεις προσδιορίζονται σαν εκφράσεις που έχουν προσδιορισμένη σημασιολογία και τα κάθε κείμενο αναπαρίσταται σαν ένα διάνυσμα δυαδικών τιμών. Δυστυχώς, όμως αυτό το μοντέλο έχει πολλά μειονεκτήματα. Καταρχήν, η στρατηγική της ανάκτησης βασίζεται στο κριτήριο της δυαδικής απόφασης, δηλαδή ένα κείμενο προβλέπεται να είναι είτε σχετικό είτε άσχετο, χωρίς να λαμβάνει υπόψη την ιδέα της κλιμακούμενης κατάταξης, που επιφέρει καλύτερη απόδοση στην ανάκτηση. Με άμεσο αποτέλεσμα το μοντέλο αυτό να αποτελεί περισσότερο ένα μοντέλο ανακτηθέντων δεδομένων παρά ένα ανακτηθείσας πληροφορίας. Δεύτερον, είναι πολύ δύσκολο να μετασχηματιστεί η επιθυμητή πληροφορία σε έκφραση Boolean, λόγω της σημασιολογική τους ακρίβεια. Παρά τα παραπάνω μειονεκτήματα το μοντέλο αυτό αποτελεί μια καλή αρχή στο χώρο της ανάκτησης.

Το δυαδικό μοντέλο, όπως είναι φανερό και από το όνομα του, θεωρεί ότι οι όροι υπάρχουν ή απουσιάζουν από το κείμενο, έτσι τα βάρη τους είναι όλα δυαδικά. Μια επερώτηση αποτελείται από όρους που συνδέονται μεταξύ τους με τους ακόλουθους συνδέσμους: το «όχι» ( $\neg$ ) το «και» ( $\wedge$ ) και το «ή» ( $\vee$ ). Ουσιαστικά, η επερώτηση είναι μια τυπική Boolean έκφραση, που μπορεί να αναπαρασταθεί ως μια διάξευξη συνδετικών διανυσμάτων. Για παράδειγμα, η επερώτηση  $q = k_a \wedge (k_b \wedge \neg k_c)$  μπορεί να γραφτεί σύμφωνα με την ακόλουθη κανονική διαζευκτική μορφή (disjunctive normal form, DNF)  $q_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$ , που κάθε συστατικό της είναι ένα διάνυσμα που αποδίδει το δυαδικό βάρος, σύμφωνα με την τριάδα  $(k_a, k_b, k_c)$ . Η ομοιότητα ενός κειμένου με την επερώτηση ορίζεται ως

$$sim(d_j, q) = \begin{cases} 1, & \text{εάν } \exists \vec{q}_{cc} \text{ έτσι ώστε } (\vec{q}_{cc} \in \vec{q}_{dnf}) \cap (\forall_{k_i}, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0, & \text{διαφορετικά} \end{cases}$$

Εάν  $sim(d_j, q) = 1$ , τότε το μοντέλο προβλέπει ότι το κείμενο  $d_j$  είναι σχετικό με την επερώτηση  $q$ , διαφορετικά δεν είναι σχετικό.

Αναγνωρίζοντας τις περιορισμένες ιδιότητες της χρήσης των δυαδικών βαρών, το επόμενο μοντέλο IR που προτάθηκε ήταν το διανυσματικό (Vector-Based) μοντέλο. Αυτό παρέχει ένα πλαίσιο εργασίας που είναι δυνατή η μερική αντιστοίχηση των επερωτήσεων με τα κείμενα, χρησιμοποιώντας θετικά, μη δυαδικά βάρη στους όρους τους. Πιο συγκεκριμένα, αυτά τα βάρη χρησιμοποιούνται στον υπολογισμό της ομοιότητας κάθε κειμένου που είναι αποθηκευμένο στο σύστημα με την επερώτηση του χρήστη. Ταξινομώντας, στην συνέχεια τα ανακτηθέντα κείμενα σε φθίνουσα σειρά του βαθμού ομοιότητας, το διανυσματικό μοντέλο λαμβάνει υπόψη του κείμενα που είναι μερικώς σχετικά με τους όρους της επερώτησης, με αποτέλεσμα τα αποτελέσματα να είναι ακριβέστερα από αυτά του Boolean μοντέλου. Κάθε κείμενο και κάθε επερώτηση αναπαρίσταται ως ένα διάνυσμα πραγματικών τιμών, τιμών που αντιστοιχούν στο βάρος των όρων που περιέχει κάθε ένα αντίστοιχα. Για παράδειγμα, εάν το  $w_{i,q}$  αντιστοιχεί στο βάρος που σχετίζει το ζευγάρι  $[k_i, q]$  με  $w_{i,q} \geq 0$ , τότε το διάνυσμα της επερώτησης ορίζεται ως  $\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$ , όπου  $t$  το συνολικό πλήθος των όρων στο σύστημα. Αντίστοιχα το διάνυσμα των κείμενων αναπαρίσταται ως  $\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$ . Η συσχέτιση των δύο παραπάνω διανυσμάτων μπορεί να ποσοτικοποιηθεί, για παράδειγμα με το

$$\text{συνημίτονο της γωνίας τους, δηλαδή ως } sim(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^t w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{i=1}^t w_{i,q}^2}}. \text{ Ο όρος } q$$

δεν επηρεάζει την κατάταξη, γιατί είναι ίδιος για όλα τα κείμενα. Όπως είναι φανερό, το μέτρο της ομοιότητας παίρνει τιμές  $[0, +1]$ .

Έχουν προταθεί πολλές τεχνικές για τον υπολογισμό των πραγματικών αριθμών που ανατίθενται σε κάθε ένα από τους όρους («βάρη»), με τον πιο κοινό να είναι τα βάρη TF-IDF, που ορίζονται με βάση το γινόμενο των τιμών του TF και του IDF. Για την επεξήγηση των παραπάνω μέτρων γίνεται η υπόθεση ότι ο συνολικός αριθμός των κειμένων στο σύστημα είναι  $N$  και το πλήθος των κειμένων που εμφανίζεται ο όρος  $k_i$  είναι  $n_i$ . Ο όρος  $TF_{i,j}$  ορίζεται με βάση το πλήθος συχνότητας του όρου  $k_i$  στο κείμενο  $d_j$ . Το κανονικοποιημένο  $TF_{i,j} = \frac{TF_{i,j}}{\sum_{l=1}^t TF_{l,j}}$ , όπου

$t$  το πλήθος όλων των όρων που περιέχει το κείμενο  $d_j$ . Όπως είναι φυσικό, εάν ένας όρος δεν εμφανίζεται στο κείμενο, το  $TF=0$ . Το inverse document frequency (IDF) για τον όρο  $k_i$ ,  $IDF_i = \log \frac{N}{n_i}$ , δηλαδή είναι συνάρτηση του πλήθους όλων των κειμένων και των κειμένων που περιέχουν τον συγκεκριμένο  $k_i$  όρο. Ο στόχος του IDF είναι να μικραίνει τις συντεταγμένες των όρων που εμφανίζονται σε πολλά κείμενα και να μεγαλώσει τις συντεταγμένες των όρων που βρίσκονται σε λίγα, γιατί όλοι οι άξονες του Vector-Based μοντέλο δεν είναι εξίσου σημαντικοί.

Συμπερασματικά, τα σημαντικότερα πλεονεκτήματα του διανυσματικού μοντέλου είναι ότι το σχήμα της ανάθεσης των βαρών βελτιώνει την απόδοση της ανάκτησης. Παράλληλα, η στρατηγική της μερικής ομοιότητας των κειμένων με την επερώτηση επιτρέπει την καλύτερη προσέγγιση των ιδιοτήτων της επερώτησης. Θεωρητικά, όμως το μοντέλο αυτό έχει ένα μειονέκτημα, λόγω του ότι υποθέτει ότι οι όροι είναι αμοιβαία ανεξάρτητοι. Αυτό όμως δεν είναι ξεκάθαρο μειονέκτημα, γιατί πολλές φορές η συνολική απόδοση μειώνεται, όταν δεν εμποδίζεται η εφαρμογή των τοπικών εξαρτήσεων των όρων στα κείμενα της συλλογής.

Τέλος, θα γίνει μια σύντομη αναφορά και σε ένα άλλο μοντέλο IR, το Probabilistic Retrieval μοντέλο [1] [32]. Τα βάρη των όρων στο μοντέλο αυτό είναι όλα δυαδικά. Για παράδειγμα,  $w_{i,j} \in \{0,1\}$ ,  $w_{i,q} \in \{0,1\}$ . Η επερώτηση  $q$  αποτελεί ένα υποσύνολο των όρων. Υποθέτοντας ότι  $R$  είναι το σύνολο των κειμένων που είναι γνωστά (ή αρχικά έχουν υποτεθεί) να είναι σχετικά και το  $R'$  να είναι το συμπληρωματικό του σύνολο, που περιέχει τα μη σχετικά, το  $P(R | \vec{d}_j)$  αντιστοιχεί στην πιθανότητα ότι το κείμενο  $d_j$  είναι σχετικό με την επερώτηση  $q$  και το  $P(R' | \vec{d}_j)$  ορίζει την πιθανότητα το κείμενο να είναι μη σχετικό. Η ομοιότητα  $sim(\vec{d}_j, q)$  του κειμένου  $d_j$  με την επερώτηση  $q$ , ορίζεται ως ο λόγος  $sim(\vec{d}_j, q) = \frac{P(R | \vec{d}_j)}{P(R' | \vec{d}_j)} = \frac{P(\vec{d}_j | R) \times P(R)}{P(\vec{d}_j | R') \times P(R')}$

όπου  $P(\vec{d}_j | R)$  είναι η πιθανότητα να επιλεγεί τυχαία το κείμενο  $d_j$  από το σύνολο των σχετικών κειμένων,  $R$  και  $P(R)$  η πιθανότητα ένα κείμενο να επιλεγεί τυχαία από ολόκληρη την συλλογή που είναι σχετική. Αντίστοιχα ορίζονται και οι υπόλοιπες πιθανότητες. Η παραπάνω επέκταση του τύπου έγινε σύμφωνα με τον κανόνα του Bayes. Επειδή  $P(R) = P(R')$ ,

$$\text{συνεπάγεται } sim(\vec{d}_j, q) \square \frac{P(\vec{d}_j | R)}{P(\vec{d}_j | R')} \square \frac{\left( \prod_{g_i(\vec{d}_j)=1} P(k_i | R) \right) \times \left( \prod_{g_i(\vec{d}_j)=0} P(k_i' | R) \right)}{\left( \prod_{g_i(\vec{d}_j)=1} P(k_i | R') \right) \times \left( \prod_{g_i(\vec{d}_j)=0} P(k_i' | R') \right)}, \quad \text{λόγω της}$$

ανεξαρτησίας των όρων. Η πιθανότητα  $P(k_i | R)$  αντιστοιχεί στην πιθανότητα ο όρος  $k_i$  να βρίσκεται στο κείμενο που θα επιλεγεί τυχαία από το σύνολο  $R$ , αντίστοιχα ορίζονται και οι υπόλοιπες πιθανότητες. Λογαριθμίζοντας, λαμβάνοντας υπόψη την σχέση  $P(k_i | R) + P(k_i' | R) = 1$  και αγνοώντας τους όρους που είναι σταθεροί για όλα τα κείμενα, ο

τύπος της ομοιότητας μια επερώτησης με κάποιο κείμενο, καταλήγει στον ακόλουθο τύπο:

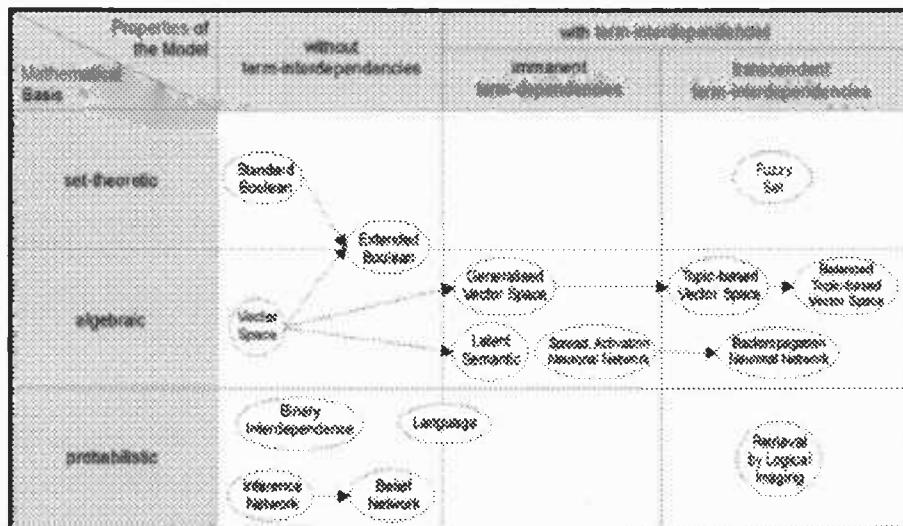
$$sim(\vec{d}_j, q) = \sum_{i=1}^t w_{i,q} \times w_{i,j} \times \left( \log \frac{P(k_i | R)}{(1 - P(k_i | R))} + \log \frac{1 - P(k_i | R')}{P(k_i | R')} \right). \quad \text{Αφού δεν είναι δυνατόν να}$$

είναι γνωστό το σύνολο  $R$ , πρέπει να εφευρεθεί μια μέθοδο για τον υπολογισμό των πιθανοτήτων  $P(k_i | R)$  και  $P(k_i | R')$ . Υπάρχουν πολλές εναλλακτικές για τον υπολογισμό, η πιο

απλή είναι να αντιστοιχηθεί η τιμή 0,5 για την  $P(k_i | R)$  και  $P(k_i | R') = \frac{n_i}{N}$ . Περισσότερες

λεπτομέρειες περιγράφονται στα [1] και [32]. Το κύριο πλεονέκτημα του μοντέλου είναι ότι η κατάταξη των κειμένων γίνεται σύμφωνα με την πιθανότητα τους να είναι σχετικά. Τα μειονεκτήματα του είναι ότι θα πρέπει κανείς να μαντέψει τον αρχικό διαχωρισμό των κειμένων σε σχετικά και μη και ότι η μέθοδος δεν λαμβάνει υπόψη της την συχνότητα των όρων που έχει κάθε κείμενο καθώς και ότι υιοθετεί την υπόθεση ότι οι όροι είναι ανεξάρτητοι, όπως στο διανυσματικό μοντέλο

Ολοκληρώνοντας την υποενότητα αυτή είναι σημαντικό να επισημανθεί ότι υπάρχουν και άλλα εναλλακτικά μοντέλα IR, όπως το επεκταμένο Boolean μοντέλο το μοντέλο του ασαφούς συνόλου, το γενικευμένο διανυσματικό μοντέλο, το μοντέλο των νευρωνικών δικτύων, τα Bayesian δίκτυα κ.α. που δεν γίνεται περεταίρω αναφορά στην συγκεκριμένη εργασία [1] [32] [33]. Η παρακάτω εικόνα, στοχεύει στην κατηγοριοποίηση όλων των υπάρχων IR μοντέλων, σύμφωνα με το μαθηματικό υπόβαθρο και τις ιδιότητες των μοντέλων. (Dominik Kuropka)



## 2.4 Αξιολόγηση των Συστημάτων Ανάκτησης Πληροφορίας

Είναι ευρέως γνωστό ότι τα κοινά μέτρα της απόδοσης συστημάτων είναι ο χρόνος και ο χώρος που καταλαμβάνουν στην μνήμη. Όσο πιο σύντομος είναι ο χρόνος που ανταποκρίνονται τα συστήματα, τόσο πιο μικρός είναι ο χώρος που καταλαμβάνουν και τόσο καλύτερα θεωρούνται τα συστήματα. Όπως, λοιπόν, είναι φυσικό, και στα συστήματα ανάκτησης πληροφοριών κάποιο από τα παραπάνω μέτρα υιοθετούνται για την αξιολόγησή τους.[1][32] Ωστόσο, εξαιρετικό ενδιαφέρον παρουσιάζει, τόσο η απόδοση της δομής του ευρετηρίου, μιας και βοηθάει στην επιτάχυνση της εύρεσης, όσο και η καθυστέρηση που πραγματοποιείται στα κανάλια επικοινωνίας.

Εξαιτίας του ότι τα αποτελέσματα των συστημάτων της ανάκτησης κατατάσσονται

σύμφωνα με την ομοιότητα των κειμένων με την επερώτηση, εκτός από τα παραπάνω μέτρα, ενδιαφέρον παρουσιάζουν κι άλλα όπως, η ακρίβεια του συνόλου των απαντήσεων. Αυτό το μέτρο είναι πολύ σημαντικό, αφού η ερώτηση του χρήστη είναι ασαφής από την φύση της και είναι πολύ δύσκολο να βρεθούν κείμενα που να έχουν την ακριβή απάντηση.

Καταρχήν, αυτό που είναι βασικό να οριστεί για την αξιολόγηση της ανάκτησης είναι το ανακτηθέν αντικείμενο που επιθυμείται να αξιολογηθεί. Για παράδειγμα, το ανακτηθέν αντικείμενο μπορεί να αποτελείται, είτε απλά από μια διαδικασία επεξεργασίας επερωτήσεων σε μια φάση, στην οποία, δηλαδή, ο χρήστης δίνει την επερώτηση και του επιστρέφεται η απάντηση, είτε από ένα αλληλεπιδρών σύστημα, στο οποίο ο χρήστης καθορίζει τις πληροφορίες που επιθυμεί μέσω μιας σειράς βημάτων που αλληλεπιδρούν με το σύστημα, είτε από τον συνδυασμό των δύο παραπάνω στρατηγικών. Στο σύστημα της μιας φάσης, σημαντικό ρόλο παίζει η ποιότητα της απάντησης, ενώ στο αλληλεπιδρών σύστημα απαιτείται έλεγχος και μέτρημα πολλών χαρακτηριστικών, όπως ο σχεδιασμός της διεπαφής, η καθοδήγηση που υπάρχει από το σύστημα και η διάρκεια του.

Οι βασικές μέθοδοι αξιολόγησης της διαδικασίας ανάκτησης είναι το μέτρο της ανάκλησης (Recall) και το μέτρο της ακρίβειας (Precision).

*Recall* είναι ο λόγος των σχετικών κειμένων που έχουν ανακτηθεί προς το σύνολο των σχετικών κειμένων και

*Precision* είναι ο λόγος των ανακτηθέντων κειμένων που είναι σχετικά προς τον αριθμό των ανακτηθέντων κειμένων.

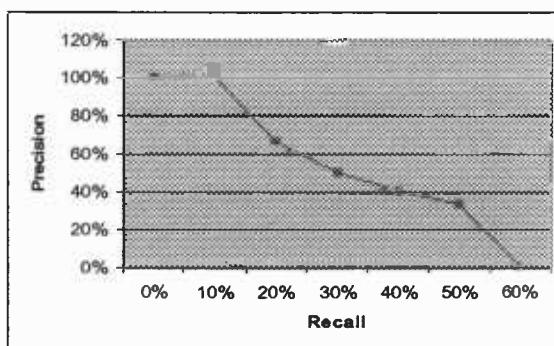
Για την έκφραση των παραπάνω μέτρων, γίνεται η υπόθεση ότι εξετάστηκαν εξαρχής όλα τα κείμενα του απαντητικού συνόλου. Ωστόσο, συνήθως ο χρήστης δεν εξετάζει όλα τα κείμενα του απαντητικού συνόλου αμέσως. Αντίθετα, τα κείμενα που επιστρέφονται πρώτα, ταξινομούνται σύμφωνα με τον βαθμό ομοιότητας τους με την επερώτηση και στην συνέχεια ο χρήστης εξετάζει την ταξινομημένη λίστα των κειμένων, από το κείμενο που βρίσκεται στην αρχή της λίστας. Σύμφωνα με αυτή την διαδικασία, τα μέτρα recall και precision τροποποιούνται τις τιμές τους, καθώς ο χρήστης επεξεργάζεται το σύνολο των απαντήσεων. Παρόλο που για να πραγματοποιηθεί μια τέλεια ανάκτηση, θα πρέπει και τα δύο μέτρα να έχουν σαν τιμή την μονάδα, στην πράξη δεν μπορεί να πραγματοποιηθεί αυτό, αφού αυτά τα δύο μέτρα είναι αντιστρόφως ανάλογα. Το recall βελτιστοποιείται όταν επιστραφεί όλη η συλλογή των κειμένων, ενώ το precision βελτιστοποιείται όταν επιστραφούν πολύ λίγα κείμενα

Είναι σημαντικό να επισημανθεί ότι η σωστή αξιολόγηση απαιτεί την αντιπαραβολή (trade-off) μεταξύ του βαθμού ανάκλησης και του βαθμού ακρίβειας, μέσω της δημιουργίας της καμπύλης precision-recall. Για να γίνει κατανοητή η διαδικασία της δημιουργίας της παραπάνω καμπύλης παρουσιάζεται το παρακάτω παράδειγμα. Υποθέτοντας την ύπαρξη μια συλλογής και ενός συνόλου παραδειγμάτων πληροφορίας που ζητούνται, καταρχήν, η προσοχή εστιάζεται στον σχηματισμό της επερώτησης που αντιπροσωπεύει την πληροφορία που απαιτείται κάθε φορά. Υποθέτοντας ότι διατίθεται και το σύνολο των κειμένων που είναι σχετικά με την επερώτηση και μάλιστα χωρίς να χαθεί η γενικότητα υποτίθεται ότι το σύνολο αυτό (Rq) αποτελείται από τα εξής δέκα κείμενα: {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123} μπορεί να αρχίσει η εκτέλεση ενός νέου αλγορίθμου ανάκτησης πληροφορίας και να γίνει η σύγκριση των αποτελεσμάτων τους. Θεωρώντας ότι η απάντηση του αλγορίθμου είναι το ακόλουθο ταξινομημένο απαντητικό σύνολο:

1.d123 (x), 2.d84, 3.d56 (x), 4.d6, 5.d8, 6.d9 (x), 7.d511, 8.d129, 9.d187, 10.d25 (x), 11.d38, 12.d48, 13.d250, 14.d113 και 15.d3 (x)

και σημειώνοντας τα κείμενα που είναι σχετικά με την επερώτηση με το σύμβολο

εκτελούνται οι ακόλουθες παρατηρήσεις, αρχίζοντας την εξέταση της ταξινομημένης λίστας από το πρώτο κείμενο: Καταρχήν, το κείμενο d123 είναι το πρώτο σχετικό κείμενο και αντιπροσωπεύει το 10% όλων των σχετικών κειμένων στο σύνολο Rq. Άρα, η τιμή του precision είναι 100% (1 κείμενο στα 1 είναι σχετικά), όταν η τιμή του recall είναι 10% (1 στα 10 κείμενα που είναι σχετικά έχουν εξεταστεί). Το κείμενο d56 που κατατάσσεται στην τρίτη θέση των σχετικών κειμένων τροποποιεί την τιμή του precision στο 66% (2 κείμενα στα 3 είναι σχετικά), όταν η τιμή του recall είναι 20% (2 στα 10 κείμενα που είναι σχετικά έχουν εξεταστεί). Εξετάζοντας και τα υπόλοιπα κείμενα η καμπύλη precision-recall που προκύπτει είναι η ακόλουθη:

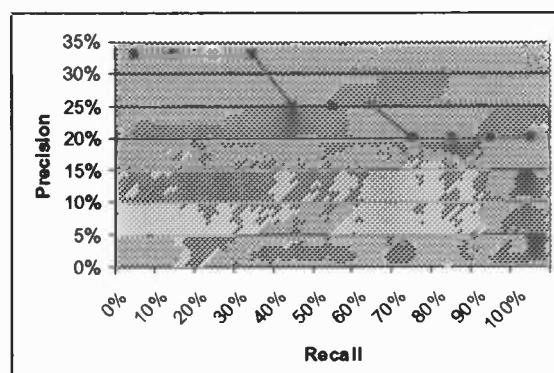


Η ακρίβεια (precision) σε επίπεδα που το recall είναι μεγαλύτερο από 50% πέφτει στο μηδέν, γιατί δεν έχουν ανακτηθεί όλα τα κείμενα. Η καμπύλη precision-recall βασίζεται συνήθως σε 11 (αντί για 10) συγκεκριμένα επίπεδα του recall, που είναι το 0%, 10%, 20%, ..., 100%. Για το επίπεδο 0% η τιμή του precision καθορίζεται με βάση της διαδικασίας της παραβολής.

Με σκοπό την περιγραφή της διαδικασίας της παρεμβολής χρησιμοποιούται το παραπάνω παράδειγμα με την διαφορά ότι αυτή την φορά το Rq σύνολο είναι το {d3, d56, d129}. Το πρώτο σχετικό με την επερώτηση κείμενο είναι το d56 και η τιμή του recall του αντιστοιχεί στο 33,3%. Το δεύτερο σχετικό κείμενο είναι το d129 και η τιμή του recall του αντιστοιχεί στο 66,6%. Τέλος, το κείμενο d3 είναι το τρίτο σχετικό κείμενο και το επίπεδο recall που του αντιστοιχεί είναι το 100%.

Θέτοντας το  $r_j$ ,  $j \in \{0, 1, 2, \dots, 10\}$  να είναι μια αναφορά στο  $j$ -οστό επίπεδο recall, ο κανόνας της παρεμβολής ισχυρίζεται ότι η τιμή του precision που προκύπτει από την παρεμβολή στο  $j$ -οστό επίπεδο recall είναι η μεγαλύτερη γνωστή τιμή precision σε κάθε recall επίπεδο ανάμεσα στο  $j$ -οστό και  $(j+1)$ -οστό.

Δηλαδή, στο παραπάνω παράδειγμα προκύπτει η ακόλουθη καμπύλη



Άλλα μέτρα που εξετάζουν την απόδοση της ανάκτησης είναι ο αρμονικός μέσος (Harmonic mean). Είναι ένα απλό μέτρο που συνδυάζει το recall και το precision και υπολογίζεται για κάθε j-οστό κείμενο της κατάταξης ως

$$\text{Harmonic mean} = \frac{2}{\frac{1}{\text{recall}(j)} + \frac{1}{\text{precision}(j)}}$$

Οι τιμές του είναι ανάμεσα στο διάστημα [0,1]. Αν είναι μηδέν τα κείμενα που ανακτήθηκαν δεν είναι σχετικά ενώ εάν είναι μονάδα είναι όλα σχετικά.

Ένα άλλο μέτρο που κι αυτό συνδυάζει το recall και το precision είναι το E και προτάθηκε από τον Rijksbergen. Η ιδέα που στηρίζεται είναι η ευελιξία του χρήστη να καθορίσει το μέγεθος του ενδιαφέροντος για το μέτρο του recall ή του precision. Ο ορισμό του μέτρου αυτού είναι

$$E = 1 - \frac{1+b^2}{\frac{b^2}{\text{recall}(j)} + \frac{1}{\text{precision}(j)}}$$

Εάν οι τιμές του b είναι μικρότερες ή μεγαλύτερες από την μονάδα καθορίζουν το πόσο ενδιαφέρεται ο χρήστης για το recall και το precision αντίστοιχα.

Μέχρι τώρα έγινε αναφορά σε μέτρα που το σύνολο των ανακτηθέντων κειμένων για την επερώτηση είναι ίδια, ανεξάρτητα από τον χρήστη. Όμως, οι διαφορετικοί χρήστες μπορεί να ερμηνεύουν διαφορετικά την ομοιότητα των κειμένων με την επερώτηση. Έτσι, στην συνέχεια αναφέρονται κάποια μέτρα που είναι προσανατολισμένα στον χρήστη.

Το coverage ratio (λόγος κάλυψης) ορίζεται ως ο λόγος των σχετικών κειμένων που αναγνωρίζει ο χρήστης και που ανακτήθηκαν προς τα σχετικά κείμενα που ο χρήστης αναγνωρίζει. Μια μεγάλη τιμή στο μέτρο αυτό δείχνει ότι το σύστημα βρίσκει τα περισσότερα κείμενα που ο χρήστης περιμένει να δει.

Το novelty (καινοτομία) ορίζεται ως ο λόγος των κειμένων που είναι σχετικά, αλλά ο χρήστης δεν τα αναγνωρίζει, προς όλα τα σχετικά κείμενα. Μια μεγάλη τιμή στον λόγο αυτό δείχνει ότι το σύστημα ανακάλυψε πολλά σχετικά κείμενα που προηγουμένως δεν τα ήξερε.

Όμοια ορίζεται το relative recall (σχετική ανάκληση), ως ο λόγος ανάμεσα στο πλήθος των σχετικών κειμένων που βρίσκονται από το σύστημα και το πλήθος των σχετικών κείμενων που ο χρήστης περίμενε να βρει και το recall effort που είναι ο λόγος ανάμεσα στο πλήθος των σχετικών κειμένων που ο χρήστης περίμενε να βρει και το πλήθος των κειμένων που εξετάστηκαν στην προσπάθεια να εντοπιστούν τα αναμενόμενα σχετικά κείμενα.

Άλλα μέτρα που μπορεί να έχουν ενδιαφέρουν είναι το αναμενόμενο μήκος της αναζήτησης (expected search length) που μπορεί να διαπραγματευτεί σύνολα κειμένων που είναι αδύνατον να καταταχτούν, το μέτρο της ικανοποίησης (satisfaction) που λαμβάνει υπόψη του μόνο τα σχετικά κείμενα και το μέτρο της απογοήτευσης (frustration) που λαμβάνει υπόψη του μόνο τα μη σχετικά κείμενα.

### 3. Επερωτήσεις

#### 3.1 Περιγραφή των τρόπων σχηματισμού των επερωτήσεων

Σε αυτή την υποενότητα γίνεται μια σύντομη αναφορά στα διαφορετικά είδη επερωτήσεων που προτείνονται τα συστήματα ανάκτησης κειμένων [1]. Τα είδη αυτά είναι μερικώς εξαρτώμενα από το μοντέλο που υιοθετούν τα συστήματα. Για παράδειγμα, τα συστήματα που χρησιμοποιούν το μοντέλο που κρατάει ολόκληρα τα κείμενα, δεν μπορεί να απαντά τις ίδιες επερωτήσεις με τα συστήματα που βασίζονται στην κατάταξη των λέξεων-κλειδιών(keywords).

Όπως είναι φυσικό, ανάλογα με το διαχωρισμό της ανακτηθείσας πληροφορίας και των ανακτηθέντων δεδομένων, διαχωρίζονται και οι διαφορετικές γλώσσες που χρησιμοποιούνται για τον σχηματισμό των επερωτήσεων. Γλώσσες που επιτρέπουν την κατάταξη των αποτελεσμάτων τους ανήκουν στις γλώσσες που προορίζονται για την ανάκτηση πληροφορίας. Αντίθετα, οι γλώσσες, που δεν μπορούν εύκολα να κατατάξουν τα αποτελέσματα τους, ανήκουν στις γλώσσες που είναι κατάλληλες για την ανάκτηση δεδομένων. Επιπρόσθετα, υπάρχουν και κάποιες γλώσσες, που δεν απευθύνονται στους τελικούς χρήστες, αλλά σε προγραμματιστικά πακέτα υψηλότερου επιπέδου, που επικοινωνούν με on-line βάσεις ή CD-ROM. Σε αυτήν την περίπτωση δεν μιλάμε για απλές γλώσσες, αλλά για πρωτόκολλα. Είναι προφανές ότι ανάλογα με την εμπειρία του χρήστη μπορεί να χρησιμοποιηθεί διαφορετική γλώσσα για την επερώτηση. Για παράδειγμα, εάν ο χρήστης ξέρει ακριβώς τι θέλει, η εργασία της ανάκτησης είναι ευκολότερη και μπορεί να μην χρειαστεί ούτε καν η κατάταξη των αποτελεσμάτων.

Ένα σημαντικό θέμα είναι ότι οι γλώσσες που χρησιμοποιούνται για το σχηματισμό των επερωτήσεων προσπαθούν να χρησιμοποιήσουν το περιεχόμενο (π.χ. την σημασιολογία) και τη δομή των κειμένων (π.χ. την σύνταξη) για να βρουν τα σχετικά κείμενα. Η νοοτροπία αυτή δεν βιοθάει πάντα στην ανάκτηση των σχετικών απαντήσεων, όποτε κρίνεται απαραίτητο η χρησιμοποίηση κάποιων τεχνικών, ώστε να προάγουν την χρησιμότητα των υπαρχουσών επερωτήσεων. Χαρακτηριστικά παραδείγματα είναι η επέκταση λέξεων, είτε από το σύνολο των συνωνύμων τους, είτε με τη χρήση θησαυρών που διαθέτουν λίστες σημαντικών λέξεων και για κάθε λέξη της παραπάνω λίστας, διθέντος ενός πεδίο γνώσης, κρατάει ένα σύνολο σχετικών λέξεων (π.χ. λέξεις που πηγάζουν από τις συνωνυμικές σχέσεις), η αποκοπή καταλήξεων, ώστε να τοποθετούνται όλα μαζί τα παράγωγα της ίδιας λέξης (π.χ. οι λέξεις connected, connecting, connection και connections έχουν όλες σαν λεξιλογική ρίζα την λέξη connect) και η απαλοιφή των τετριμένων λέξεων, που διαθέτουν πολύ μικρές τιμές διάκρισης για τους σκοπούς τις ανάκτησης(π.χ. the, a κ.τ.λ.).

Όπως έχει αναφερθεί και παραπάνω, η επερώτηση είναι ο σχηματισμός που αποδίδει την πληροφορία που χρειάζεται ο χρήστης. Η απλούστερη μορφή της αποτελείται από λέξεις-κλειδιά και είναι προφανές ότι και τα κείμενα πρόκειται να περιέχουν λέξεις σαν κι αυτές που αναζητούνται. Η μορφή αυτή είναι ιδιαίτερα δημοφιλής, διότι είναι διαισθητική, εκφράζεται εύκολα και επιτρέπει τη γρήγορη κατάταξη των αποτελεσμάτων. Μια επερώτηση τέτοιας μορφής μπορεί να είναι μια απλή λέξη, ή μπορεί να είναι ένας πολύπλοκός συνδυασμός πράξεων με λέξεις.

Πολλά συστήματα υλοποιούν τις επερωτήσεις μιας λέξης, με δυνατότητα να αναζητούν λέξεις, σε ένα διθέν περιβάλλον, που είναι κοντά σε άλλες λέξεις. Λέξεις που μπορεί να εμφανίζονται κοντά σε άλλες, ίσως προσδίδουν μια μεγαλύτερη πιθανότητα ομοιότητας από όταν εμφανίζονται ξεχωριστά. Για παράδειγμα, είναι δυνατόν να σχηματίζονται επερωτήσεις από φράσεις λέξεων, δηλαδή μια σειρά από επερωτήσεις μιας λέξης. Χαρακτηριστικό είναι το παράδειγμα της αναζήτησης λέξεων, όπως το enhance και το retrieval, που αναγνωρίζονται οι

επερωτήσεις από φράσεις ότι οι διαχωριστές των λέξεων δεν είναι πάντα το κενό, όπως στις επερωτήσεις, αλλά και λέξεις χωρίς σημασία, δεν θα πρέπει να αποπροσανατολίζουν την ανάκτηση και θα πρέπει να είναι δυνατόν να θεωρήσουν σχετικά κείμενα, όπως το «...enhance the retrieval...»). Παράλληλα, σχηματίζονται επερωτήσεις που αναζητούν λέξεις που είναι κοντινές σε ένα κείμενο, δηλαδή δίνεται μια σειρά από επερωτήσεις μιας λέξης ή και φράσεων, μαζί με την μεγαλύτερη επιτρεπτή απόσταση μεταξύ τους. Σύμφωνα, δηλαδή, με το παραπάνω παράδειγμα, με μέγιστη απόσταση το 4, πρόκειται να αναζητούν κείμενα, όπως «...enhance the power of retrieval...» ).

Η παλαιότερη μορφή για το συνδυασμό επερωτήσεων με λέξεις-κλειδιά είναι η χρήση των πράξεων Boolean. Μια τέτοιου είδους ερώτησης έχει ένα συντακτικό που αποτελείται από «άτομα» (π.χ. βασικές επερωτήσεις) και Boolean πράξεις (or, and, but) που εκτελούνται πάνω στους τελεστές τους (που είναι σύνολα κειμένων) και λαμβάνουν σύνολα κειμένων. Η κατάταξη των αποτελεσμάτων δεν παρέχεται, έτσι ή ένα κείμενο ικανοποιεί την επερώτησης ή όχι. Αυτό το πρόβλημα μπορεί να ξεπεραστεί, εάν οι ιδιότητες της ανάκτησης «χαλαρώσουν». Ένας τρόπος είναι να υιοθετηθούν οι «fuzzy Boolean» πράξεις. Η ιδέα που στηρίζονται οι πράξεις αυτές είναι στη διαφορετική ερμηνεία που προσδίδουν στον τελεστή or και and. Για παράδειγμα αντί ένα στοιχείο να εμφανίζεται σε όλους τους τελεστές (and), τα στοιχεία, που αναζητήθηκαν μπορεί να εμφανίζονται μόνο σε μερικούς τελεστές. Είναι δυνατόν με αυτούς να οριστεί και η κατάταξη των κειμένων, σύμφωνα με το πλήθος των στοιχείων που είναι κοινά με την επερώτηση.

### 3.2 Παρουσίαση πράξεων και λειτουργιών των επερωτήσεων

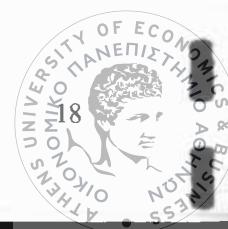
Εξαιτίας του ότι οι χρήστες δυσκολεύονται πολύ να σχηματίσουν επερωτήσεις που να αντιπροσωπεύουν τους σκοπούς της ανάκτησης, συνήθως επιβάλλεται ο μετασχηματισμός της. Αυτός ο μετασχηματισμός αποτελείται από δύο βασικά βήματα : (1) την επέκταση της αρχικής επερώτησης με νέους όρους και (2) με την τροποποίηση του βάρους των όρων της επεκταμένης επερώτησης. [1] [32]

Υπάρχουν τρεις κατηγορίες για την βελτίωση της αρχικής επερώτησης μέσω της επέκτασης και της τροποποίησης του βάρους. Η πρώτη αφορά την ανάδραση (feedback) της πληροφορίας από το χρήστη, η δεύτερη βασίζεται στην πληροφορία που πηγάζει από το σύνολο των κειμένων που αναζητήθηκαν με την εκτέλεση της αρχικής επερώτησης και η τελευταία στην συνολική πληροφορία που απορρέει από ολόκληρη την συλλογή των κειμένων.

#### Κύκλος της Ανάδρασης (relevant feedback)

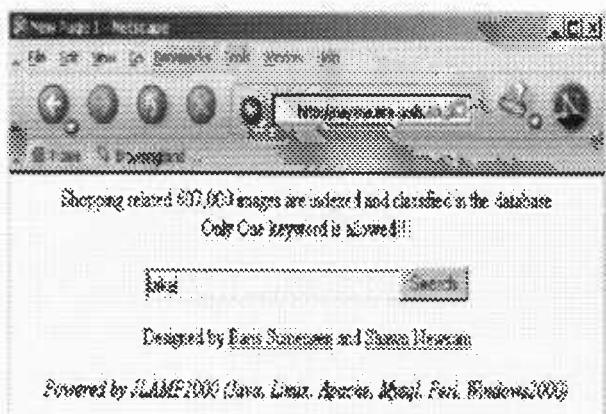
Η κύρια ιδέα στον κύκλο της ανάδρασης (relevant feedback) είναι η σύνδεση του χρήστη με την διαδικασία της ανάκτησης. Η βασική διαδικασία είναι :

- Ο χρήστης εκτελεί μια επερώτηση
- Το σύστημα επιστρέφει ένα αρχικό σύνολο ανακτηθέντων αποτελεσμάτων
- Ο χρήστης προσδιορίζει από το σύνολο των ανακτηθέντων κειμένων ποια είναι σχετικά με την επερώτηση και ποια μη σχετικά
- Το σύστημα προσπαθεί με βάση τις καινούριες πληροφορίες να υπολογίσει μια καλύτερη αναπαράσταση της πληροφορίας που ζητείται
- Το σύστημα εμφανίζει ένα νέο σύνολο ανακτηθέντων αποτελεσμάτων



Η παραπάνω διαδικασία μπορεί να εκτελεστεί μια ή περισσότερες φορές. Όπως είναι φανερό, ο κύκλος της ανάδρασης εκμεταλλεύεται την ιδέα ότι είναι πολύ δύσκολο να σχηματίσει κάνεις μια καλή επερώτηση, όταν δεν γνωρίζει καλά την συλλογή, αλλά είναι εύκολο να κρίνει τα αποτελέσματα (κείμενα) που ανακτώνται ψηλά στην κατάταξη. Σημειώνοντας, στην συνέχεια ποια είναι πραγματικά σχετικά με την αρχική επερώτηση μπορεί να προσδιορίσει τις πληροφορίες που επιζητά. Στόχος της λοιπόν είναι να επιλεγούν σημαντικοί όροι ή εκφράσεις από τα κείμενα που έχουν χαρακτηριστεί ως σχετικά και να βοηθήσουν στην ενίσχυση της σημαντικότητας των όρων στο νέο σχηματισμό της επερώτησης. Το αναμενόμενο αποτέλεσμα είναι η τροποποίηση της επερώτησης να αναδείξει τα σχετικά κείμενα και να απομακρύνει τα μη σχετικά.

Ένα χαρακτηριστικό παράδειγμα είναι η αναζήτηση εικόνων, αφού ο χρήστης δεν μπορεί εύκολα να σχηματίσει μια επερώτηση για το θέμα που επιζητά, όμως μπορεί να προσδιορίσει ποιες εικόνες είναι σχετικές και ποιες όχι. Οι παρακάτω εικόνες παρουσιάζουν το κύκλο της ανάδρασης ενός χρήστη που αναζητά εικόνες «ποδήλατων» στο σύστημα: <http://nayana.ece.ucsb.edu/imsearch/imsearch.html> (Newsam et al. 2001)



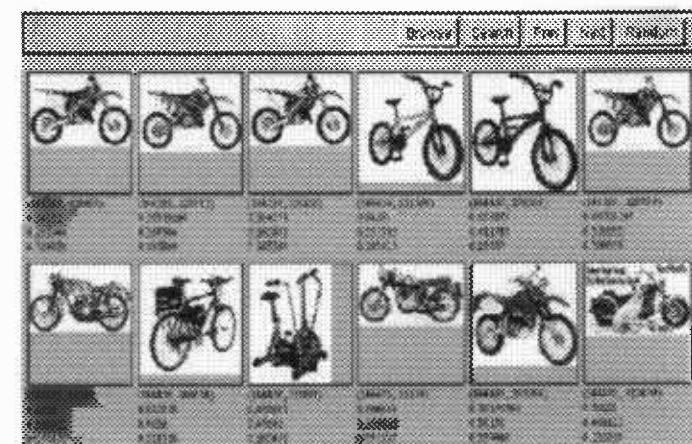
**Εικόνα 1: Ο χρήστης εισάγει την αρχική επερώτηση στο σύστημα**



**Εικόνα 2: Ο χρήστης λαμβάνει τα αποτελέσματα της αρχικής επερώτησης**



**Εικόνα 3: Ο χρήστης σημειώνει τις σχετικές εικόνες**



**Εικόνα 4: Ο χρήστης λαμβάνει το καινούριο σύνολο αποτελεσμάτων**

Παρακάτω παρουσιάζεται ακόμα ένα παράδειγμα βασισμένο αυτή την φορά σε αναζήτηση κειμένων, αντί για εικόνες. Ο χρήστης, όπως φαίνεται και στη παρακάτω εικόνα, επιζητά πληροφορίες για νέους χώρους διαστημικών εφαρμογών. Αφού, επιστραφούν από το

σύστημα τα σχετικά κείμενα, σημειώνονται με (+) τα κείμενα που είναι όντως σχετικά, σύμφωνα με τον χρήστη. Στην συνέχεια, η αρχική επερώτηση επεκτείνεται σύμφωνα με τους 18 όρους που εμφανίζονται παρακάτω μαζί με τα βάρη τους. Τέλος, εμφανίζονται τα αποτελέσματα που προκύπτουν από την επέκταση της επερώτησης και σημειώνονται με (\*) τα κείμενα που είναι σχετικά.

**Query: New space satellite applications**

- + 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
- + 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
- 3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
- 4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
- 5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
- 6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
- 7. 0.516, 04/13/87, ArianeSpace Receives Satellite Launch Pact From Telesat Canada
- + 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

2.074 new 15.106 space  
 30.816 satellite 5.660 application  
 5.991 nasa 5.196 eos  
 4.196 launch 3.972 aster  
 3.516 instrument 3.446 arianeSpace  
 3.004 bundespost 2.806 ss  
 2.790 rocket 2.053 scientist  
 2.003 broadcast 1.172 earth  
 0.836 oil 0.646 measure

- \* 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
- \* 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
- 3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
- 4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
- \* 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies
- 6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
- 7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
- 8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost \$90 Million

Είναι σημαντικό να τονιστεί ότι η επιτυχία του κύκλου της ανάδρασης εξαρτάται άμεσα από κάποιες προϋποθέσεις. Αρχικά, ο χρήστης θα πρέπει να έχει την απαραίτητη γνώση, ώστε να σχηματίσει μια επερώτηση που θα είναι σημασιολογικά κοντά στα κείμενα που επιθυμεί να ανακτήσει. Οι περιπτώσεις, που ο κύκλος της ανάδρασης δεν προσφέρει ικανοποιητικά αποτελέσματα παρατηρούνται όταν, οι όροι χρησιμοποιούνται ή προφέρονται με διαφορετικό τρόπο από ότι χρησιμοποιούνται ή προφέρονται στα κείμενα της συλλογής. Παράλληλα, η ανάδραση απαιτεί η διανομή των όρων στα κείμενα που είναι σχετικά, να είναι ίδια με των κειμένων, που ο χρήστης χαρακτηρίζει ως σχετικά, ενώ η διανομή των όρων στα μη σχετικά να είναι διαφορετική από αυτή των σχετικών. Τέλος, η ανάδραση έχει εκτός των άλλων και κάποια τεχνικά προβλήματα, όπως μια πολύ μεγάλη ερώτηση σε μήκος δεν είναι αποτελεσματική σε ένα κλασσικό σύστημα IR, αλλά και ότι ο χρήστης δεν είναι πάντα πρόθυμος να συμμετάσχει στον κύκλο της ανάδρασης.

Στην περίπτωση λοιπόν που το σύνολο με τα σχετικά κείμενα είναι πολύ μεγάλο και δεν είναι δυνατόν η ανάδραση να γίνει με την βοήθεια του χρήστη, εκτελείται αυτόματα και η τροποποίηση της επερώτησης πραγματοποιείται περιλαμβάνοντας αναγνωριστικούς όρους που είναι σχετικοί με τους όρους της επερώτησης. Αυτοί οι αναγνωριστικοί όροι μπορεί να είναι



συνώνυμα, που να προέρχονται από θησαυρούς λέξεων, λέξεις από stemming ή λέξεις που ήταν κοντά στους όρους της επερώτησης στα κείμενα της συλλογής. Είναι σημαντικό να αναφερθεί ότι υπάρχουν δύο στρατηγικές που μπορούν να υιοθετηθούν: οι τοπικές (local) και οι συνολικές (global).

### Τοπική στρατηγική (Local Methods)

Όταν ακολουθείται η τοπική στρατηγική, τα κείμενα που έχουν ανακτηθεί, εξετάζονται για να καθορίσουν τους όρους για την επέκταση της επερώτησης, παραδείγματος χάριν επιλογή των 10 πιο συχνών εμφανιζόμενων λέξεων των κορυφαίων 5 εγγράφων. Στην περίπτωση που αυτή η διαδικασία γίνεται χωρίς την παρέμβαση του χρήστη, δύο στρατηγικές μπορούν να υιοθετηθούν για την πραγματοποίησή της είτε η τοπική ομαδοποίηση (local clustering) είτε η τοπική ανάλυση των συμφραζόμενων (local context analysis).[1] [32]

Υιοθετώντας την τοπική ομαδοποίηση, ιδιαίτερη έμφαση δίνεται στο σύνολο των κειμένων που ανακτήθηκαν από την αρχική ερώτηση και συγκεκριμένα στα κείμενα που κατατάχθηκαν πιο ψηλά, ώστε να ομαδοποιηθούν οι γειτονικοί τους όροι. Αυτές οι ομαδοποιήσεις στηρίζονται στους όρους που συνυπάρχουν μέσα στο κείμενο. Οι όροι που αποδεικνύονται ότι περιγράφουν καλύτερα κάθε όρο της επερώτησης χρησιμοποιούνται για την επέκταση της επερώτησης.

Όσο αφορά την local context analysis είναι ένα συνδυασμός της τοπικής και της συνολικής ανάλυσης (αναζήτηση για συσχετίσεις όρων σε ολόκληρη την συλλογή). Αυτή η προσέγγιση βασίζεται στην χρήση ομάδων ουσιαστικών (π.χ. ενός ουσιαστικού, δύων ή τριών γειτονικών ουσιαστικών των κειμένων), αντί στην χρήση απλών λέξεων-κλειδιών, όπως συμβαίνει στον ορισμό της έννοιας των κειμένων (documents concepts). Για την επέκταση της επερώτησης, οι έννοιες που επιλέγονται από τα κείμενα που είναι ψηλά στην κατάταξη (όπως γίνεται στην τοπική ανάλυση) στηρίζονται στην συνύπαρξη (co-occurrence) τους με τους όρους της επερώτησης και όχι σε λέξεις που προέρχονται από stemming, όπως παραπάνω. Ωστόσο, αντί για ολόκληρα τα κείμενα, χρησιμοποιούνται τμήματα των κειμένων για να καθορίσουν το είδος της συνύπαρξης.

### Γενική Στρατηγική (Global Methods)

Οι διαδικασίες global analysis χρησιμοποιούν πληροφορία από ολόκληρο το σύνολο των κειμένων της συλλογής. Παραδείγματος χάριν, υπολογισμός πινάκων συσχέτισης (association matrices) που ποσοτικοποιούν την ομοιότητα μεταξύ των όρων ανάλογα με το πόσο συχνά συνυπάρχουν. Αυτή η προσέγγιση προσδιορίζεται από δύο διαφορετικές στρατηγικές που στηρίζονται στην δημιουργία θησαυρού: (1) Επέκταση επερώτησης βασισμένη στην ομοιότητα του θησαυρού (similarity thesaurus) και (2) Επέκταση επερώτησης βασισμένη σε στατιστικό θησαυρό (statistical thesaurus).

Στην πρώτη στρατηγική θα πρέπει να τονιστεί ότι ο θησαυρός ομοιότητας (similarity thesaurus) βασίζεται σε σχέσεις μεταξύ όρων και όχι σε έναν πίνακα συνύρταξης, όπως στην περίπτωση της τοπικής στρατηγικής. Οι σχέσεις αυτές προέρχονται από την παρατήρηση, ότι οι όροι είναι έννοιες που βρίσκονται μέσα σε ένα εννοιολογικό χώρο (concept space). Σε αυτόν τον εννοιολογικό χώρο, κάθε όρος συντάσσεται σε ένα ευρετήριο, μέσω του κειμένου που εμφανίζεται. Έτσι, οι όροι αναλαμβάνουν τον αρχικό ρόλο των κειμένων, ενώ τα κείμενα ερμηνεύονται ως στοιχεία του ευρετηρίου. Δοθέντος του θησαυρού ομοιότητας, τα βήματα για την επέκταση της επερώτησης είναι τα ακόλουθα: (α) η επερώτηση αναπαρίσταται στο εννοιολογικό χώρο που έχουν αναπαρασταθεί οι όροι του ευρετηρίου (β) υπολογίζεται η ομοιότητα μεταξύ των όρων συσχετιζόμενοι, τόσο με ολόκληρη την επερώτηση, όσο

μεμονωμένα με όρους της επερώτησης, σύμφωνα με τον θησαυρό και (γ) επεκτείνεται η επερώτησης με τους όρους που είναι πιο ψηλά στην κατάταξη σύμφωνα με την ομοιότητα που υπολογίστηκε στο (β) βήμα.

Στην δεύτερη στρατηγική, ο θησαυρός αποτελείται από τάξεις που ομαδοποιούν τους όρους που συσχετίζονται με το περιβάλλον ολόκληρης της συλλογής. Αυτοί οι όροι, στην συνέχεια, μπορούν να χρησιμοποιηθούν για την επέκταση της επερώτησης. Όσο υψηλές είναι οι τιμές διαφοροποίησης των όρων, δηλαδή έχουν μικρότερη συχνότητα, τόσο πιο αποτελεσματικοί θα είναι. Εξαιτίας της δυσκολίας να ομαδοποιηθούν οι όροι με την μικρότερη συχνότητα, λόγω της μικρής πληροφορίας που κατέχουν, ομαδοποιούνται τα κείμενα σε τάξεις. Παράλληλα, χρησιμοποιούν την μικρή συχνότητα των όρων στα συγκεκριμένα κείμενα, για να ορίσουν τις τάξεις του θησαυρού, με άμεσο αποτέλεσμα την δημιουργία μικρών και πυκνών ομάδων. Αναλυτικά η διαδικασία που ακολουθείται είναι: (1) τοποθέτηση κάθε κείμενου σε μια διακριτή ομάδα (2) υπολογισμός της ομοιότητας μεταξύ όλων των δυνατών ζευγαριών των ομάδων (3) καθορισμός του ζευγαριού των ομάδων με την μεγαλύτερη εσωτερική ομοιότητα (4) συγχώνευση ομάδων (5) επανάληψη της διαδικασίας από το βήμα (2) έως να ικανοποιηθεί η συνθήκη τερματισμού και (6) επιστροφή των ιεραρχημένων ομάδων. Έχοντας, τώρα, τις τάξεις του θησαυρού μπορούν να χρησιμοποιηθούν για την επέκταση της επερώτησης.

Τέλος, παρουσιάζεται ένα παράδειγμα επέκτασης της αρχικής επερώτησης με βάση το θησαυρό PubMed:

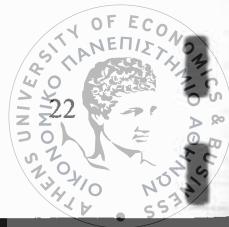
- User query: cancer
- PubMed query: ("neoplasms"[TIAB] NOT Medline[SB]) OR "neoplasms"[MeSH Terms] OR cancer[Text Word]
- User query: skin itch
- PubMed query: ("skin"[MeSH Terms] OR "integumentary system"[TIAB] NOT Medline[SB]) OR "integumentary system"[MeSH Terms] OR skin[Text Word]) AND {"pruritus"[TIAB] NOT Medline[SB]} OR "pruritus"[MeSH Terms] OR itch[Text Word])

## 4. Word Sense Disambiguation

### 4.1 Εισαγωγή

Ένα από τα πρώτα προβλήματα που είχαν να αντιμετωπίσουν τα συστήματα επεξεργασίας φυσικής γλώσσας είναι η διφορούμενη έννοια των λέξεων, που μπορεί να είναι είτε συντακτική, είτε σημασιολογική. Η λύση για την συντακτική αμφιβολία (syntactic ambiguity) των λέξεων έχει προσεγγισθεί με την επεξεργασία της γλώσσας μέσω ετικετών, που προσδιορίζουν το μέρος του λόγου που είναι η κάθε λέξη, προβλέποντας την συντακτική κατηγορία των λέξεων στο κείμενο με μεγάλη ακρίβεια. Το πρόβλημα της λύσης της σημασιολογικής αμφιβολίας είναι γνωστό ως word sense disambiguation και έχει αποδειχτεί ότι είναι πολύ πιο δύσκολο πρόβλημα από την συντακτική βεβαιότητα (syntactic disambiguation).

Πιο συγκεκριμένα, το πρόβλημα είναι ότι συχνά οι λέξεις έχουν περισσότερες από μια ερμηνείες, κάποιες φορές αρκετά όμοιες και μερικές φορές εντελώς διαφορετικές. Η ερμηνεία της κάθε λέξης σε μια συγκεκριμένη χρήση της μπορεί να καθοριστεί μόνο από την εξέταση του περιβάλλοντος της. Αυτό μπορεί να είναι τετριμμένο για το σύστημα επεξεργασίας της γλώσσας του ανθρώπου, όμως για τον υπολογιστή έχει αποδειχτεί ότι είναι πολύ δύσκολο και ίσως να αποτελεί ένα δυσεπίλυτο πρόβλημα.



Αυτή την απαισιόδοξη συμπεριφορά, ότι δηλαδή δεν μπορεί να λυθεί η σημασιολογική αμφιβολία από τον ηλεκτρονικό υπολογιστή, είχε ο σκεπτικιστής Bar-Hillel [7]. Τον παραπάνω ισχυρισμό τον απέδειξε με την χρήση του παραδείγματος που ακολουθεί:

*Little John was looking for his toy box.  
Finally he found it.  
The box was in the pen.  
John was very happy.*

που περιείχε την λέξη pen, που έχει πολλές σημασιολογικές ερμηνείες. Χρησιμοποιώντας μόνο δύο από τις ερμηνείες του pen, την “writing implement” και την “enclosure” κατέληξε στο ότι δεν υπάρχει τρόπος από τον υπολογιστή να αποφασίσει ανάμεσα στις δύο ερμηνείες.

Ωστόσο, η κατάσταση δεν είναι τόσο άσχημη, όσο υποστήριξε, αφού υπάρχουν αρκετά πλεονεκτήματα στο word sense disambiguation και η σημασιολογική αμφιβολία στα κείμενα μπορεί να λυθεί σε ένα ικανοποιητικό βαθμό ακρίβειας.

## 4.2 Η χρησιμότητα του Word Sense Disambiguation

Κατά την επεξεργασία της γλώσσας διακρίνονται οι τελικές και οι ενδιάμεσες εργασίες. Οι τελικές εργασίες αποτελούνται από αυτές που όταν εκτελούνται έχουν την δική τους χρησιμότητα. Σημαντικά παραδείγματα αυτών είναι οι μηχανές μετάφρασης, τα συστήματα αυτόματης περιληψης και εξαγωγής πληροφοριών. Ενώ οι ενδιάμεσες εργασίες είναι αυτές που όταν εκτελεστούν βοηθούν τις τελικές εργασίες, παραδείγματα τέτοιων εργασιών είναι η δημιουργία ετικετών που δείχνουν το μέρος του λόγου, η ανάλυση της γραμματικής των λέξεων των κειμένων και το word sense disambiguation. Παραδοσιακά, το word sense disambiguation μπορεί να φανεί πολύ χρήσιμο σε εργασίες, όπως είναι η ανάκτηση πληροφοριών και οι μηχανές μετάφρασης.

### Ανάκτηση Πληροφοριών

Το word sense disambiguation αποτελεί ένα βασικό τμήμα της ανάκτησης πληροφορίας. Εάν ένα σύστημα ανάκτησης δημιουργήσει ένα ευρετήριο από κείμενα των οποίων οι ερμηνείες τόσο των λέξεων που περιέχουν, όσο και των εκτιμώμενων λέξεων που περιέχει η επερώτηση προσδιοριστούν, τότε δεν θα ανακτηθούν μη σχετικά κείμενα που περιέχουν μεν τις λέξεις της επερώτησης, αλλά τους αποδίδουν διαφορετική ερμηνεία.

Κάποιοι ερευνητές βρήκαν ότι το word sense disambiguation οδηγεί σε μικρή, εάν όχι σε καθόλου, βελτίωση της απόδοσης της ανάκτησης. Ανάμεσά τους ήταν ο Weiss[8], που χρησιμοποιώντας τον disambiguator που υλοποίησε για να επιλύσει τις ερμηνείες πέντε διφορούμενων λέξεων της συλλογής ADI, παρουσίασε μόλις 1% βελτίωση στην απόδοση της ανάκτησης.

Μια πιο εκτεταμένη έρευνα στην αμφιβολία και στην ανάκτηση πληροφοριών παρουσίασαν οι Krovetz και Croft[9], χρησιμοποιώντας τις συλλογές CACM και TREC. Για κάθε επερώτηση στις συγκεκριμένες συλλογές παρουσίασαν μια ανάκτηση και για κάθε ανάκτηση εξέτασαν το ταίριασμα μεταξύ της ερμηνείας κάθε επερώτησης και της ερμηνείας των λέξεων σε ένα πλήθος ανακτηθέντων κειμένων. Αυτή η χειρονακτική έρευνα περιλάμβανε την μελέτη εκατομμυρίων ταιριασμάτων των ερμηνειών ερωτήσεων/κειμένων. Αυτή η μελέτη κατέληξε στο ότι η μη αντιστοιχία των ερμηνειών συμβαίνει συχνότερα όταν τα κείμενα δεν είναι σχετικά με την επερώτηση και κατά επέκταση αυτή η μη αντιστοιχία είναι πιθανότερο να

συμβαίνει, όταν υπάρχει μικρό πλήθος κοινών λέξεων μεταξύ του κειμένου και της επερώτησης. Ολοκληρώνοντας την μελέτης τους, λοιπόν, υποστήριξαν ότι η επίδραση της αμφιβολίας των εννοιών στην ανάκτηση πληροφορίας δεν είναι δραματική, άλλα ο προσδιορισμός της ερμηνείας αποδίδει πλεονεκτήματα στην ανάκτηση, κυρίως όταν υπάρχουν λίγες κοινές λέξεις μεταξύ του κειμένου και της ερώτησης.

Η Voorhees[10] και ο Wallis[11] εφάρμοσαν ένα disambiguator σε σύστημα ανάκτησης πληροφοριών με μεγάλης κλίμακα δεδομένα. Η Voorhees[10] κατασκεύασε έναν disambiguator που εκμεταλλεύτηκε την σημασιολογία του WordNet, ώστε να βελτιώσει την αποδοτικότητα της ανάκτησης δημιουργώντας ευρετήριο με τις ερμηνείες των λέξεων (word sense) αντί με την προέλευση των λέξεων (word stems). Ο αλγόριθμος εφαρμόστηκε στις συλλογές CACM, CISI, CRAN, MED και TIME. Δυστυχώς όμως τα αποτελέσματα έδειξαν ότι αυτή η αποτελεσματικότητα των πινάκων της *technique* disambiguation είναι χειρότερη από την χρήση των word stem πινάκων και για τις πέντε συλλογές. Η υποβίβαση της αποτελεσματικότητας οφείλεται κυρίως στην δυσκολία προσδιορισμού των ερμηνειών των λέξεων στις μικρές επερωτήσεις, γιατί το μικρό περιβάλλον που διαθέτουν, είτε δεν είναι ικανό να επιλύσει τον προσδιορισμό της ερμηνείας των διφορούμενων λέξεων, είτε επιλέγει λάθος ερμηνείες. Όπως είναι αντιληπτό και στις δύο περιπτώσεις, η επερώτηση δεν οδηγεί σε κείμενα που η ερμηνεία τους επιλύθηκε σωστά. Τα αποτελέσματα επίσης, έδειξαν ότι η απουσία ταιριάσματος της ερώτησης με τα κείμενα μειώνει την απόδοση πολύ περισσότερο από ότι η απομάκρυνση των λανθασμένων ταιριασμάτων, που βοηθούν στην ανάκτηση μικρών και ομοιογενών συλλογών. Το θηικό δίδαγμα αυτής της έρευνας ήταν ότι ακόμα κι αν μια τέτοια μέθοδο μπορεί να αντιγραφεί στις μικρές ερωτήσεις, μπορεί να μην είναι χρήσιμη για την ανάκτηση.

Ο Wallis[11] χρησιμοποίησε τον disambiguator, σαν μέρος ενός πιο πολύπλοκου πειράματος, στο οποίο αντικαθιστούσε τις λέξεις της συλλογής με κείμενο από τους ορισμούς τους σε λεξικό. Αυτή η διαδικασία ακολουθήθηκε, έτσι ώστε οι συνώνυμες λέξεις να αναπαρασταθούν με παρόμοιο τρόπο και να ανακτηθούν μαζί και κείμενα που περιέχουν αυτές τις συνώνυμες λέξεις. Όταν μια λέξη αντικαθίσταται με τον ορισμό της, ο disambiguator έπρεπε να επιλέξει τον ορισμό που προσδιόριζε την ερμηνεία της περισσότερο. Οι συλλογές που εκτέλεσε τα πειράματά του ήταν οι CACM και TIME, αλλά δεν βρήκε καμιά σημαντική βελτίωση στην απόδοση της ανάκτησης.

Πρέπει όμως να επισημανθεί ότι τα αποτελέσματα των Voorhees και Wallis είναι εντυπωσιακά, καθώς φαίνεται λογικό ότι εάν λυθεί η αμφιβολία των ερμηνειών, η απόδοση των IR θα αυξηθεί, αφού το κύριο πρόβλημα που αντιμετώπισαν είναι η έλλειψη των αξιόπιστων χαρακτηριστικών για τους disambiguators που κατασκεύασαν.

Ο Sanderson[12] παρουσίασε παρόμοια πειράματα στα οποία εισήγαγε τεχνητά αμφιβολία στις συλλογές και αποκάλυψε ότι η απόδοση αυξάνεται μόνο για πολύ μικρές επερωτήσεις (λιγότερο από 5 λέξεις). Ο λόγος που οφείλεται το παραπάνω είναι ότι οι στατιστικοί αλγόριθμοι που στηριζόταν η ανάκτηση της πληροφορίας ήταν όμοιες με κάποιες προσεγγίσεις του word sense disambiguation, καθώς και ότι οι λέξεις σε επερωτήσεις μεγάλου μήκους βοηθούν να επιβεβαιωθούν οι ερμηνείες των άλλων λέξεων με σεβασμό στα κείμενα. Ο Sanderson, επίσης, επισήμανε ότι η απόδοση των συστημάτων ανάκτησης δεν είναι τόσο ευαίσθητοι στην αμφιβολία (ambiguity), όσο είναι στην λανθασμένη επιβεβαίωση των ερμηνειών (disambiguation).

Μια άλλη έρευνα που αναφέρεται στο WSD στην ανάκτησης της πληροφορίας είναι των Stokoe, Oakes και Tait[13] που προσδιόρισαν την ερμηνεία πολυσήμαντων λέξεων, χρησιμοποιώντας σε βήματα την σύνταξη, τα στατιστικά συνύπαρξης και την συχνότητα των κύριων ερμηνειών. Η τακτική αυτή προσπαθεί να μειώνει την επίδραση λανθασμένων μη διφορούμενων εννοιών στην επίδοση της ανάκτησης και υποστηρίζει ότι έχει θετικά



αποτελέσματα, τόσο στην ακρίβεια (collocation και co-occurrence), όσο και στην ανάκληση (raw sense frequency statistics). Τα κείμενα και οι επερωτήσεις αναπαρίστανται με διανύσματα ερμηνειών και τα κείμενα που ανακτώνται χρησιμοποιούν την παραδοσιακή μέθοδο υπολογισμού του βάρους των όρων, την  $t \times idf$ . Υπάρχουν όμως δύο βασικά προβλήματα στο σύστημα τους: πρώτον ο αλγόριθμος είναι επιβλεπόμενης μάθησης και χρησιμοποιεί την συλλογή SemCor που έχει προσδιορισμένες της ερμηνείες, καθώς και η αποδοτικότητα της ανάκτησης είναι μειωμένη.

### Μηχανές μετάφρασης

Η αποδοτική διαδικασία του word sense disambiguation αποτελούσε την λύση των προβλημάτων κι αυτού του πεδίου. Οι Hutchins και Sommers[14] έδειξαν ότι υπάρχουν δύο τύποι λεξιλογικής σημασιολογικής αμφιβολίας που οι μηχανές μετάφρασης θα έπρεπε να αντιμετωπίζουν. Ένας τύπος είναι η αμφιβολία που υπάρχει στην γλώσσα πηγής στην οποία η ερμηνεία της λέξης μπορεί να μην είναι προφανής αμέσως και ο άλλος είναι η αμφιβολία στην γλώσσα που προκύπτει όταν η λέξη δεν είναι διφορούμενη στην γλώσσα πηγής, αλλά έχει δύο πιθανές μεταφράσεις στην γλώσσα προορισμού.

Ο Brown et. al.[15] κατασκεύασε ένα αλγόριθμο word sense disambiguation για ένα σύστημα μετάφρασης από Αγγλικά σε Γαλλικά χρησιμοποιώντας μια διαγλωσσική παράλληλη συλλογή, με προτάσεις ευθυγραμμισμένες. Αυτή η προσέγγιση έλυνε μόνο τον πρώτο τύπο αμφιβολίας. Ο Brown απέδειξε ότι το 45% των μεταφράσεων ήταν αποδεκτές, όταν χρησιμοποιούνταν μηχανές disambiguation, ενώ μονό 37% όταν δεν χρησιμοποιούνταν. Αυτό που είναι εμφανής από τον παραπάνω συμπέρασμα είναι η χρησιμότητα του word sense disambiguation στις μηχανές μετάφρασης.

Σε αυτόν τον τομέα έχουν γίνει παρά πολλές έρευνες. Κάποιοι άλλοι εκπρόσωποι είναι οι Dagan και Itai [16] που χρησιμοποίησαν ένα διαγλωσσικό λεξικό και αναλυτές λέξεων (parsers) σε συνδυασμό με μια διαγλωσσική παράλληλη συλλογή. Οι Kali και Morimoto [17] χρησιμοποίησαν, και αυτοί, διαγλωσσικό λεξικό, αλλά η μέθοδος τους απαιτούσε μια διαγλωσσική συγκρίσιμη συλλογή. Παράλληλα, χαρτογράφησαν τις συσχετίσεις μεταξύ των ερμηνειών και τις χρησιμοποιούσαν σε μια μαθηματική φόρμουλα. Αυτές τις σχέσεις τις εντόπισαν χρησιμοποιώντας το λεγόμενο context window.

Αργότερα, οι Oliveira, Wong, Li, Zheng [18] παρουσίασαν έναν στατιστικό WSD με εφαρμογή σε μηχανές μετάφρασης από πορτογαλικά σε κινέζικα. Εξαιτίας της περιορισμένης διαθεσιμότητας από πηγές πορτογαλικών-κινέζικων ψηφιακών συλλογών και σχολιασμών, εφαρμόστηκε μια unsupervised learning και μη ευθυγραμμισμένη διαγλωσσική συλλογή. Η προτεινόμενη μέθοδος αρχικά αναγνώριζε λέξεις που σχετίζονται με κάθε μια από τις διφορούμενες λέξεις με βάση τις λέξεις του περιβάλλοντος τους και τις σχετικές αποστάσεις τους. Στην συνέχεια, ένα μαθηματικό μοντέλο εφαρμόζεται για να μπορέσει να αναγνωρίσει την περισσότερο κατάλληλη ερμηνεία μιας διφορούμενης λέξης, σε σχέση με τις λέξεις που σχετίζεται. Όλες οι ερμηνείες που βρίσκονται μετατρέπονται σε ένα σύνολο κανόνων σε μια βάση γνώσης, ώστε να χρησιμοποιηθούν στην διαδικασία του disambiguation και της μετάφρασης. Κάποια βασικά αποτελέσματα έδειξαν μια βελτίωση 6% από την βασική μέθοδο, δηλαδή χωρίς τη χρήση της παραπάνω τεχνικής.

### **4.3 Περιγραφή των Προσεγγίσεων του Word Sense Disambiguation**

Για το WSD έχουν προταθεί δύο διαφορετικές μεθοδολογίες, η επιβλεπόμενη (Supervised Disambiguation), στην οποία κατά την εκπαίδευση χρησιμοποιούνται δεδομένα πάρο

είναι γνωστή η ερμηνεία τους (*sense-tagged data*) και η μη-επιβλεπόμενη (*Unsupervised Disambiguation*), στην οποία κατά τη διάρκεια της εκπαίδευσης, οι ετικέτες της ερμηνείας της κάθε λέξης δεν είναι γνωστές. Ωστόσο, η προσέγγιση της επιβλεπόμενης μάθησης δεν έχει πρακτική σημασία, αφού καταρχήν το κόστος δημιουργίας τέτοιων συλλογών είναι τεράστιο και ο χρόνος που χρειάζεται για να γίνει ο χειρονακτικός προσδιορισμός των ερμηνειών είναι απαγορευτικός.

Πιο συγκεκριμένα, όλες οι προσεγγίσεις του προβλήματος αυτού μπορούν να κατηγοριοποιηθούν σε μια από τις τρεις γενικές στρατηγικές: αυτών που βασίζονται στην γνώση (*knowledge based*), αυτών που βασίζονται σε συλλογές (*corpus based*) και των υβριδικών (*hybrid*).

### Knowledge based

Σε αυτήν την προσέγγιση το *disambiguation* εκτελείται χρησιμοποιώντας πληροφορίες από ένα σαφή λεξικό ή βάση γνώσης. Το λεξικό μπορεί να είναι ένα λεξικό μηχανής ή ένας θησαυρός (ιεραρχικός ή μη). Αυτή είναι η πιο δημοφιλής προσέγγιση και έχουν γίνει πάρα πολλές έρευνες, χρησιμοποιώντας λεξιλογικές πηγές γνώσης, όπως το WordNet, το LDOCE και τον διεθνή θησαυρό του Roget.

Η πληροφορία αυτών των πηγών χρησιμοποιήθηκαν με διάφορους τρόπους. Οι Wilks και Stevenson [19], καθώς και οι Harley και Glennon [20] χρησιμοποίησαν μεγάλα λεξικά και πληροφορίες που σχετίζονται με τις ερμηνείες, όπως ετικέτες του μέρους του λόγου και τοπικούς οδηγούς, ώστε να κατορθώσουν να επιδείξουν την σωστή ερμηνεία. Άλλη μια προσέγγιση είναι το κείμενο να συμπεριφέρεται σαν μια ομάδα λέξεων, χωρίς κάποια ιδιαίτερη σειρά, σε αυτή την μορφή τα μέτρα ομοιότητας υπολογίζονται εξετάζοντας την σημασιολογική ομοιότητα, όπως αυτή υπολογίζεται από την πηγή γνώσης, ανάμεσα σε όλες τις λέξεις στο παράθυρο άσχετα με την θέση τους. Η παραπάνω προσέγγιση υιοθετήθηκε από τον Yarowsky [21].

Οι Mihalcea, Tarau και Figa [22] πρότειναν ένα μη επιβλεπόμενο αλγόριθμο για WSD βασισμένο στην προσέγγιση αυτή. Προσπάθησαν να εφαρμόσουν αλγόριθμους ίδιου στυλ με τους Page Rank, σε εννοιολογικά γραφήματα βασισμένα στο WordNet, που οι ερμηνείες των λέξεων είναι κορυφές και οι σχέσεις είναι ακμές. Ο αλγόριθμος Page Rank που εφαρμόστηκε σε αυτό το γράφημα είχε σαν σκοπό να πάρει το κόμβο με το υψηλότερο βαθμό. Επιπρόσθετα, μαζί με τον αλγόριθμο Page Rank χρησιμοποιήθηκαν και οι ορισμοί των λέξεων που βρισκόντουν στο ίδιο περιβάλλον, καθώς και η πρώτη ερμηνεία για κάθε λέξη στο WordNet. Η ακρίβεια που διαπιστώθηκε σε αυτήν την μελέτη ήταν 70,32%.

Οι Patwardhan, Banerjeev και Pedersen [23] παρουσίασαν μια ομάδα από WSD αλγόριθμους βασισμένους στις σημασιολογικές σχέσεις που μετρήθηκαν από το WordNet. Ο αλγόριθμος με την καλύτερη απόδοση ήταν μια επέκταση του Lesk αλγορίθμου που χρησιμοποιούσε τους ορισμούς των λέξεων, για να προσδιοριστούν οι ερμηνείες τους, καθώς και τους ορισμούς των υπονύμων (*hyponyms*) τους. Η ακρίβεια που προσδιορίστηκε ήταν 39,1%.

Για την καλύτερη κατανόηση των παραπάνω γίνεται μια παρένθεση περιγράφοντας τον Lesk αλγόριθμο. Αυτός προσδιορίζει την ερμηνεία (*disambiguate*) μιας πολυσήμαντης λέξης-στόχου με το να συγκρίνει την επεξήγησή της με αυτή των λέξεων που βρίσκονται στο περιβάλλον της. Η λέξη-στόχος προσδιορίζεται από την ερμηνεία, που η επεξήγηση της έχει την μεγαλύτερη επικάλυψη ή μοιράζονται πολλές λέξεις με τις επεξηγήσεις των γειτονικών λέξεων. Εξαιτίας του ότι οι επεξηγήσεις των λεξικών είναι αρκετά συχνά πολύ σύντομες και

περιλαμβάνουν αποτελεσματικό λεξιλόγιο για να προσδιοριστούν οι συσχετίσεις των ερμηνειών, οδηγήθηκαν στην επέκταση του αλγορίθμου χρησιμοποιώντας το WordNet. Εκμεταλλεύτηκε την εννοιολογική ιεραρχία του WordNet, ώστε να επιτρέπει στις επεξηγήσεις των ερμηνειών των λέξεων να συσχετίζονται, αλλά και να συγκρίνονται, με τις λέξεις του περιβάλλοντος,

Τέλος, αναφέρετε μια έρευνα που αντί να προσπαθεί να προσδιορίσει τις ερμηνείες των πολυσήμαντων λέξεων ενός κειμένου, όπως όλες οι παραπάνω μελέτες, απευθύνεται στις κεντρικές λέξεις των επερωτήσεων. Οι Shuang Lui, Clement Yu και Weiyi Meng [24] χρησιμοποίησαν το WordNet με έναν πιο πολύπλοκο τρόπο από τους παραπάνω, αφού ερευνούσαν συνώνυμα, υπόνυμα, τους ορισμούς τους, υπέρνυμα και πληροφορίες του domain. Επιπλέον, όταν ένας όρος μπορούσε να ερμηνευτεί με παραπάνω από έναν τρόπο, ένα πολύπλοκο σχήμα επιδρούσε για να καθοριστεί η κατάλληλη ερμηνεία. Παράλληλα, γινόταν χρήση της επικρατέστερης ερμηνείας, καθώς και του WEB, σε περιπτώσεις που δεν μπορούσε να καθοριστεί η ερμηνεία του. Η ακρίβεια που κατάφεραν να επιτύχουν ήταν 90%.

### Corpus based

Αυτή η προσέγγιση χρησιμοποιεί την πληροφορία που παρέχεται από κάποιες συλλογές εκπαίδευσης, αντί να παίρνει την πληροφορία κατευθείαν από μια σαφή πηγή γνώσης. Η εκπαίδευση μπορεί να γίνει σε συλλογές που είναι είτε disambiguated, δηλαδή υπάρχει η σημασιολογία σε κάθε πολυσήμαντο αντικείμενο του λεξιλογίου, είτε απλές, δηλαδή συλλογές που δεν έχουν υποστεί κάποια επεξεργασία.

Πριν από την αναλυτική περιγραφή των συλλογών, μελετήθηκαν κάποιοι αλγόριθμοι για επιβλεπόμενο WSD, που βασίζονται σε αυτήν την προσέγγιση. Καταρχήν, ο Naïve-Bayes αποτελεί μια απλή αναπαράσταση μεθόδου στατιστικής μάθησης. Υποθέτοντας την ανεξαρτησία των χαρακτηριστικών, μπορεί να ταξινομήσει ένα νέο παράδειγμα με το να προσδιορίσει την κλάση, που μεγιστοποιεί την εξαρτημένη της πιθανότητα, δοθέντος της παρατηρούμενης σειράς των χαρακτηριστικών του παραδείγματος. Τα πιθανοτικά μοντέλα υπολογίζονται κατά την διάρκεια της διαδικασίας εκπαίδευσης χρησιμοποιώντας τις σχετικές συχνότητες.

Ο ταξινομητής βασισμένος σε πρότυπα (Exemplar-based Classifier) αποθηκεύει απλά τα παραδείγματα στην μνήμη και η ταξινόμηση νέων παραδειγμάτων βασίζεται στην ομοιότητα τους με τα πρότυπα που είναι αποθηκευμένα. Ένας από τους τρόπους να υπολογιστεί αυτή η ομοιότητα είναι χρησιμοποιώντας τον αλγόριθμο k-Nearest-Neighbours και την απόσταση Hamming.

Η αρχιτεκτονική SNoW αποτελεί έναν άλλο αλγόριθμος που εφαρμόζεται για τον προσδιορισμό των στατιστικών μοντέλων και των ταξινομητών μιας συλλογής, για να εκτελεστεί το WSD. Το SNoW είναι ένα αραιό δίκτυο γραμμικών διαχωριστών που κάνει χρήση του αλγόριθμο μάθησης Winnow. Σε αυτήν την αρχιτεκτονική, υπάρχει ένας κόμβος winnow για κάθε κλάση που μαθαίνει να διαχωρίζει την συγκεκριμένη κλάση από τις υπόλοιπες. Κατά την διάρκεια της εκπαίδευσης, κάθε παράδειγμα αποτελεί ένα θετικό παράδειγμα για τον winnow κόμβο που συσχετίζεται με την κλάση του και αρνητικό για όλους του υπόλοιπους. Το σημείο κλειδί για να γίνεται γρήγορα η διαδικασία της μάθησης είναι ότι οι winnow κόμβοι δεν συνδέονται με όλα τα χαρακτηριστικά, αλλά μόνο με αυτά που είναι σχετικά με την κλάση τους. Όταν ταξινομεί ένα νέο παράδειγμα, το SNoW είναι όμοιο με ένα νευρωνικό δίκτυο που παίρνει σαν είσοδο χαρακτηριστικά και εξάγει την κλάση με την μεγαλύτερη ενεργοποίηση.

Οι λίστες απόφασης (Decision Lists) είναι λίστες χαρακτηριστικών που εξάγονται από τα παραδείγματα εκπαίδευσης και ταξινομούνται με το μέτρο της λογαριθμικής πιθανότητας. Αυτό το μέτρο υπολογίζει πόσο ισχυρό είναι ένα συγκεκριμένο χαρακτηριστικό για να επιδεικνύει την συγκεκριμένη ερμηνεία. Κατά την διάρκεια της εκπαίδευσης, η λίστα απόφασης ελέγχεται με την σειρά και το χαρακτηριστικό με το μεγαλύτερο βάρος που ταιριάζει με το παράδειγμα εκπαίδευσης χρησιμοποιείται για να επιλεγεί η ερμηνεία.

Τέλος, αναφέρονται οι LazyBoosting αλγόριθμοι, που η κύρια ιδέα τους είναι να συνδυάσουν πολλές απλές υποθέσεις ακρίβειας σε ένα απλό, υψηλής ακρίβειας ταξινομητή. Οι «αδύνατοι» ταξινομητές εκπαιδεύονται σειριακά και καθένας από αυτούς, εννοιολογικά, εκπαιδεύεται σε παραδείγματα που είναι πολύ δύσκολο να ταξινομηθούν από προηγούμενους «αδύνατους» ταξινομητές. Αυτές οι υποθέσεις αδυναμίας συνδυάζονται γραμμικά σε απλούς κανόνες.

Έχοντας γίνει πολλές συγκρίσεις μεταξύ τέτοιων αλγορίθμων σύμφωνα με τον Escudero, τον Marquez και τον Rigau [25] τα δύο βασικά συμπεράσματα είναι ότι οι αλγόριθμοι LazyBoosting υπερτερούν από τους άλλους τέσσερις σε ότι αφορά την ακρίβεια και την ικανότητα να εναρμονίζονται σε νέα domains, καθώς και ότι η εξάρτηση των domains από τα συστήματα WSD φαίνεται να είναι ισχυρή.

Στην συνέχεια περιγράφονται αναλυτικά οι συλλογές που μπορεί να γίνει η εκπαίδευση.

### *Disambiguated συλλογές*

Αυτό το σύνολο τεχνικών απαιτούν μια συλλογή εκπαίδευσης στην οποία έχουν αποσαφηνιστεί οι ερμηνείες των πολυσήμαντων λέξεων. Γενικά, ο αλγόριθμος μηχανικής μάθησης εφαρμόζεται σε κάποια κεντρικά χαρακτηριστικά που εξάγονται από την συλλογή και χρησιμοποιούνται για να σχηματίζονται μια αναπαράσταση για καθεμία από τις ερμηνείες. Αυτές οι αναπαραστάσεις μπορούν στην συνέχεια να χρησιμοποιηθούν για να αποσαφηνιστούν νέα περιστατικά. Διαφορετικές έρευνες έχουν χρησιμοποιήσει διαφορετικά σύνολα χαρακτηριστικών, χαρακτηριστικά παραδείγματα είναι η τοπική σύνταξη την μελέτη του Brown, ή πιο συνηθισμένα, όλες τις λέξεις σε ένα παράθυρο λέξεων ανάμεσα στις διφορούμενες λέξεις, με στόχο να επεξεργάζεται το περιβάλλον σαν μια ομάδα από λέξεις χωρίς σειρά.

Μια άλλη προσέγγιση είναι η χρήση μοντέλων Markov, τα οποία έχουν αποδειχτεί πολύ επιτυχημένα στον προσδιορισμό των ετικετών του μέρους του λόγου.

### *Τεχνικές Συλλογές*

Η δυσκολία απόκτησης συλλογών που προσδιορίζουν την ερμηνεία τους με ετικέτες οδήγησαν τους ερευνητές να βρουν καινοτόμους τρόπους για την δημιουργία τεχνικών συλλογών, οι οποίες περιέχουν κάποιους τύπους σημασιολογικών ετικετών.

Ο πρώτος τύπος τεχνικής συλλογής που χρησιμοποιήθηκε ήταν οι παράλληλες (parallel) συλλογές. Οι διαγλωσσικές συλλογές αποτελούνται από δύο συλλογές που περιέχουν τα ίδια κείμενα στις διαφορετικές γλώσσες, αυτά μπορεί να είναι μετάφραση του ενός στο άλλο, ή να έχουν δημιουργηθεί από οργανισμούς όπως η Ευρωπαϊκή Ένωση που έχουν αντίγραφα συνδιαλέξεων σε διάφορες γλώσσες. Υπάρχουν διάφοροι αλγόριθμοι για να εκτελέσουν την διαδικασία λήψης μιας τέτοιας συλλογής και το ταίριασμα των προτάσεων, που μεταφράζουν η μια την άλλη με μεγάλη ακρίβεια. Όταν εκτελεστεί αυτή η διαδικασία, η διαγλωσσική συλλογή

μετατρέπεται σε μια ευθυγραμμισμένη παράλληλη συλλογή. Αυτή η μορφή συλλογής χρησιμοποιείται από την διαδικασία του word sense disambiguation, για τον προσδιορισμό των εννοιών των λέξεων που μεταφράζονται διαφορετικά από τις γλώσσες.

Υπάρχουν δύο τρόποι δημιουργίας τέτοιου είδους συλλογής: ο πρώτος είναι να αποσαφηνίστούν οι λέξεις με τρόπο όπως στις παράλληλες συλλογές, κι ο δεύτερος τρόπος προσέγγισης είναι να προστεθεί αμφιβολία (ambiguity) στην συλλογή και να βρεθεί ένας αλγόριθμος που θα προσπαθεί να επιλύσει αυτή την αμφιβολία στην αρχική συλλογή. Ο Yarowsky [26] χρησιμοποίησε την παραπάνω τεχνική δημιουργώντας μια συλλογή που περιείχε ψευδό-λέξεις. Οι ψευδό-λέξεις αυτές προέρχονταν από τη επιλογή δύο λέξεων και την αντικατάστασή τους σε κάθε χρήση τους με την ένωσή τους. Παραδείγματος χάρη, όταν συναντούσε τις λέξεις «crocodile» και «shoes» τις αντικαθιστούσε με την «crocodile/shoes». Η πηγή των περιστατικών που χρησιμοποιούσε ο disambiguator (π.χ. λεξικό) θα έπρεπε να ενημερωθεί, ώστε να αντανακλά την ένωση των δύο λέξεων. Στην συνέχεια, ο disambiguator εφαρμοζόταν σε κάθε εμφάνιση της καινούργιας λέξης. Όσο αφορά, την αξιολόγηση των αποτελεσμάτων του disambiguator είναι ασήμαντης ουσίας, καθώς ήταν γνωστή η σωστή ερμηνεία κάθε εμφάνισης της λέξης προκαταβολικά. Ωστόσο, υπήρχαν κάποιοι περιορισμοί. Η μέθοδος που επιλέχτηκε να σχηματίζει τις ψευδό-λέξεις από τις ατομικές λέξεις είναι τυχαίας επιλογής. Συνεπώς, οι διάφορες ερμηνείες μιας ψευδό-λέξης είναι απίθανο να συσχετίζονται στενά και αυτό διαφέρει από την αντιστοιχία των πραγματικών διφορούμενων λέξεων, που οι ερμηνείες σχετίζονται με κάποιον τρόπο. Η σημασία αυτής της διαφοράς δεν είναι ξεκάθαρη και συνεπώς δεν μπορεί να ισχυριστεί ότι η αμφιβολία που εισήγαγε ταιριάζει ακριβώς με την αμφιβολία που βρίσκεται στις πραγματικές καταστάσεις.

### Ακατέργαστες (Raw) Συλλογές

Η δυσκολία, τόσο του να αποκτήσεις λεξιλογικές πηγές, όσο και οι δυσκολίες να αποκτήσεις μη διφορούμενα κείμενα από supervised disambiguation οδήγησαν στην μελέτη τρόπων εξερεύνησης ακατέργαστων, δηλαδή χωρίς σχολιασμό, συλλογών για την παρουσίαση unsupervised disambiguation. Αυτός ο τρόπος αποσαφήνισης των λέξεων δεν παρέχει μια συγκεκριμένη ετικέτα στους όρους σαν αναφορά της συγκεκριμένης ερμηνείας τους, γιατί απαιτεί περισσότερη πληροφορία από αυτή που διατίθεται, αντίθετα προσφέρει την διάκριση των ερμηνειών των λέξεων. Συγκεκριμένα, ομαδοποιεί τα περιστατικά μιας λέξης σε διακριτές κατηγορίες, χωρίς να χρησιμοποιεί κατηγορίες από ένα λεξικό.

Χαρακτηριστικό παράδειγμα αυτής της κατηγορία είναι η τεχνική δυναμικού ταιριάσματος του Radford, που μελέτησε όλα τα περιστατικά, δοθέντος ενός όρου στην συλλογή και σύγκρινε το περιβάλλον που σύμβαινε για να εντοπίσει κοινές λέξεις και συντακτικούς τύπους. Έτσι, σχηματίζεται ένας πίνακας ομοιότητας που υπόκειται στην ανάλυση των ομάδων, για τον καθορισμό των σχετικών σημασιολογικών συνόλων των περιστατικών των όρων.

Άλλο ένα παράδειγμα είναι η εργασία του Pedersen [27] που σύγκρινε τρεις διαφορετικούς απρόβλεπτους αλγορίθμους μάθησης με 13 διαφορετικές λέξεις. Κάθε αλγόριθμος εκπαιδεύτηκε σε κείμενα με τοποθετημένες ετικέτες είτε από το WordNet, είτε από το LDOCE, για τον καθορισμό των ερμηνειών των λέξεων. Ο αλγόριθμος είχε πρόσβαση στο πλήθος των ερμηνειών για κάθε λέξη και κάθε αλγόριθμος χωρίζε τα περιστατικά κάθε λέξης στο κατάλληλο πλήθος ομάδων. Στην συνέχεια, αυτές οι ομάδες χαρτογραφήθηκαν με βάση την στενότερη ερμηνεία από το κατάλληλο λεξικό. Δυστυχώς όμως, τα αποτελέσματα δεν ήταν πολύ ενθαρρυντικά, ο Pedersen ανέφερε ότι το 65-66% της σωστής αποσαφήνισης εξαρτάται από τον αλγόριθμο μάθησης που χρησιμοποιείται. Αυτό το αποτέλεσμα οφείλει να συγκριθεί

όμως με τον ισχυρισμό ότι για την συλλογή που χρησιμοποιήθηκε το 73% των περιπτώσεων θα μπορούσε να ταξινομηθεί σωστά με το να επιλέγει απλά η συχνότερη ερμηνεία.

### Hybrid

Αυτές οι προσεγγίσεις χρησιμοποιούν τμήματα κι από τις δυο παραπάνω προσεγγίσεις. Ένα χαρακτηριστικό παράδειγμα είναι το σύστημα Luk [29] που χρησιμοποιεί τους ορισμούς των ερμηνειών που αναφέρονται σε κείμενα από ένα λεξικό μηχανής (LDOCE), ώστε να αναγνωρίσει σχέσεις ανάμεσα σε ερμηνείες και επιπλέον χρησιμοποιεί μια συλλογή για να υπολογίσει την αμοιβαία πληροφορία των σκορ ανάμεσα στις σχετικές ερμηνείες, ώστε να ανακαλύψει την πιο χρήσιμη. Αυτή η αρχιτεκτονική επιτρέπει στον Luk να παράγει ένα σύστημα που χρησιμοποιεί την πληροφορία των λεξιλογικών πηγών σαν ένα τρόπο να μειώσει τον όγκο των κειμένων που χρειάζονται στις συλλογές εκπαίδευσης.

Ένα άλλο παραδείγματα της προσέγγισης αυτής είναι ο unsupervised αλγόριθμος του Yarowsky [30]. Αυτός παίρνει ένα μικρό αριθμό «օρισμών» (WordNet synsets ή ορισμούς από κάποιο λεξικό) των ερμηνειών κάποιων λέξεων και τους χρησιμοποιεί για να ταξινομήσει τις προφανείς περιπτώσεις σε μια συλλογή. Στην συνέχεια, χρησιμοποιεί λίστες απόφασης για να κάνουν γενικεύσεις, οι οποίες βασίζονται σε περιστατικά της συλλογής που έχουν ταξινομηθεί. Στην συνέχεια, αυτές οι λίστες ξανά εφαρμόζονται στην συλλογή για να ταξινομήσουν περισσότερα περιστατικά. Η μάθηση συνεχίζει με αυτόν τον τρόπο μέχρι να ταξινομηθούν όλα τα περιστατικά της συλλογής. Συμπερασματικά, θα πρέπει να τονιστεί ότι ο αλγόριθμος του Yarowsky δουλεύει βασιζόμενος σε αρκετές δυνατές και εμπειρικά παρατηρούμενες ιδιότητες της γλώσσας, δηλαδή την δυνατή τάση των λέξεων να παρουσιάζουν μόνο μια ερμηνεία ανά σύνταξη και ανά κείμενο. Αυτός προσπαθεί να παράγει τη μέγιστη ισχύ των συντακτικών σχέσεων και χρησιμοποιεί περισσότερο διακριτές πληροφορίες από άλλους αλγόριθμους που επεξεργάζονται τα κείμενα ως ομάδες από λέξεις που αγνοούν την σχετική θέση και σειρά των λέξεων. Ένα σημαντικό χαρακτηριστικό του αλγορίθμου είναι η ευαισθησία του στην ευρεία έκταση των λεπτομερειών της γλώσσας, παρά στην τυπική εστίαση στους στατιστικούς αλγορίθμους αποσαφήνισης. Τέλος, ο Yarowsky αναφέρει ότι το σύστημα ταξινομεί σωστά το 96% των ερμηνειών. Αυτός ο ισχυρισμός δείχνει ότι το κόστος των μεγάλων συλλογών εκπαίδευσης με προσδιορισμένες τις ερμηνείες, ίσως να μην είναι απαραίτητο για να κατορθωθεί η ακρίβεια του WSD.

## 5. WordNet

### 5.1 Εισαγωγή

Το WordNet αποτελεί ένα λεξικογραφικό σύστημα, το οποίο έχει κατασκευαστεί χειρονακτικά από τον George Miller και τους συναδέλφους του στο πανεπιστήμιο του Princeton.

Στόχος αυτού του συστήματος ήταν να αναπτυχθεί ένα λεξικό, στο οποίο να είναι δυνατή η εννοιολογική αναζήτηση, εκτός από την αλφαριθμητική. Το WordNet, τελικά, εξελίχτηκε σε ένα σύστημα που αντανακλά τις τρέχουσες ψυχογλωσσολογικές θεωρίες, για το πώς οι άνθρωποι οργανώνουν τις λεξικολογικές τους μνήμες.

Αν και τα περιεχόμενα του WordNet περιλαμβάνουν σύνθετες λέξεις, φραστικά ρήματα, ιδιωματικές εκφράσεις και παραθέσεις, τη βασική του δομή αποτελεί η λέξη. Όμως, η δομή του δεν περιλαμβάνει ούτε ανάλυση των λέξεων σε μικρότερες σημαντικές δομές, ούτε οργανωτικές δομές μεγαλύτερες από τις λέξεις, όπως κείμενα ή frames (περιλαμβάνουν όλες τις λεξικογραφικές έννοιες που σχετίζονται με κάποια είδη καταστάσεων), που συνήθως αποτελούν

την δομή κατασκευής των λεξικών. Όμως, κάποιες από τις παραπάνω σχέσεις αντανακλώνται από τις ισχύουσες σχέσεις του WordNet.

Γνωρίζοντας, τον διαχωρισμό που επικρατεί μεταξύ της λεξιλογικής γνώσης και της εγκυκλοπαιδικής, το WordNet δεν προσπαθεί να συμπεριλάβει την εγκυκλοπαιδική γνώση. Εντούτοις, διαθέτει ορισμούς που περιλαμβάνουν πληροφορία για την ερμηνεία των λέξεων, που δεν αποτελούν μέρος των λεξιλογικών του δομών.

Όσο αφορά τον σχεδιασμό του WordNet, αυτός μοιάζει περισσότερο με αυτών των θησαυρών. Το βασικό του αντικείμενο είναι ένα σύνολο συνωνύμων, που λέγεται synset. Κάθε διαφορετικό synset, στο οποίο εμφανίζεται η ίδια λέξη, διαθέτει διαφορετική ερμηνεία της λέξης αυτής. Όμως, το WordNet κάνει περισσότερα από να ταξινομεί τις ερμηνείες (concepts) των λέξεων στα synsets. Ταυτόχρονα, συνδέει μεταξύ τους τα synsets με σχέσεις συνεπαγωγής, υπονύμων (hyponymy), μερονύμων (meronymy) κ.τ.λ. Με αυτό τον τρόπο, το WordNet διαχωρίζει τις θεμελιώδεις από τις λεξιλογικές σχέσεις.

Σε αντίθεση με τους θησαυρούς, στο WordNet, οι σχέσεις μεταξύ ερμηνειών και λέξεων είναι ξεκάθαρες και προσδιορισμένες. Ο χρήστης μπορεί να επιλέξει τη σχέση που τον οδηγεί από μια ερμηνεία στην επόμενη, καθώς και να επιλέξει την κατεύθυνση που θα ακολουθήσει στο χώρο των ερμηνειών (concepts).

Το WordNet δίνει ορισμούς και δείγματα προτάσεων για τα περισσότερα synsets, σαν ένα παραδοσιακό λεξικό. Ο ορισμός είναι έγκυρος για όλα τα συνώνυμα εντός του synset, αφού εκφράζει την ίδια σημασιολογική ερμηνεία. Σε περίπτωση, όμως που τα παραδείγματα δεν εκφράζουν καλά όλα τα συνώνυμα, είναι δυνατόν να υπάρχουν διαφορετικές προτάσεις για τα διαφορετικά μέλη του synset. Παράλληλα, περιέχει πληροφορίες για τη μορφολογική σχέση των λέξεων και τη σύνδεση σχετικών επίθετων με συγκεκριμένα ουσιαστικά.

Όσο αφορά τις σχέσεις που διαθέτει το WordNet, αυτό δεν περιέχει συντακτικές σχέσεις, συνδέοντας λέξεις από διαφορετικές συντακτικές κατηγορίες. Ο βασικός λόγος που συμβαίνει αυτό είναι ότι οι τέσσερις βασικές συντακτικές κατηγορίες που διαθέτει για κάθε ουσιαστικό, ρήμα, επίθετο και επίρρημα επεξεργάζονται ξεχωριστά. Ο διαχωρισμός αυτός οφείλεται στις σημασιολογικές διαφορές που υπάρχουν μεταξύ των σχέσεων που συνδέουν τις λέξεις και τις ερμηνείες των τεσσάρων κατηγοριών.

Εν κατακλείδι, το WordNet συνδέει τις λέξεις και τις ερμηνείες (concepts) των λέξεων μέσω μιας ποικιλίας από σημασιολογικές σχέσεις που βασίζονται στην ομοιότητα και την αντίθεση και όχι μέσω της σημασιολογίας που καθορίζει το εκάστοτε κείμενο ή η ομιλία. Εξαιτίας αυτών, το WordNet δεν περιλαμβάνει σχέσεις που προέρχονται από λέξεις που χρησιμοποιούνται για την διατύπωση ενός θέματος ομιλίας.

## 5.2 Περιγραφή της λεξιλογικής βάσης του WordNet

### Η οργάνωση των ουσιαστικών

Το WordNet (version 1.5) περιλαμβάνει σχεδόν 80,000 ουσιαστικά που οργανώνονται σε 60,000 λεξιλογικές αρχές (concepts). Πολλά από αυτά τα ουσιαστικά είναι παραθέσεις και μόνο λίγα από αυτά έχουν τεχνικά την ιδιότητα να είναι κατάλληλα για την συγκεκριμένη κατηγοριοποίηση.



Ο στόχος του WordNet διαφέρει ελαφρώς από τα πρότυπα των λεξικών και η σημασιολογία του είναι βασισμένη στην ιδέα της ερμηνείας των λέξεων, που οι λεξικογράφοι χρησιμοποιούν στα λεξικά. Πολλές πληροφορίες των λεξικών που παραλείπονται όταν αυτά γίνονται ηλεκτρονικά, παραλείπονται και από το WordNet. Χαρακτηριστικά είναι τα παραδείγματα της μη εμφάνισης της προφοράς, της μορφολογίας των παραγόμενων λέξεων, της ετοιμολογίας και των σημειώσεων χρήσης. Παρόλο αυτά γίνεται σημαντική προσπάθεια να μετασχηματίσει τις σημασιολογικές σχέσεις μεταξύ των ερμηνειών των λέξεων, σε πιο κατανοητές και πιο εύκολες στην χρήση.

Η βασική σημασιολογική σχέση είναι τα συνώνυμα, τα οποία οργανώνονται σε σύνολα (synsets). Η ιδέα των συνωνύμων δεν συνεπάγεται την εναλλαγή τους σε όλα τα περιβάλλοντα, αφού εκ των προτέρων οι φυσικές γλώσσες έχουν λίγα συνώνυμα. Έτσι, τα synsets συνδέονται σύμφωνα με τις σημασιολογικές σχέσεις και καθένα τους έχει μόνο μια απλή επεξήγηση. Για να εκφραστεί η πολυσημία μιας λέξης, θα πρέπει η λέξη αυτή να βρίσκεται σε πολλά διαφορετικά synsets.

Παρόλο που τα συνώνυμα αποτελούν μια σημασιολογική σχέση μεταξύ των λέξεων, η σημασιολογική σχέση που είναι σημαντική στην οργάνωση των ουσιαστικών στο WordNet είναι η σχέση μεταξύ των λεξιλογικών αρχών (concepts). Συγκεκριμένα, η λεξιλογική ιεραρχία του WordNet οφείλεται στη σχέση των υπονύμων(hyponyms).

Η λεξιλογική ιεραρχία μπορεί να ξανακατασκευαστεί και να ακολουθήσει και την κλίμακα των υπερνύμων (hyperonyms). Πιο συγκεκριμένα, τα υπέρνυμα είναι σχέσεις που δημιουργούνται μεταξύ συγκεκριμένων ερμηνειών της λέξης. Έτσι, η σχέση αυτή στο WordNet αναπαρίσταται από ένα δείκτη ανάμεσα στα ανάλογα synsets. Όσο αφορά την σχέση των υπονύμων, αυτή ορίζεται με δείκτες της αντίθετης κατεύθυνσης από πριν. Στο WordNet η κίνηση και προς τις δύο κατευθύνσεις είναι πολύ εύκολη και ο συνδυασμός των γενικών και των ειδικών όρων μιας λέξης εξασφαλίζουν μια λίστα όρων σχετικών με την συγκεκριμένη λέξη.

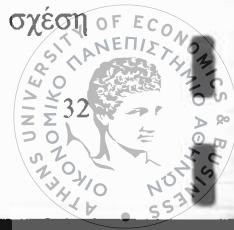
Τα ουσιαστικά στο WordNet σχηματίζουν ένα σύστημα λεξιλογικής κληρονομικότητας που σαν σκοπό του έχει να συνδέσει τα υπόνυμα με τα υπέρνυμά τους.

Το WordNet προϋποθέτει μια γλωσσολογική γνώση της αναφορικής σχέσης, δηλαδή ένα υπέρνυμο μπορεί να αντικαταστήσει έναν πιο συγκεκριμένο όρο, όποτε τα συμφραζόμενα εξασφαλίζουν ότι η αντικατάσταση δεν θα δημιουργήσει σύγχυση των εννοιών και των ερμηνειών.

Το WordNet χωρίζει τα ουσιαστικά σε διαφορές ιεραρχίες, με διαφορετική μοναδική έναρξη. Αυτή η πολλαπλή ιεραρχία αντιπροσωπεύει τα διαφορετικά σημασιολογικά μέρη του λεξιλογίου. Αφού τα χαρακτηριστικά που προσδιορίζουν την μοναδική έναρξη είναι κληρονομικά μέσω όλων των υπονύμων τους, η μοναδική έναρξη αντιπροσωπεύει ένα πρωταρχικό σημασιολογικό συστατικό στην θεωρία της λεξιλογικής σημασιολογίας.

Η λεξιλογική πηγή των αρχείων στο WordNet χρησιμοποιούν ένα σύνολο από 25 μοναδικές αφετηρίες. Αυτές οι ιεραρχίες διαφοροποιούνται σε μέγεθος και δεν είναι αμοιβαία αποκλειόμενες, αλλά στο σύνολο τους καλύπτουν ξεχωριστά λεξιλογικά και εννοιολογικά πεδία ορισμού(domain).

Επιπρόσθετα, στα αρχεία των ουσιαστικών του WordNet αναπαρίσταται μια σχέση ουσιαστικό με ουσιαστικό που εκφράζει την σχέση των σύνθετων (meronymy). Αυτή η σχέση



μπορεί να είναι μεταξύ ενός ουσιαστικού που δηλώνεται στο σύνολο του και ουσιαστικών που δηλώνουν μέρη του συγκεκριμένου ουσιαστικού.

Μια άλλη σχέση που αναπαρίσταται είναι στο WordNet είναι η σημασιολογική αντίθεση (antonymy) των λέξεων Αυτή η σχέση όμως δεν κληρονομείται σε όλα τα υπόνυμα τους.

Όσο αφορά την πολυσημία των ουσιαστικών στο WordNet, αυτό οργανώνει τις έννοιες του σύμφωνα με την συχνότητα που συμβαίνουν στα κείμενα που έχει γίνει σημασιολογικός προσδιορισμός.

Παράλληλα, προστέθηκε στο σύστημα και η ομαδοποίηση όμοιων ερμηνειών. Η ιδέα της ομαδοποίησης βασίζεται σε δύο σύνολα: των «αδελφιών» (sisters) και των «διδύμων» (twins). Το πρώτο σύνολο προσδιορίζεται όταν χρησιμοποιούνται οι δομές των δέντρων και σχηματίζεται όταν και οι δύο λέξεις είναι υπόνυμα του ίδιου κόμβου. Το δεύτερο σύνολο βασίζεται στην υπόθεση ότι εάν δύο synsets μοιράζονται τις ίδιες λέξεις (τουλάχιστον τρεις), τότε οι ερμηνείες τους είναι όμοιες. Πρέπει να τονιστεί ότι η σχέση της ομαδοποίησης είναι στο επίπεδο της υλοποίησης.

### Η οργάνωση των επίθετων και των επιρρημάτων

Τα επίθετα είναι λέξεις που έχουν σαν σκοπό να τροποποιούν τα ουσιαστικά. Το WordNet περιέχει 16,428 synsets επίθετων, περιλαμβάνοντας πολλά ουσιαστικά, μετοχές και εμπρόθετες εκφράσεις, που λειτουργούν σαν τροποποιητές. Το WordNet χωρίζει τα επίθετα ασαφώς σε δύο κατηγορίες: τα περιγραφικά (descriptive) και τα συσχετισμένα (relational).

Τα περιγραφικά επίθετα αποδίδουν στο ουσιαστικό την αξία ενός χαρακτηριστικού. Το WordNet περιέχει δείκτες μεταξύ αυτής της κατηγορίας επίθετων και των ουσιαστικών με τα οποία τα κατάλληλα χαρακτηριστικά λεξικογραφούνται. Στην συνέχεια, θα αναφερθούν σχέσεις επίθετων όπως τα αντίθετα (antonymy), τα διαβαθμισμένα (gradation), τα σημειωμένα (markedness), τα πολύσημα (polysemy), τα ποσοτικοποιημένα (quantifiers) και τα μετοχικά (participial).

Η βασική σημασιολογική σχέση, ανάμεσα στα περιγραφικά επίθετα είναι τα αντίθετα. Στο WordNet η σχέση αυτή περιγράφεται με δείκτες, που η ερμηνεία τους εκφράζεται από την ετικέτα «είναι αντίθετο με».

Πρέπει να δοθεί ιδιαίτερη έμφαση στα σοβαρά προβλήματα που δημιουργούνται στο WordNet με τις διαπιστώσεις ότι ενώ σε δύο επίθετα οι ερμηνείες τους μπορεί είναι πολύ κοντά, μπορεί να μην έχουν τα ίδια αντίθετα και ότι ενώ η σχέση των αντιθέτων είναι πολύ σημαντική, μερικά επίθετα φαίνεται να μην έχουν αντίθετα. Για την επίλυση αυτού του προβλήματος το WordNet αναπαριστά τα αντίθετα με ψηφία, γιατί τα αντίθετα είναι μια σχέση που υλοποιείται ανάμεσα σε λέξεις και μια λέξη με δύο διαφορετικές έννοιες έχει δύο διαφορετικές μορφές στο WordNet.

Παράλληλα, τα επίθετα οργανώνονται κι αυτά σε ομάδες synsets, σύμφωνα με την σημασιολογική ομοιότητα του επίθετου που αποτελεί την εστία. Ταυτόχρονα, όμως η ομάδα αυτή συσχετίζεται και με την ομάδα που περιέχει τα αντίθετα. Συμφωνά με αυτόν τον σχηματισμό, όλα τα επίθετα έχουν αντίθετα, αφού τα επίθετα που δεν έχουν άμεσα αντίθετά έχουν έμμεσα μέσω των ομάδων. Αυτή η στρατηγική είναι πολύ επιτυχημένη στα επίθετα τις αγγλικής γλώσσας. Το WordNet περιέχει παραπάνω από 1,732 ομάδες, μια για κάθε ζευγάρι αντιθέτων, συνεπώς υπάρχουν 3,464 μισές ομάδες με ομοιότητα στις ερμηνείες.



Μια σημασιολογική σχέση οργάνωσης της λεξιλογικής μνήμης των επιθέτων αποτελεί και η διαβάθμιση (gradation). Για μερικά χαρακτηριστικά, η διαβάθμιση μπορεί να εκφραστεί, ως η διάταξη των συμβολοσειρών (strings) των επιθέτων που υποδεικνύουν το ίδιο ουσιαστικό στο WordNet.

Είναι σημαντικό να κατανοηθεί ότι τα περισσότερα χαρακτηριστικά έχουν έναν προσανατολισμό. Έτσι είναι φυσικό, να θεωρηθούν ως διαστάσεις σε ένα υπέρ-διάστημα που το τέλος κάθε διάστασης να αγκυροβολεί σε ένα σημείο στο κανονικό χώρο. Το σημείο αυτό είναι η αναμενόμενη τιμή και κάθε απόκλιση από αυτό αποτελεί λάθος. Αυτό το γλωσσολογικό φαινόμενο αποτελεί το markedness (σημειωμένη τιμή του χαρακτηριστικού). Το markedness δεν κωδικοποιείται ξεκάθαρα στο WordNet, παρόλο αυτά, πάντα υπάρχει ένας δείκτης που συνδέει τα ουσιαστικά που προσδιορίζουν ένα χαρακτηριστικό και τα επίθετα που εκφράζουν τις τιμές του χαρακτηριστικού.

Τα επίθετα είτε είναι πολύσημα, είτε όχι, επιλέγονται από τα ουσιαστικά που τροποποιούν. Ο γενικός κανόνας είναι ότι εάν η αναφορά που προσδιορίζεται από το ουσιαστικό δεν έχει το χαρακτηριστικό που η τιμή εκφράζεται μέσω του επίθετου, τότε ο συνδυασμός του επίθετου με το ουσιαστικό απαιτεί τροποποίηση ή ιδιωματική ερμηνεία.

Η σημασιολογική συνεισφορά των επιθέτων εξαρτάται άμεσα από τα ουσιαστικά, αφού πολλά επίθετα έχουν διαφορετική ερμηνεία όταν προσδιορίζουν διαφορετικά ουσιαστικά. Το WordNet υποθέτει ότι η αλληλεπίδραση των επιθέτων με τα ουσιαστικά δεν είναι από πριν αποθηκεμένη, αλλά υπολογίζεται με βάση μια διερμηνευτική διαδικασία.

Οι ποσοτικοποιητές είναι λέξεις που αναφέρονται στην ποσότητα ενός ουσιαστικού (π.χ. all, some, few). Μερικοί γλωσσολόγοι περιλαμβάνουν αυτή την υποκατηγορία λέξεων στην τάξη των προσδιοριστών (the, this,...), αφού μοιράζονται κάποια χαρακτηριστικά, αλλά το WordNet τα διαχωρίζει. Οι ποσοτικές λέξεις βοηθούν στον καθορισμό της αναφοράς μιας φράσης ουσιαστικών, έτσι αφού μοιάζουν με τα επίθετα συμπεριλαμβάνονται στο αρχείο των περιγραφικών επιθέτων.

Τέλος, υπάρχει η κατηγορία των επιθέτων που σχηματίζονται από τις μετοχές. Το WordNet διατηρεί ένα ξεχωριστό αρχείο που τα εισάγει ατομικά και όχι σε ομάδες και στον αριθμό είναι περίπου 88. Για να συσχετίζονται σημασιολογικά με τις άλλες λέξεις της γλώσσας κάθε επίθετο είναι αναφορά του κατάλληλου ρήματος μέσω του δείκτη «κύριο\_μέρος\_του».

Στην δεύτερη κατηγορία των επιθέτων περιλαμβάνονται αυτά που συσχετίζονται σημασιολογικά και μορφολογικά με τα ουσιαστικά. Τυπικά, ένα τέτοιο επίθετο έχει ένα παρόμοιο ρόλο με αυτό των τροποποιητών-ουσιαστικών και συναρτήσεων σαν τους ταξινομητές. Το WordNet διατηρεί ένα ξεχωριστό αρχείο με δείκτες στα ουσιαστικά που αντιστοιχούν και το αρχείο αυτό περιέχει γύρω στα 2,832 synsets επιθέτων.

Όσο αφορά τα επιρρήματα, τα περισσότερα παράγονται από επίθετα μέσω της προσθήκης καταλήψεων. Τα επιρρήματα στο WordNet συνδέονται με τις ερμηνείες των επιθέτων, με βάση τους δείκτες που έχουν σαν ερμηνεία «προέρχομαι\_από». Εξαιτίας του ότι πολλά επίθετα έχουν πολλές διαφορετικές ερμηνείες και τα παραγόμενα επιρρήματα συνήθως κληρονομούν την ερμηνεία του βασικού επιθέτου στο WordNet αυτά συνδέονται.

Το WordNet περιέχει 3,242 synsets επιρρημάτων. Η σημασιολογική οργάνωση των επιρρημάτων είναι απλή και ακριβής. Δεν υπάρχει δεντρική δομή, όπως στα ουσιαστικά και στα ρήματα, ούτε υπάρχει μια δομή ομαδοποίησης όπως στα επίθετα.

## Η οργάνωση των ρημάτων

Για τον σκοπό της οργάνωσης του λεξικού των ρημάτων σαν ένα δίκτυο συσχετίσεων, η καλύτερη λύση ήταν να χωριστεί το λεξικό σε σημασιολογικά τμήματα, ώστε να μπορούν να ανακαλυφτούν σχέσεις που οργανώνουν τα ρήματα και τις ερμηνείες τους.

Αυτά χωρίστηκαν σε 15 ομάδες (ρήματα αισθήσεως, νοήματος, επαφής, επικοινωνίας, ανταγωνισμού, αλλαγής, νόησης, κατανάλωσης, δημιουργίας, συναισθήματος, κατοχής, σωματικής φροντίδας και λειτουργίας, ρήματα που αναφέρονται σε κοινωνική συμπεριφορά και αλληλεπίδραση, στατικά, βιοθητικά και ρήματα ελέγχου) είναι επαρκή για να εξυπηρετήσουν όλα τα synsets των ρημάτων.

Η διαίρεση του λεξικού των ρημάτων σε σημασιολογικά τμήματα, όχι μόνο επιτρέπει την οργάνωση μεγάλων μεγέθους δεδομένων, αλλά και επιβάλει την απουσία μιας απλής ρίζας για ολόκληρο το λεξικό. Έτσι υιοθετεί την ιδέα για την δημιουργία περισσότερων ριζών για τα 15 σημασιολογικά τμήματα.

Τα ρήματα όπως ακριβώς και τα ουσιαστικά και τα επίθετα οργανώνονται σε σύνολα συνωνύμων. Το WordNet δεν λαμβάνει υπόψη του ρήματα, τα οποία μπορεί να έχουν παρόμοια ερμηνεία, αλλά διαφέρουν στην χρήση έτσι, αυτά ανήκουν στο ίδιο synset. Οι ιδιωματικές εκφράσεις και οι μεταφορικές έννοιες με τις λογοτεχνικές τους ερμηνείες περιλαμβάνονται και αυτές στα κατάλληλα synsets.

Όσο αφορά τις λεξιλογικές σχέσεις που υπάρχουν ανάμεσα στα ρήματα και στα synsets, διαθέσιμες είναι η συνεπαγωγή, η σημασιολογική αντίθεση και η σχέση αιτίας αποτελέσματος. Η συνεπαγωγή μεταξύ δύο ρημάτων συμβαίνει όταν η πρόταση που περιέχει το ένα από τα δύο ρήματα λογικά συνεπάγει την πρόταση με το άλλο ρήμα. Η σχέση αυτή δεν είναι αμφίδρομη και η άρνηση της σχέσης αντιστρέφει την πορεία της συνεπαγωγής. Αυτή η σχέση έρχεται σε συμφωνία με την σχέση των μερονύμων στα ουσιαστικά.

Η εξέταση των υπονύμων (hyponyms) και των υπερνύμων (superordinate) των ρημάτων φανερώνουν ότι η λεξικολογία περιλαμβάνει πολλά είδη σημασιολογικής επεξεργασίας, μέσω των διαφορετικών τμημάτων. Στο WordNet, τα διαφορετικά είδη επεξεργασίας που διακρίνει ένα υπόνυμο ρήματος από το υπέρνυμό του έχει συγχωνευτεί σε αυτό που ονομάζεται τροπόνυμο (troponymy). Ο σχηματισμός που το υλοποιεί είναι «το ρήμα1 είναι το ρήμα2 κατά ένα συγκεκριμένο τρόπο» και έτσι μπορούν να συσχετίσουν τα υπέρνυμα τους, σύμφωνα με πολλές σημασιολογικές διαστάσεις. Τα τροπόνυμα αποτελούν ένα είδος συνεπαγωγής. Κάθε τροπόνυμο v1 ενός πιο γενικού ρήματος v2, συνεπάγει και το v2. Η ιεραρχία των ρημάτων κατασκευάζεται σύμφωνα με τις σχέση των τροπονύμων

Ο δείκτης με ετικέτα «αντίθετα» στο λεξικό των ρημάτων, στην πραγματικότητα εκφράζει μια πολύπλοκη σχέση περικλείοντας αρκετά διαφορετικού υποτύπους της σημασιολογικής αντίθεσης. Τέλος στο WordNet υπάρχει και ο κατάλληλος δείκτης που συνδέει τα λεξιλογικά ζευγάρια αιτίας-αποτελέσματος.

Η πολυσημία, όσα αφορά τα ρήματα είναι πολύ μεγαλύτερη από αυτή των ουσιαστικών, παρόλο που το πλήθος τους είναι πολύ μικρότερο. Τα πιο πολυσύχναστα ρήματα είναι πολυσήμαντα και η ερμηνεία τους εξαρτάται έντονα από τα ουσιαστικά που συνδέονται.

Η ανάλυση των ρημάτων σε σημασιολογικούς και εννοιολογικούς όρους μπορεί να ανακαλύψει πολλές συντακτικές ιδιότητες. Έχει παραπορηθεί ότι τα ρήματα που

αναγνωρίζονται από όρους συγκεκριμένων ερμηνευτικών συστατικών έχουν την τάση να μοιράζονται τις συντακτικές συμπεριφορές.

### Σχεδιασμός και Υλοποίηση της λεξιλογικής βάσης και του λογισμικού αναζήτησης

Το WordNet αποτελείται από τέσσερα μέρη, τα λεξιλογικά αρχεία πηγής, τους Grinder που αποτελούν το λογισμικό που μετατρέπει αυτά τα αρχεία στην λεξιλογική βάση, την λεξιλογική βάση του WordNet και διάφορα εργαλεία λογισμικού που χρησιμοποιούνται για την πρόσβαση της βάσης. Στην συνέχεια θα μελετηθούν αναλυτικά αυτά τα μέρη.

Καταρχήν, τα αρχεία πηγής του WordNet έχουν γραφτεί από λεξικογράφους και είναι αποτέλεσμα μιας αναλυτικής ανάλυσης της λεξιλογικής σημασιολογίας. Στο WordNet, τα ουσιαστικά, τα ρήματα, τα επίθετα, τα επιφρήματα οργανώνονται σε σύνολα συνωνύμων, τα οποία στην συνέχεια οργανώνονται σε σύνολα λεξιλογικών αρχείων πηγής σύμφωνα με την συντακτική κατηγορία και άλλα οργανωτικά κριτήρια.

Ένα λεξιλογικό αρχείο περιλαμβάνει synsets από μόνο μια συντακτική κατηγορία και κάθε synsets αποτελείται από συνώνυμες λέξεις, δείκτες συσχέτισης, μια γρήγορη επεξήγηση και παραδείγματα. Χαρακτηριστικό είναι το παρακάτω παράδειγμα

**Σύνολο συνώνυμων:** {car, auto, automobile, machine, motorcar}  
**Ορισμός:** “*wheeled motor vehicle; usually propelled by an internal combustion engine.*”

Η μορφή των λέξεων στην βάση του WordNet είναι είτε μια ορθογραφική αναπαράσταση ατομικών λέξεων, είτε συμβολοσειρές ατομικών λέξεων που έχουν μια απλή ερμηνεία. Η μορφή των λέξεων μπορεί να αυξηθεί με απαραίτητες πληροφορίες, για την σωστή επεξεργασία και ερμηνεία των δεδομένων. Μια λέξη σε ένα synsets αναπαρίσταται από την ορθογραφική μορφή της λέξης, την συντακτική κατηγορία, το σημασιολογικό τομέα και ένα αναγνωριστικό αριθμό. Όλα αυτά τα χαρακτηριστικά σχηματίζουν το «κλειδί της ερμηνείας» που είναι μοναδικό για κάθε ζευγάρι λέξης/ερμηνείας στην βάση.

Όσο αφορά τις σχέσεις που περιλαμβάνει το WordNet στοχεύουν στην διάκριση του από τις λεξιλογικές πηγές και τα λεξικά. Κατασκευάζοντας ο λεξικογράφος ένα synset αναγνώρισε κάποιες σχετικές σχέσεις και τις κωδικοποίησε. Οι σχέσεις αναπαρίστανται, είτε μέσω σημασιολογικών, είτε μέσω λεξιλογικών δεικτών. Ένας δείκτης κατασκευάζεται καθορίζοντας την μορφή της λέξης από το synset στόχο, ακολουθώντας το ένα κόμμα και ολοκληρώνοντας το ένα σύμβολο που αντιστοιχεί στην επιθυμητή σχέση.

Παρακάτω παρουσιάζονται συνοπτικά οι σχέσεις που διαθέτει το WordNet:

**Antonym:** front → back  
**Attribute:** benevolence → good (noun to adjective)  
**Pertainym:** alphabetical → alphabet (adjective to noun)  
**Similar:** unquestioning → absolute  
**Cause:** kill → die  
**Entailment:** breathe → inhale  
**Holonym:** chapter → text (part-of)  
**Meronym:** computer → cpu (whole-of)  
**Hyponym:** tree → plant (specialization)  
**Hypernym:** fruit → apple (generalization)

Κάθε synset που αποτελείται από ρήματα έχει μια λίστα από σχηματισμένες προτάσεις (sentence frames), που προσδιορίζει τους τύπους των απλών προτάσεων, που μπορούν να χρησιμοποιηθούν από το κάθε synset. Η λίστα από τα frames των ρημάτων μπορεί να είναι αυστηρά ένα απλό ρήμα στο σύνολο συνωνύμων ή μπορεί να είναι κατάλληλα σε όλα τα ρήματα του. Τα frames των προτάσεων των ρημάτων μπορεί να αναπαρασταθούν από τους αριθμούς των frames και στα λεξιλογικά αρχεία πηγής και στην βάση.

Οι συμβολοσειρές στα λεξιλογικά αρχεία πηγής ακολουθούν τους παρακάτω συντακτικούς κανόνες που αναγνωρίζονται από τους Grinder σαν synsets. (1) Κάθε synset ξεκινάει και τελειώνει με αγκύλη { και } αντίστοιχα. (2) Κάθε synset περιέχει μια λίστα από μια ή περισσότερες μορφές λέξεων ακολουθούμενες από κόμμα και κενό. (3) Για να κωδικοποιηθούν οι σχέσεις, οι λίστες των λέξεων ακολουθούνται από μια λίστα δεικτών συσχέτισης με βάση το εξής συντακτικό: μια λέξη από το synset στόχο, ακολουθεί ένα κόμμα, ακολουθεί το σύμβολο του δείκτη συσχέτισης και ένα κενό. (4) Για τα synsets των ρημάτων, τα frames ακολουθούνται από μια λίστα από τους αριθμούς των frames χωρισμένη με κόμματα. Η λίστα των frames των ρημάτων ακολουθούν όλους τους δείκτες συσχέτισης. (5) Για την κωδικοποίηση των λεξιλογικών σχέσεων μια μορφή λέξης και οι λίστες του βήματος 3 και 4 εσωκλείονται σε [...] (6) Η σύντομη περιγραφή εσωκλείεται σε παρενθέσεις και (7) Για την κωδικοποίηση των ομάδων των επιθέτων, κάθε μέρος της ομάδας χωρίζεται από τα άλλα μέρη των ομάδων με μια γραμμή που περιέχει μόνο παύλες. Κάθε ομάδα εσωκλείεται σε [...].

Όσο αφορά το σύστημα του αρχείου είναι το μέρος που διατηρούνται τα λεξιλογικά αρχεία πηγής και βασίζεται στο Unix Revision Control System για την διαχείριση πολλαπλών εκδόσεων των αρχείων κειμένων. Οι λόγοι που δημιουργείται το αρχείο συστήματος είναι για να επιτρέπει την επαναδημιουργία της βάσης σε κάθε έκδοσης του WordNet, να διατηρεί τις αλλαγές των λεξιλογικών αρχείων, να εμποδίζει τους λεξικογράφους από το να κάνουν συγκρουόμενες αλλαγές σε ίδια αρχεία, καθώς και για να εξασφαλίζει ότι είναι σε θέση να παράγει μια ενημερωμένη έκδοση της βάσης. Τα προγράμματα στο αρχείο του συστήματος είναι UNIX shell scripts που περικλείονται εντολές με τέτοιο τρόπο, ώστε να διατηρείται ο επιθυμητός έλεγχος στα λεξιλογικά αρχεία πηγής, ενώ παρέχουν μια απλή διεπαφή για τους λεξικογράφους.

Η «καρδιά» του συστήματος WordNet είναι η λεξιλογική βάση. Το ASCII format έχει επιλεγεί για τα αρχεία της βάσης, ώστε να διευκολύνει την πρόσβαση από διαφορετικές γλώσσες προγραμματισμού. Έχουν αναπτυχθεί αρκετές διεπαφές για να ανταποκρίνονται στις ερωτήσεις του χρήστη, στην απλή εξαγωγή των δεδομένων από την βάση και στην τροποποίησή τους για να επιστραφούν στον χρήστη. Απλά εργαλεία έχουν κατασκευαστεί και για την αναζήτηση αντικειμένων της βάσης. Είναι σημαντικό να τονιστεί ότι για να δημιουργηθεί η βάση απαιτεί την χρήση των Grinder, ώστε να επεξεργαστούν όλα τα λεξιλογικά αρχεία της εισόδου την ίδια στιγμή.

Οι διαφορετικές ερμηνείες για μια δοθείσα λέξη οργανώνονται ιεραρχικά από την περισσότερη στην λιγότερη συχνή στην χρήση ερμηνεία. Η σειρά καθορίστηκε από την χρήση τους σε κάποιες συγκεκριμένες συλλογές. Σε περίπτωση που δεν υπάρχει η παραπάνω ιεραρχία, η σειρά είναι τυχαία.

Ένα σημαντικό γλωσσολογικό γεγονός για τα πνευματικά λεξικά είναι ότι μερικές λέξεις είναι πιο σύνηθες από άλλες. Η οικειότητα μιας λέξης επηρεάζει μια ευρεία κλίμακα στην απόδοση μεταβλητών, όπως η ταχύτητα διαβάσματος, η ταχύτητα κατανόησης, η ευκολία ανάκλησης και η πιθανότητα να χρησιμοποιηθεί. Η αγνόηση αυτής της μεταβλητής σε μια λεξιλογική βάση που προσπαθεί να ανακλά φυσικές γλωσσολογικές αρχές είναι αδιανότητο. Έτσι, για να γίνεται η ενσωμάτωση των διαφορών της οικειότητας(familiarity) στο WordNet,

σχετίζεται με κάθε μορφή λέξης ένας δείκτης που υποδηλώνει την οικειότητα. Σαν δείκτη οικειότητας στο WordNet χρησιμοποιείται η πολυσημία. Αυτό το μέτρο μπορεί να καθοριστεί από το λεξικό που διαβάζεται από την μηχανή με το ακόλουθο τρόπο: εάν η τιμή του δείκτη είναι 0, τότε οι λέξεις δεν ανήκουν στο λεξικό. Αντίθετα οι λέξεις ανήκουν στο λεξικό, όταν η τιμή είναι 1 ή μεγαλύτερη, σύμφωνα με το πλήθος των ερμηνειών που έχει η λέξη. Η τιμή του δείκτη οικειότητας είναι διαθέσιμη για κάθε λέξη, σε κάθε συντακτική κατηγορία.

Η βάση κατασκευάζεται περιοδικά από τα λεξιλογικά αρχεία εισόδου. Είναι απαραίτητη η ανακατασκευή των synsets από μια μορφή κατάλληλη για την σύνταξη σε ένα σύνολο αλληλοσυσχετισμένων αρχείων αναζήτησης, στα οποία επιλύονται οι δείκτες συσχέτισης και υπολογίζεται το πλήθος των ερμηνειών και της πολυσημίας. Η διαίρεση των synsets σε συντακτικές κατηγορίες διατηρείται, όμως τα δεδομένα κάθε συντακτικής κατηγορίας αναπαρίστανται από δύο αρχεία: το ευρετήριο και το αρχείο των δεδομένων.

Η πληροφορία του κάθε synset αρχίζει από ένα συγκεκριμένο byte offset στο αρχείο δεδομένων και συνεχίζει μέχρι να συναντήσει το χαρακτήρα καινούριας γραμμής. Οι δείκτες συσχέτισης αναπαρίστανται στα αρχεία δεδομένων μέσω των συμβόλων των δεικτών και την διεύθυνση των synset. Τα ευρετήρια είναι αλφαριθμητικές λίστες όλων των μορφών λέξεων του WordNet και αποτελούν τα κύρια μονοπάτια των αρχείων δεδομένων. Κάθε είσοδος-λέξης στο αρχείο του ευρετηρίου περιλαμβάνει μια λίστα από τις διευθύνσεις των synsets για όλες τις ερμηνείες της λέξης. Η αναζήτηση όλων των ερμηνειών μιας λέξης περιλαμβάνει την εκτέλεση δυαδικής αναζήτησης της βασικής μορφής της λέξης στο αρχείο του ευρετηρίου για μια συντακτική κατηγορία και την μετακίνηση στην λίστα των διευθύνσεων των συνόλων συνωνύμων. Για κάθε διεύθυνση, το synset διαβάζεται από το αντιπροσωπευτικό αρχείο δεδομένων. Οι δείκτες ανιχνεύονται από το αρχείο δεδομένων κινούμενοι από το synset εισόδου στον αντικειμενικό στόχο μέσω των διευθύνσεων των synsets.

Ένα αρχείο ευρετηρίου αρχίζει με διάφορες γραμμές που περιλαμβάνουν σημειώσεις για τα δικαιώματα του δημιουργού, τον αριθμό της έκδοσης και την άδεια. Οι υπόλοιπες γραμμές περιλαμβάνουν πληροφορίες, όπως την μορφή της λέξης, τον μετρητή πολυσημίας, μια λίστα από σύμβολα για όλους τους δείκτες, που χρησιμοποιούνται στα synsets, που περιέχουν την λέξη, την λίστα των διευθύνσεων των synsets, για κάθε ερμηνεία της λέξης.

Το αρχείο δεδομένων ξεκινάει και αυτό με την ίδια εισαγωγή, όπως το ευρετήριο και στην συνέχεια, ακολουθεί μια λίστα από λεξιλογικά ονόματα αρχείων. Κάθε γραμμή που μένει αντιπροσωπεύει ένα σύνολο συνωνύμων και είναι μια κωδικοποίηση της πληροφορίας που έχει εισαχθεί από τους λεξικογράφους με δείκτες που επιλύονται στις διευθύνσεις των synsets.

Το ευρετήριο των ερμηνειών παρέχει στο WordNet μια εναλλακτική μέθοδο πρόσβασης στα synsets και στις λέξεις της βάσης. Το αρχείο είναι μια ταξινομημένη λίστα όλων των λέξεων όλων των synsets της βάσης. Μια δυαδική αναζήτηση στο αρχείο ανακτά γρήγορα πληροφορίες, όπως την διεύθυνση και τον αριθμό της ερμηνείας που αντιπροσωπεύει την ερμηνεία της λέξης, καθώς και επιτρέπει την άμεση πρόσβαση σε ένα synset μέσω των διευθύνσεων των synsets. Το ευρετήριο αυτό δεν χρησιμοποιείται από το λογισμικό αναζήτησης του WordNet, αλλά είναι χρήσιμο σε εφαρμογές που αναζητούν συγκεκριμένες ερμηνείες.

Το επόμενο μέρος του WordNet που θα περιγράφεται είναι οι Grinder που έχουν πολύπλευρη χρησιμότητα, με πρωταρχικό σκοπό την μεταγλώττιση των λεξιλογικών αρχείων εισόδου σε τέτοια μορφή που να διευκολύνει την μηχανή ανάκτησης της πληροφορίας του. Επίσης χρησιμοποιείται σαν ένα εργαλείο επαλήθευσης, ώστε να εξασφαλιστεί η συντακτική ακεραιότητα των λεξιλογικών αρχείων.

Το πρόγραμμα Grinder είναι ένας μεταγλωττιστής για synsets που εκτελείται με πολλαπλά περάσματα. Το πρώτο πέρασμα προσπαθεί να βρει όσα περισσότερα συντακτικά και λάθη δομής είναι δυνατόν. Συντακτικά λάθη μπορούν να συμβούν, όταν το αρχείο εισόδου αποτύχει να συμμορφωθεί με τις γραμματικές προδιαγραφές εισόδου. Τα λάθη δομής αναφέρονται στους δείκτες συσχέτισης που δεν μπορούν να επιλυθούν για κάποιο λόγο, συνήθως γιατί έχουν γίνει τυπογραφικά λάθη, όπως δημιουργία δείκτη σε ένα αρχείο που δεν υπάρχει. Στο δεύτερο πέρασμα επιλύει τους λεξιλογικούς και σημασιολογικούς δείκτες. Οι δείκτες που καθορίζονται σε κάθε synset εξετάζονται στην σειρά με σκοπό την εύρεση του αντικειμενικού στόχου κάθε δείκτη. Στο επόμενο πέρασμα εξετάζει την λίστα των μορφών των λέξεων, αναθέτοντας στην καθεμία το πλήθος της πολυσημίας για κάθε συντακτική κατηγορία. Το τελευταίο πέρασμα δημιουργεί τα αρχεία της βάσης. Αρχικά, καθορίζεται το byte offset κάθε synset και οι εσωτερικές δομές των δεδομένων, που ενημερώνονται μέσω των διευθύνσεων των synset, με άμεσο αποτέλεσμα την δημιουργία των δεδομένων, του ευρετήριο και του ευρετηρίου ερμηνειών.

Όπως είναι φυσικό, για να έχει πρόσβαση ο χρήστης στις πληροφορίες της βάσης, απαιτείται μια διεπαφή. Το λογισμικό της διεπαφής του WordNet δημιουργεί γρήγορα τις απαντήσεις στις απαιτήσεις του χρήστη. Οι πληροφορίες του αποθηκεύονται σε τέτοια μορφή που να είναι ασήμαντες για ένα συνηθισμένο χρήστη. Στόχος της διεπαφής είναι να παρέχει στο χρήστη μια ποικιλία τρόπων για να ανακτήσει και να εκθέσει τις λεξιλογικές πληροφορίες. Κάποιες διεπαφές χρηστών στο WordNet αναπτύχθηκαν σε παραθυρικά περιβάλλοντα, όπως το X Windows, Microsoft Windows, όμως υπάρχει και η εναλλακτική διεπαφή στην γραμμή εντολών.

Η διαδικασία της ανάκτησης των πληροφοριών είναι ίδια, ανεξάρτητα του τύπου αναζήτησης που απαιτείται. Στο πρώτο βήμα διαβάζεται η είσοδος της λέξης που αναζητείται από το κατάλληλο αρχείο του ευρετηρίου, ώστε να εξασφαλίσει τις διευθύνσεις των synset για όλες τις ερμηνείες της λέξης. Στην συνέχεια, σε καθένα από τα synsets του αρχείο δεδομένων αναζητείται η πληροφορία που απαιτείται. Η αναζήτηση είναι μια πολύπλοκη διαδικασία, αφού στην πραγματικότητα κάθε synset που περιέχει την λέξη που αναζητείται, περιέχει επίσης δείκτες με άλλα synsets του αρχείου δεδομένων, που ίσως πρέπει να ανακτηθούν και να εμφανιστούν, ανάλογα τον τύπο αναζήτησης.

Οι διεπαφές των χρηστών στο WordNet βασίζονται σε μια βιβλιοθήκη συναρτήσεων για να διαχειριστούν τα αρχεία της βάσης και τα περιεχόμενά τους. Ο δομημένος και ευέλικτος σχεδιασμός της βιβλιοθήκης παρέχει έναν εύκολο προγραμματισμό της διεπαφής. Η βιβλιοθήκη παρέχει ένα περιεκτικό σύνολο συναρτήσεων, εκτελώντας αναζητήσεις και ανακτήσεις, μετασχηματίζοντας την μορφολογία και διαθέτοντας κάποιες συναρτήσεις γενικού σκοπού. Οι συναρτήσεις μπορούν να κατηγοριοποιηθούν στις πολύπλοκες (εκτέλεση πραγματικών δεδομένων ανάκτησης, μετασχηματισμούς μορφολογίας και τροποποίηση των αποτελεσμάτων αναζήτησης για την αναπαράστασή τους στον χρήστη), στις χαμηλού επιπέδου (πρόσβαση στις λεξιλογικές πληροφορίες του ευρετηρίου και των αρχείων δεδομένων) και στις χρήσιμες (επιτρέπουν τον χειρισμό των συμβολών αναζήτησης, ανοίγουν και κλείνουν αρχεία της βάσης).

Η βασική συνάρτηση αναζήτησης λαμβάνει σαν είσοδο τέσσερις παραμέτρους: την μορφή της λέξης, την συντακτική κατηγορία, έναν κωδικό που αντιπροσωπεύει τον τύπο αναζήτησης και τον αριθμό της ερμηνείας. Οι περισσότερες αναζητήσεις αντιστοιχούν σε δείκτες συσχέτισης. Πολλές φορές, ο χρήστης αναζητά μια ιεραρχική αναζήτηση ή μια αναζήτηση που διασχίζει μόνο ένα επίπεδο του δέντρου. Όταν ένα synset ανακτηθεί από την βάση, τότε μορφοποιείται όπως ορίστηκε στον τύπο αναζήτησης σε ένα buffer εξόδου. Ο buffer που έχει το αποτέλεσμα περιέχει όλα τα μορφοποιημένα synsets για όλες τις ερμηνείες που

απαιτούνται από την λέξη που αναζητείται και με μια άλλη συνάρτηση τυπώνει τα περιεχόμενά του.

Το WordNet περιέχει μια βιβλιοθήκη που επεξεργάζεται τις μορφολογικές συναρτήσεις, Morphy, που χειρίζεται μια μεγάλη κλίμακα μορφολογικών μετασχηματισμών. Η Morphy χρησιμοποιεί δύο διαδικασίες στην προσπάθειά της να μετατρέψει μια λέξη στην μορφή που βρίσκεται στην βάση του WordNet: κανόνες αποκοπής (detachment), που περιλαμβάνουν λίστες ελέγχου κλιτών καταλήψεων βασισμένες στην συντακτική κατηγορία, με σκοπό τον χωρισμό τους από την λέξη και λίστες εξαίρεσης για κάθε συντακτική κατηγορία, που αποτελούνται από ταξινομημένες λίστες κλιτών μορφών λέξεων ακολουθούμενες από μια ή περισσότερες βασικές μορφές και χρησιμοποιούνται κυρίως στις ανώμαλες κλίσεις. Η Morphy επιθεωρεί δύο αντικείμενα την μορφή της λέξης και την συντακτική κατηγορία.

Οι απλές λέξεις επεξεργάζονται εύκολα. Η Morphy κοιτάει πρώτα για την μορφή της λέξης στις λίστες εξαίρεσης που αντιστοιχούν στο μέρος του λόγου. Εάν υπάρχει επιστρέφεται η βασική μορφή. Εάν δεν υπάρχει εκεί, εφαρμόζονται οι κανόνες αποκοπής για την συντακτική κατηγορία, εάν ταιριάζουν οι καταλήψεις προστίθεται η αντίστοιχη κατάληξη και στην συνέχεια το WordNet μελετά εάν η λέξη που προκύπτει βρίσκεται στην βάση και επιστρέφει την μορφή της.

Αντίθετα, μια σύνθετη λέξη ή μια φράση έχει αρκετό ενδιαφέρον για να μετασχηματιστεί σε μια βασική μορφή που να υπάρχει στο WordNet. Γενικά, μόνο βασικές λέξεις αποθηκεύονται στο WordNet ακόμα κι αν αυτές σχηματίζουν σύνθετα. Οι φράσεις ρημάτων που περιέχουν προθέσεις είναι πολύ δύσκολο να επεξεργαστούν. Όπως και στις απλές λέξεις γίνεται πρώτα αναζήτηση στην λίστα των εξαιρέσεων. Αν η φράση δεν υπάρχει, η Morphy καθορίζει εάν υπάρχει μια πρόθεση. Εάν υπάρχει γίνεται μια προσπάθεια να βρεθεί η βασική μορφή κατασκευάζοντας μια συμβολοσειρά με τις βασικές μορφές των λέξεων χωρίς την πρόθεση. Εάν δεν υπάρχει, τότε βρίσκονται οι βασικές μορφές κάθε λέξης της φράσης.

Με συμβολοσειρές που έχουν προστεθεί παύλες είναι εξίσου δύσκολο να γίνει αναζήτηση στην βάση. Όταν η Morphy σπάει μια συμβολοσειρά σε λέξεις ελέγχει και για κενά και για παύλες σαν χαρακτήρες αρχής και τέλους. Παράλληλα, αναζητά και τα μέρη της συμβολοσειράς και τα διαγράφει εάν δεν υπάρχει ακριβής ταίριασμα στην βάση.

### 5.3 Εφαρμογές του WordNet

#### Κατασκευή Semantic Concordances

Με τη χρήση του WordNet (version 1.5) δημιουργήθηκε το λεξικό που χρησιμοποίησαν στο σημασιολογικό αλφαριθμητικό λεξικό (semantic concordances). Σαν semantic concordances ορίστηκε μια συλλογή βασισμένη σε κείμενα και ένα λεξικό, έτσι ώστε να συνδέεται κάθε θεμελιώδη λέξη του κείμενο με την κατάλληλη έννοια(sense) στο λεξικό.

Τα semantic concordances κατασκευάστηκαν χειρονακτικά, ώστε το λογισμικό που θα χρησιμοποιηθεί να είναι σχεδιασμένο ειδικά γι' αυτόν τον σκοπό. Το ConText είναι ένα σημασιολογικό tagging πρόγραμμα που χρησιμοποιήθηκε για να σχολιάσει τον αγγλικό πεζό λόγο με τις έννοιες του WordNet. Αυτό αρχικά έδειχνε το κείμενο και τόνιζε κάθε λέξη του αντικειμενικού σκοπού στην σειρά. Κάτω από το κείμενο, παρουσιαζόταν το σύνολο των έννοιών του WordNet για τις λέξεις που τόνιζε. Στην συνέχεια ένας άνθρωπος διάβαζε το

κείμενο και επέλεγε την καταλληλότερη ερμηνεία για την λέξη. Αποθήκευε τις ερμηνείες μαζί με το κείμενο, παρουσιάζοντας τις συνδέσεις μεταξύ αυτών και της βάσης του WordNet.

Όταν η λέξη ήταν μονοσήμαντη, δεν παρουσιάζεται κανένα ιδιαίτερο πρόβλημα για τον άνθρωπο που καθορίζει τις ερμηνείες. Απλώς έκανε τον έλεγχο και άφηνε το σύστημα να φτιάξει τις συνδέσεις. Στην περίπτωση όμως που η λέξη είναι πολυσήμαντη, ο άνθρωπος έπρεπε να καθορίσει την σωστή ερμηνεία. Παρόλο που το ποσοστό των λέξεων που έχουν πολλές ερμηνείες στο WordNet είναι σχετικά μικρό (μόνο 18%), το πρόβλημα αυτό αντιμετωπίζεται συχνά στις συλλογές, κυρίως γιατί οι λέξεις που χρησιμοποιούνται πιο συχνά έχουν πολλές διαφορετικές ερμηνείες.

Παράλληλα με την δημιουργία των semantic concordances, η διαδικασία του tagging βοηθάει και στην βελτίωση της κάλυψης ελλείψεων του WordNet. Είναι δυνατόν να μην υπάρχουν κάποιες λέξεις ή κάποιες ερμηνείες λέξεων που συναντιούνται στα κείμενα, έτσι οι λεξικογράφοι προσαρμόζουν αυτά τα νέα στοιχεία στο WordNet.

#### *Απόδοση και σιγουριά στον σχολιασμό σημασιολογικής αποστολής*

Σε συνέχεια των παραπάνω, κατά την αντιστοίχηση της χρήσης της λέξης με την ανταπόκριση της σημασία της από το WordNet, έρχεται η προσπάθεια να εντοπίζεται ο βαθμός του πόσο σωστά η σημασία από το σύστημα αντιπροσωπεύει το επιθυμητό νόημα. Οι διαθέσιμες έννοιες σχεδόν πάντα περιλαμβάνουν μια ερμηνεία που ταιριάζει με την δοθείσα χρήση, όμως σε κάποιες περιπτώσεις υπάρχουν κοντινές σημασιολογικά ερμηνείες, που είναι δύσκολο να ξεχωρίσει κανείς, ενώ σε άλλες περιπτώσεις οι διαφορετικές ερμηνείες είναι διαχωρίσιμες, είναι πολύ εύκολο να αναγνωριστεί το ταίριασμα. Σημαντικό φαίνεται ότι όσο πιο δύσκολη είναι η επιλογή της σωστής ερμηνείας, τόσο απίθανο είναι να βρεθεί η επιθυμητή ερμηνεία και ένας υψηλός βαθμός σιγουριάς για την επιλογή της σωστής ερμηνείας.

Όσο αφορά τα ουσιαστικά η ερμηνείας τους είναι σχετικά σταθερή κι έτσι δεν υπάρχει το παραπάνω πρόβλημα. Στην περίπτωση, όμως, των επιθέτων και των επιρρημάτων υπάρχει η τάση η ερμηνεία τους να αλλάζει ανάλογα με την σημασία της λέξης που προσδιορίζουν. Η ερμηνεία των επιρρημάτων εξαρτάται από την θέση τους στην πρόταση, ενώ η ερμηνεία των ρημάτων εξαρτάται από τα ουσιαστικά που βρίσκονται στις προτάσεις.

Στην ανάλυση του κατά πόσο συμφωνεί η πραγματική ερμηνεία μιας λέξης με αυτήν που της αποδίδεται, όσο αυξάνεται η πολυσημία, τόσο χειρότερα είναι τα αποτελέσματα, αφού είναι δυσκολότερο να επιλέξει κανείς την επιθυμητή ερμηνεία, ιδιαίτερα εάν οι ερμηνείες του WordNet είναι πολύ κοντινές.

Όταν οι ερμηνείες των λέξεων κατατάσσονται με σειρά αυτής που χρησιμοποιείται πιο συχνά τις περισσότερες φορές επιλέγουν την σωστή ερμηνεία. Όμως, όταν η σειρά είναι τυχαία, τα πράγματα διαφέρουν κι αυτό γιατί η επιλογή είναι πιο δύσκολη από το να επιλέξεις την πρώτη σε συχνότητα, που είναι πιο γενική και αποδίδει μια ισορροπία στην συλλογή. Σε κάθε περίπτωση, επιθυμείται ο βαθμός σιγουριάς να αντανακλά την συμφωνία των αποτελεσμάτων και για να γίνει αυτό στην περίπτωση της τυχαίας σειράς, θα πρέπει να εξεταστεί προσεκτικά ολόκληρη η λεξιλογική είσοδος, χωρίς να είναι αναγκαίο να υιοθετηθεί το αποτέλεσμα της πιθανότητα, που δηλώνει ότι η πιο συχνή ερμηνεία είναι πιθανόν η καλύτερη επιλογή.



Είναι ξεκάθαρο ότι οι πρωταρχικές γνώσεις της γλώσσας, όπως λεξιλόγια, γραμματικές έχουν έναν πολύ σημαντικό ρόλο, παρά τις αλλαγές μεγάλης κλίμακας που έχουν γίνει. Οι προσπάθειες που γίνονται στην απόκτηση τέτοιων γνώσεων αυτόμata δεν είναι ώριμες και πολλές πετυχημένες προσπάθειες απαιτούν τουλάχιστον αρχικά σχολιασμό από χειρονακτικά δεδομένα.

Παρόλο που μερικές μέθοδοι που βασίζονται σε συλλογές περιγράφονται σαν «αμιγής» στατιστικές, δεν ισχύει αφού εμπεριέχουν πρωταρχική γνώση. Παραδείγματος χάρη, στην αλγεβρική δομή που αποτελεί την βάση για τα πιθανοτικά μοντέλα, στους μη ποσοτικούς όρους γίνονται κάποιες υποθέσεις για την φύση των δεδομένων. Παράλληλα αποδεικνύουν το παραπάνω και τα Markov μοντέλα που χρησιμοποιούνται στο tagging του μέρους του λόγου.

Οι τάξεις των λέξεων είτε βασίζονται στην γνώση, είτε αντλούνται από την διανομή, είναι πολύ σημαντικές στις εργασίες των γλωσσών που βασίζονται σε συλλογές, αφού οι συλλογές κειμένων είναι μικρές για να παρέχουν αρκετή πληροφορία για την χρήση μη πολυσύχναστων λέξεων. Μια λύση είναι να εξάγονται συμπεράσματα για την συμπεριφορά των λέξεων αυτών με βάση κάποιων λέξεων που τους μοιάζουν. Για την δημιουργία και την χρήση των τάξεων παρόμοιων λέξεων χρησιμοποιείται η λεξιλογική διανομή του κείμενο της συλλογής. Προϋπόθεση είναι η σχετικότητα των λέξεων να αντανακλάται από την ομοιότητα στην διανομή του περιβάλλον τους (συναφών εκφράσεων).

Για την αναπαράσταση της ανάλυσης της διανομής απαραίτητο είναι ο καθορισμός των λέξεων που ενδιαφέρουν τον χρήστη, η παρουσίαση των λέξεων που παρατηρούνται στο περιβάλλον τους όταν χρησιμοποιούνται στο κείμενο και κάθε συνδυασμός των παραπάνω, καθώς και ο υπολογισμός της συχνότητας που μια λέξη ανήκει σε αντίστοιχο περιβάλλον.

Αυτή η ανάλυση όμως έχει αρκετά μειονεκτήματα, καθώς οι τάξεις που προκύπτουν μέσω των τεχνικών διανομής δεν αναγνωρίζονται από κανένα είδος συμβολικών ετικετών (label). Παράλληλα, επικεντρώνεται στα token και όχι στις έννοιες των λέξεων. Έτσι, εάν μια λέξη σχετίζεται με ένα απλό σημείο στο χώρο της σημασιολογίας, τότε οι συνιστώσες των πολυσήμαντων λέξεων θα συγχέουν την ατομική τους ερμηνεία έχοντας μεγαλύτερη επιρροή στην πιο συχνή ερμηνεία. Επιπρόσθετα, προβλήματα σχετικότητας αναπτύσσονται όταν δημιουργείται η σημασιολογία των τάξεων των λέξεων από τα αποτελέσματα των μεθόδων διανομής και είναι υπερευαίσθητα σε συγκεκριμένα παρατηρούμενα δεδομένα.

Όμως, όλα τα παραπάνω προβλήματα το WordNet τα διευθετεί. Αρχικά με αυτόματη διαδικασία καθορίζονται οι συμβολικές ετικέτες μέσω του σχήματος των μοναδικών synsets. Παράλληλα, δεν είναι δύσκολο να δημιουργηθεί και μια λογική συμφωνία για την ευανάγνωση από τον άνθρωπο συμβολική περιγραφή των synsets. Όσο αφορά τις διακρίσεις των ερμηνειών των λέξεων, η λύση για την ταξινόμηση είναι η βάση. Οι βασικές σχέσεις ταξινόμησης του δικτύου, συνώνυμα και υπόνυμα μπορεί να καθορίσουν σημασιολογικές περιγραφές σε όρους με διαφορετική σημασιολογική θεωρία και κληρονομική συνεπαγωγή. Τέλος, το WordNet σχεδιάστηκε για να αντανακλά γενική γνώση και όχι την ιδιοσυγκρασία συγκεκριμένων πεδίων ορισμού (domain).

Όμως, το WordNet είναι μια αποθήκη γνώσης για γλωσσικές σχέσεις και δεν κάνει χρήση της γλώσσας, όποτε έχει περιορισμούς. Εν τέλει, κατανοώντας την έννοια των λέξεων και χρησιμοποιώντας την λεξιλογική γνώση σε εμπειρικά συστήματα, απαιτεί προσοχή στο

περιβάλλον της λέξης. Αυτή η πληροφορία δεν μπορεί να βρεθεί στα λεξικά παρά μόνο στην ανάλυση της διανομής(distributional analysis).

Για την χρήση του WordNet με τις distributional μεθόδους, θα πρέπει να οριστεί ένα μοντέλο πιθανοτήτων. Χρησιμοποιώντας την προσέγγιση της διανομής δεν είναι σίγουρο ότι θα υιοθετεί το πιθανοτικό μοντέλο, αλλά έχει αρκετά πλεονεκτήματα. Χαρακτηρίζοντας το είδος της πληροφορίας που χρησιμοποιούν οι όροι, το πιθανοτικό μοντέλο καθορίζει τις πτυχές τις ταξινόμησης που παίζουν σημαντικό ρόλο στο μοντέλο και αναγνωρίζει τις πληροφορίες που θεωρούνται σχετικές.

Δεν είναι εύκολο να εργαστεί κανείς με το WordNet στο πιθανοτικό περιβάλλον, γιατί ενώ ο χώρος του μοντέλου καθορίζεται από λέξεις, η ταξινόμηση οργανώνεται σε όρους της ερμηνείας των λέξεων. Στο WordNet δεν είναι εύκολο να αναγνωριστεί το στοιχείο που αναπαριστά το αποτέλεσμα του πειράματος, αφού κάθε λέξη ανήκει σε πολλές κλάσεις στην ταξινόμηση και δεν είναι ξεκάθαρο ποια τάξη/κλάση θα επιλεγεί. Βασικό πρόβλημα στο πιθανοτικό μοντέλο, περιλαμβάνοντας την ταξινόμηση, είναι να συσχετίσει κανείς ευδιάκριτα token με μη ευδιάκριτες έννοιες.

Αυτό το πρόβλημα είναι δυσδιάστατο. Η πρώτη του διάσταση αφορά τις διφορούμενες έννοιες και η δεύτερη το επίπεδο της αφηρημένης έννοιας. Υπάρχουν διάφοροι τρόποι για να επιλύσει κανείς αυτά τα προβλήματα. Ένας είναι να καθοριστεί ένα απλό επίπεδο αφηρημένων εννοιών και να προσδιοριστεί η κατηγοριοποίηση του WordNet για τους πιθανοτικούς σκοπούς, ώστε αυτοί να είναι οι όροι των τάξεων σε αυτό το επίπεδο, αλλά δυστυχώς η λύση αυτή δεν διορθώνει το πρόβλημα των διφορούμενων εννοιών. Άλλες είναι η ισοπέδωση της ταξινομίας, προσαρμόζοντας το WordNet σε χωρισμό ολόκληρης της συλλογής των synsets σε σύνολα ισοπεδωμένων κατηγοριών, κ.τ.λ.

*Συνδυασμός του Local context και τις ομοιότητας του WordNet για την αναγνώριση των ερμηνειών*

Η αναγνώριση των εννοιών είναι η αντιστοιχία μεταξύ των λέξεων του κειμένου και της κατάλληλης ερμηνείας στο λεξικό. Οι έρευνες για αυτόματη αναγνώριση των εννοιών είναι βασισμένες σε συλλογές και τυπικά εκπαιδεύονται ένα στατιστικό ταξινομητή σε περιβάλλοντα που περιέχουν μια πολυσήμαντη λέξη με γνωστή ερμηνεία. Βασισμένος σε αυτήν την γνώση, ο ταξινομητής προσδιορίζει την ερμηνεία σε νέες εμφανίσεις πολυσήμαντων λέξεων.

Το βασικό πρόβλημα αυτής της προσέγγισης είναι η έλλειψη πολλών δεδομένων εκπαίδευσης, αφού για να είναι αξιόπιστα θα πρέπει να καθορίζεται η έννοια των δεδομένων εκπαίδευσης χειρονακτικά και η διαδικασία αυτή είναι χρονοβόρα και επίπονη.

Στόχος είναι να χρησιμοποιηθεί το WordNet και χρησιμοποιώντας τις σημασιολογικές του σχέσεις, να αυξήσει το αποδοτικό μέγεθος των δεδομένων εκπαίδευσης.

Η έρευνα για την αναγνώριση των εννοιών στις συλλογές εστιάζονται είτε στο θέμα του περιβάλλοντος(*topical context*), είτε στο τοπικό περιβάλλον(*local context*), είτε στο συνδυασμό τους. Το topical context καθορίζεται από τις θεμελιώδης λέξεις που είναι πιθανόν να συνυπάρχουν με την δοθείσα ερμηνεία της πολυσήμαντης λέξης. Το local context αποτελείται από συντακτικές και σημασιολογικές υποδείξεις στην άμεση γειτονιά της πολυσήμαντης λέξης.

Το topical context είναι πολύ αποδοτικό στην αναγνώριση των ερμηνειών των ομωνύμων που δεν σχετίζονται σημασιολογικά, π.χ. bank που σημαίνει και τράπεζα και όχθη. Παράλληλα και η τοπική πληροφορία βοηθάει στην αναγνώριση των ερμηνειών και μάλιστα

έχει βρεθεί ότι είναι αξιόπιστη σε περιπτώσεις αυτόματης αναγνώρισης των ερμηνειών. Ενώ, η topical context κοιτάζει σε μεγάλο παράθυρο γύρω από το αντικειμενικό στόχο, η local context εστιάζει σε πολύ μικρότερο παράθυρο.

Συμπερασματικά, επειδή ο local context ταξινομητής είναι πολύ ευαίσθητος στα προβλήματα μικρών δεδομένων, ένας τρόπος για να αυξηθεί ο τοπικός χώρος της εκπαίδευσης είναι να βρεθεί μια μέθοδος που να συνδυάζει την τοπική συντακτική πληροφορία με την σημασιολογική πληροφορία του WordNet. Μεγεθύνοντας τον ταξινομητή με την παραπάνω διαδικασία παρουσιάζεται μια μικρή βελτίωση στην απόδοση, ειδικά όταν η εκπαίδευση γίνεται σε μικρά σύνολα.

### *Χρήση του WordNet στην Ανάκτηση Πληροφοριών*

Στόχος της ανάκτησης πληροφορίας είναι να εντοπίσει κείμενα γραμμένα σε φυσική γλώσσα, των οποίων το περιεχόμενό τους να ικανοποιεί την ανάγκη του χρήστη να βρει κάποια συγκεκριμένη πληροφορία.

Χρησιμοποιώντας το WordNet είναι δυνατόν να ενισχυθεί η πρόσβαση κειμένων από συλλογές. Στόχος είναι να εκμεταλλευτεί η γνώση που κωδικοποιεί το WordNet, ώστε να βελτιωθούν οι επιδράσεις των συνωνύμων(synonyms) και των ομογράφων(homographs), που παρουσιάζουν τα κείμενα των συστημάτων ανάκτησης, που χρησιμοποιούν το ταίριασμα των λέξεων (word matching).

Υπάρχουν δύο εκδοχές: η πρώτη χρησιμοποιεί τα synsets του WordNet (σύνολο συνωνύμων) για να αναπαραστήσει το περιεχόμενο των κειμένων και στην δεύτερη το WordNet αποτελεί την πηγή των λέξεων που θα προστεθούν στην επερώτηση του χρήστη. Και στις δύο εκδοχές η ανικανότητα να επιλύσει αυτόματα την ερμηνεία των λέξεων με πολλές σημασίες περιορίζει τα πλεονεκτήματα της χρήσης του WordNet.

### *Προσωρινό Indexing μέσω Λεξιλογικών Αλυσίδων*

Καθώς η αποθήκευση των πληροφοριών κινείται με μεγάλη ταχύτητα μακριά από την παραδοσιακή δομή των εγγράφων και τις ιεραρχικές βάσεις, σε βάσεις που περιέχουν μη δομημένα δεδομένα, είναι απαραίτητο να αναπτυχθούν καινούργιες τεχνικές για την αποθήκευση, την ευρετηρίαση και την ανάκτηση αυτών των πληροφοριών. Οι πηγές αυτών των δεδομένων είναι ταινίες, βιντεοσκοπημένες διαλέξεις, μουσική, φωτογραφίες κ.τ.λ. που είτε είναι αδόμητες, είτε δομούνται μικρά τμήματά τους. Δεν υπάρχει μεγάλο όγκος έρευνας που να διαχειρίζεται και να δημιουργεί αυτόματη ευρετηρίαση σε τέτοια δεδομένα.

Έχει αναπτυχθεί μια μέθοδος ευρετηρίασης συνδιαλέξεων με βάση το θέμα. Αυτό είναι ένα μέρος εργασίας που έχει σαν σκοπό την σχεδίαση και την δημιουργία πρωτότυπων εργαλείων για τον μετασχηματισμό των βίντεο-συνδιαλέξεων σε αποθήκη οργανωτικής μνήμης και γνώσης, κάνοντάς το με μικρή σχετικά εμπιστοσύνη στις σημασιολογικές αναλύσεις των συνδιαλέξεων.

Η κατασκευή ενός ευρετηρίου τέτοιων πληροφοριών θα πρέπει να είναι βασισμένη σε δεδομένα που είναι εύκολα να διεξαχθούν, όπως είναι το θέμα. Αυτό μπορεί να γίνει χρησιμοποιώντας την αναγνώριση του λόγου και τις λεξιλογικές αλυσίδες, αναγνωρίζοντας τα θέματα μέσου του ακουστικού μέρους των συνδιαλέξεων.

Η ανάπτυξη του αλγορίθμου LexTree φτιάχνει ένα σύνολο από λεξιλογικά δέντρα, τα οποία χαρακτηρίζουν την σημασιολογική δομή του κειμένου. Όπως φαίνεται στη παρακάτω εικόνα, ο σχεδιασμός του λεξιλογικού δέντρου περιλαμβάνει τα εξής στάδια: (1) το σύστημα επιλέγει τις λέξεις του κειμένου που θα περιλαμβάνονται στο δέντρο, (2) καθορίζει τη καλύτερη σχέση μεταξύ της επιλεγμένης λέξης και του κόμβου του δέντρου για όλα τα δέντρα, (3) επιλέγει τα δέντρα που συνδέονται με την επιλεγμένη λέξη, σε περίπτωση που δεν υπάρχει δημιουργείται ένα καινούριο του οποίου η ρίζα είναι η λέξη αυτή. Στα περισσότερα βήματα του αλγορίθμου χρησιμοποιούνται πληροφορίες από το WordNet.

```

REPEAT
  1. READ next word from input file
  2. IF word is a possible compound word component THEN
    2.1   IF (word has a noun sense in WordNet) and
          (compound word buffer is empty) THEN
      2.1.1   PUSH word in compound word buffer
    ELSE
      2.1.2   IF (compound word buffer is not empty) THEN
        2.1.2.1   attempt to join word and item in compound
                  word buffer
      END IF
    END IF
  ELSE
    2.2   IF (word is not a general word) and
          (word has a noun sense in WordNet) THEN
      2.2.1   FOR all trees within a suitable span
        2.2.1.1   FOR all words in the tree
          CHECK WordNet for relations between word
          and tree word
        2.2.1.2   CHOOSE best relation
      END FOR
      2.2.1.2   IF (best relation > connection threshold)
        THEN
          ADD word and best relation to tree.
        END IF
      END FOR
      2.2.2   IF no relations are found THEN
        2.2.2.1   MAKE word a node in a new tree
      END IF
    END IF
  END IF
UNTIL end of input file

```

*COLOR-X* \*: Χρήση Γνώσης από το WordNet για εννοιολογική μοντελοποίηση

\*Conceptual Linguistically based Object-Oriented Representation Language for Information and Communication Systems

Σε αυτήν την εφαρμογή το WordNet αποκαλύπτει το ρόλο του στην γλωσσολογία, βασισμένη στο περιβάλλον της εννοιολογικής μοντελοποίησης. Μια μέθοδος μοντελοποίησης που αναπτύχθηκε ήταν το COLOR-X που αποτελείται από αρκετές τεχνικές γραφικής μοντελοποίησης, προερχόμενες από την αλληλεπίδραση της χρήσης της γνώσης του WordNet. Το WordNet αποτελεί μια πηγή επαναχρησιμοποιήσιμης γνώσης που μπορεί να χρησιμοποιηθεί κατά την διάρκεια της εννοιολογικής μοντελοποίησης και να διασφαλίσει εάν τα αποτελέσματα είναι σωστά. Παράλληλα, το WordNet υποστηρίζει την δημιουργία διαδικασιών φυσικής γλώσσας, κατορθώνοντας έτσι μια συνοχή τόσο ανάμεσα στα μοντέλα όσο και ανάμεσα στο μοντέλο και το πραγματικό πρόβλημα.

Το COLOR-X χρησιμοποιείται για να γεφυρώσει το χάσμα μεταξύ της ανάλυσης των απαιτήσεων και της σχεδίασης φράσεων λογισμικού στην διαδικασία ανάπτυξης λογισμικού. Κατανοώντας τις απαιτήσεις με επαρκή τρόπο, είναι δυνατόν να δημιουργηθούν αυτόματα περισσότερα μοντέλα τεχνικού σχεδιασμού.

Αυτή η εργασία είναι μέρος των Linguistically based Information and Communication Systems που ερεύνα τον τρόπο που μπορεί να χρησιμοποιηθεί η γλωσσολογική γνώση. Σε αυτά τα συστήματα παίζει σημαντικό ρόλο, ο έλεγχος του προβλήματος που αφορά την ερμηνεία των λέξεων. Τα συστήματα προσφέρουν αποτελεσματική αποθήκευση, επεξεργασία και μεταφορά

δεδομένων, όμως η μεταφορά των δεδομένων είναι χρήσιμη μόνο εάν ο αποστολέας και ο παραλήπτης συμφωνούν με την ερμηνεία των λέξεων.

Η ενσωμάτωση της γλωσσολογικής γνώσης στο ενοιολογικό μοντέλο είναι απαραίτητη για να χρησιμοποιεί λέξεις που εμφανίζονται στο μοντέλο με συνέπεια. Παράλληλα, τέτοιες τεχνικές δίνουν περισσότερη εκφραστική δύναμη, καθώς και επιτρέπουν ευκολότερα τη δημιουργία φυσικών προτάσεων ώστε να υπάρχει ανάδραση στο σύστημα.

Η ενσωμάτωση αυτή είναι και οριζόντια και κατακόρυφη. Για την επίτευξη του ακριβή βαθμού κατακόρυφης ενσωμάτωσης, δηλαδή την ενσωμάτωση μεταξύ διαφορετικών φράσεων της διεργασίας του λογισμικού, θα πρέπει να εστιαστεί στην ανταπόκριση των μοντέλων σχεδίασης και της ανάλυσης της πληροφορίας. Η επαναχρησιμοποίηση της γνώσης από το WordNet και η επιβεβαίωση από τα συστήματα σχεδίασμού είναι πιθανόν να συμβούν μέσω της γλωσσολογικής βάσης των μοντέλων.

## 6. Περιγραφή του συστήματος ανάκτησης που υλοποιήθηκε

Το σύστημα που αναπτύχτηκε είναι βασισμένο στο διανυσματικό μοντέλο IR (Vector Space Model), όπου η αναπαράσταση των κειμένων και των επερωτήσεων γίνεται μέσω διανύσματων. Οι συντεταγμένες του διανύσματος καθορίζονται από τα βάρη που αναθέτονται στους όρους που περιέχει το κάθε κείμενο ή την κάθε επερώτηση και είναι διάστασης όσο το πλήθος των όρων (keywords) ολόκληρης της συλλογής.

Πριν το καθορισμό όμως αυτών των όρων, είναι απαραίτητο να διαβαστούν τα κείμενα και οι επερωτήσεις της συλλογής και να κρατηθούν μόνο οι όροι που είναι χρήσιμοι για την ανάκτηση της πληροφορίας. Έτσι καθώς διαβάζεται το κάθε κείμενο και η κάθε επερώτηση της συλλογής, γίνεται προσπάθεια να προσδιοριστούν οι όρους(tokens) που τα αποτελεί. Στην συνέχεια από κάθε όρο αποκόπτονται τα σημεία στίξης, εάν υπάρχουν στην αρχή και στο τέλος του όρου.

Ταυτόχρονα, οι όροι εξετάζονται, ώστε αυτοί που ανήκουν στην λίστα (ArrayList) των τετριμμένων λέξεων (stopwords) να αφαιρεθούν από την πληροφορία που κρατείται για κάθε κείμενο και επερώτηση. Στην περίπτωση που ο όρος που εξετάζεται κάθε φορά δεν βρίσκεται στην λίστα των stopwords, αποκόπτεται η κατάληξη του, ώστε να μετασχηματιστεί στη ρίζα του. Με άλλα λόγια, εκτελείται η διαδικασία του Stemming, σύμφωνα με τον αλγόριθμο του Porter (Stemmer.java), ώστε να βελτιώνεται η ανάκτηση.

Παράλληλα, με χρήση της μεθόδου WSD των Μαυροειδή, Τσατσαρώνη, Βαζιργιάννη, Theobald και Weikum [3] αποσαφηνίζονται οι ερμηνείες των όρων, σύμφωνα με τις σχέσεις που παρέχει το WordNet και κυρίως των υπερνύμων. Η έκδοση του WordNet που χρησιμοποιήθηκε ήταν η 1.7.1. Συνοπτικά, η μελέτη αυτή στοχεύει στην απεικόνιση των γειτονικών λέξεων σε senses που είναι συμπαγής (μέτρο ομοιότητας για senses) στον ιεραρχικό θησαυρό. Εξαιτίας του ότι μόνο τα ουσιαστικά έχουν από μόνα τους μια ερμηνεία, σε αντίθεση με τα ρήματα που εκφράζουν τις σχέσεις μεταξύ των λέξεων, κάνει Disambiguation μόνο αυτά. Αυτή η μέθοδος WSD επιλέχτηκε, επειδή αποδεικνύεται ότι παρέχει υψηλό precision, σημαντικό κριτήριο για την Ανάκτηση Πληροφορίας, παρόλο που έχει χαμηλό recall. (SentenceWSD.java, Phrase.java, SensesCompactnessWrapperThreshold.java, JWNLWrapper.java, Candidate.java)

Αφού έχουν βρεθεί οι «χρήσιμοι» όροι που απαρτίζουν την απαραίτητη πληροφορία για κάθε κείμενο και επερώτηση, προσδιορίζονται τα βάρη που είναι απαραίτητα για την δημιουργία των διανυσμάτων τους. Τα βάρη αυτά υπολογίζονται βάση του γινομένου των τιμών

του TF και του IDF. Η ποσότητα TF αντιστοιχεί στην συχνότητα του όρου σε κάθε κείμενο ή επερώτηση και υπολογίζεται καθώς διαβάζεται το κείμενο. Η ποσότητα IDF βρίσκεται με βάση τον παρακάτω τύπο:  $\log \frac{1 + (\text{πλήθος των εγγράφων της συλλογής})}{\text{πλήθος εγγράφων που περιέχουν τον συκεκριμένο όρο}}$ . Κρατώντας αυτήν την πληροφορία (όρο και τιμή) σε μια δομή (hashtable), κατορθώνετε να δημιουργηθεί το λεξικό της συλλογής. Συνάμα διατηρείται σε αντίστοιχη δομή και η κανονικοποιημένη του μορφή. (IDFSpace.java)

Πιο συγκεκριμένα λοιπόν, για το κάθε κείμενο και για την κάθε επερώτηση διατηρείται μια δομή (hashtable) η οποία έχει τα εξής χαρακτηριστικά: (1) το αναγνωριστικό τους αριθμό, (2) μια δομή (hashtable) που κρατάει την συχνότητα του κάθε όρου (TF) στο συγκεκριμένο κείμενο ή επερώτηση, (3) μια δομή (hashtable) που κρατάει το κανονικοποιημένο TF, δηλαδή το TF κάποιου όρου, διαιρεμένο με το πλήθος όλων των συχνοτήτων, όλων των όρων στο κείμενο, (4) μια δομή που κρατάει το TF επί το IDF για κάθε όρο και (5) μια μεταβλητή με το πλήθος όλων των συχνοτήτων όλων των όρων του κειμένου. (Document.java). Κρίνεται απαραίτητο, όλα τα κείμενα και όλες οι επερωτήσεις της συλλογής να κρατούνται σε μια δομή (hashtable) στην οποία κρατείται το όνομα του κειμένου ή της επερώτησης και η παραπάνω μορφή δομής Document, για την εύκολη επεξεργασία τους. (CorpusReader.java)

Έχοντας συγκεντρώσει όλη την παραπάνω πληροφορία και υπολογίζοντας το κανονικοποιημένο tf και το idf, καθώς και το γινόμενο τους, είναι δυνατόν να υπολογιστεί η ομοιότητα των κειμένων με την επερώτηση, βάση του συνημίτονου της γωνίας που δημιουργείται ανάμεσα στα αντίστοιχα διανύσματα. Διατηρείται μια δομή (Results.java) που κρατάει το αναγνωριστικό αριθμό της επερώτησης, ένα hashtable που περιέχει το συνημίτονο της επερώτησης με κάθε κείμενο της συλλογής, ώστε να φανερώνεται η ομοιότητα της επερώτησης με το κάθε κείμενο/έγγραφο και ένα hashtable που αποθηκεύει τις τιμές της ανάκλησης (recall) και της ακρίβειας (precision), ώστε να γίνεται σωστή η αξιολόγηση της ανάκτησης της πληροφορίας, αφού έχει γίνει η κατάταξη των αποτελεσμάτων με βάση τον βαθμό ομοιότητα τους με την επερώτηση. (Main.java)

Ολοκληρώνοντας, είναι σημαντικό να διευκρινιστεί ότι στο σύστημα η επέκταση των όρων τόσο των επερωτήσεων, όσο και των κειμένων γίνεται με τρεις διαφορετικούς τρόπους. Για κάθε όρο προστίθενται είτε μόνο τα senses που προκύπτουν από την μέθοδο WSD που αναλύθηκε παραπάνω, είτε όλα τα υπέρνυμα που περιέχει το WordNet, είτε 2,4,6, ή 8 υπέρνυμα από το WordNet.

Εξαιτίας του ότι δεν είναι γνωστό το πλήθος των υπερνύμων που χρειάζονται, ώστε να επιτευχθεί η καλύτερη απόδοση στην ανάκτηση, καθώς και ότι για διαφορετικά dataset, υπάρχει διαφορετικό «βέλτιστο» πλήθος υπερνύμων, κρίνεται απαραίτητη η εύρεση μιας μεθόδου που να μην παρέχει ad-hoc Ανάκτηση Πληροφορίας. Η λύση στο παραπάνω πρόβλημα έρχεται μέσω της τροποποίησης των βαρών των senses των όρων και των υπερνύμων τους.

Για τον προσδιορισμό αυτών των βαρών χρησιμοποιείται η δημοσίευση των Xing, Ng, Jordan, Russell “Distance metric learning, with application to clustering with side-information”. Ο αλγόριθμος που προτάσσει, χρησιμοποιεί περιορισμούς (τύπου must-link, cannot-link) για να κατασκευάσει μια μετρική απόσταση στο Rn. Πιο συγκεκριμένα, βρίσκει έναν πίνακα A ανάμεσα στην ευκλείδεια απόσταση. Στην περίπτωση που ο A είναι διαγώνιος, μαθαίνει τα βάρη των διαστάσεων, ώστε να έρχονται κοντά τα must-link και μακριά cannot-link.

Προσαρμόζοντας τον παραπάνω αλγόριθμο στην Ανάκτηση της Πληροφορίας και παράγοντας τα διανύσματα των κειμένων, σύμφωνα με το tf\*idf των υπέρνυμων και των senses

της συλλογής που διαθέτει το κάθε κείμενο, οι παραπάνω περιορισμοί (τύπου must-link, cannot-link) των κειμένων προκύπτουν από την ανάδραση του χρήστη. Χαρακτηριστικό είναι το παρακάτω παράδειγμα: υποθέτοντας ότι τα αποτελέσματα του συστήματος ανάκτησης είναι τα {k1,k2,k3} και τα αποτελέσματα από την ανάδραση του χρήστη είναι τα {k1,k3}, οι περιορισμοί που προκύπτουν είναι Must-link = {(k1,k3),(k1,k1),(k3,k3)} και Cannot-link = {(k1,k2),(k2,k3)}. Εξαιτίας του ότι τα ζητούμενα ζεύγη δημιουργούνται για όλες τις επερωτήσεις της συλλογής μαζί, σε περίπτωση που ένα ζευγάρι κειμένων ανήκει για μια επερώτηση στα must-link και για μια άλλη στα cannot-link δεν συμπεριλαμβάνεται σε κανένα από τα δύο ζευγάρια. Τα αποτέλεσμα του Αλγόριθμου Xing είναι η εύρεση βαρών στις διαστάσεις του ευκλείδειου χώρου και η χρήση τους στην επέκταση της επερώτησης.

## 7. Πειράματα

### 7.1 Συλλογές

Οι συλλογές που χρησιμοποιήθηκαν στην εκτέλεση των πειραμάτων ήταν: η *MEDLINE*, η *CISI*, η *CACM* και η *CRANFIELD*. Παρακάτω παρουσιάζονται κάποια από τα γενικά χαρακτηριστικά τους.

Η συλλογή *MEDLINE* αποτελείται από άρθρα που προέρχονται από κάποιο ιατρικό περιοδικό. Το πλήθος των κειμένων είναι 1033 και το πλήθος των επερωτήσεων που εκτελούνται και διατίθενται οι απαντήσεις είναι 30.

Η συλλογή *CACM* αποτελείται από συλλογές τίτλων και περιλήψεων από το περιοδικό “Communications of the ACM” κατά την περίοδο 1958-1979. Τα κείμενα αυτά καλύπτουν ένα ευρύ πεδίο της λογοτεχνίας της επιστήμης των υπολογιστών, εξαιτίας του ότι αυτό το περιοδικό ήταν το πιο δημοφιλές στο πεδίο αυτό. Το πλήθος των κειμένων της συλλογής είναι 3204 και το πλήθος των επερωτήσεων που εκτελούνται και διατίθενται οι απαντήσεις είναι 64. Εκτός από τα κείμενα, η συλλογή περιέχει πληροφορίες για τα ονόματα των συγγραφέων, τις ημερομηνίες έκδοσης, λέξεων που έχουν αποκοπεί οι καταλήψεις από τους τίτλους και τις περιλήψεις, κατηγορίες που προέρχονται από το σχήμα της ιεραρχικής κατηγοριοποίησης, άμεσες αναφορές ανάμεσα στα άρθρα, βιογραφικές συνδέσεις και πλήθος κειμένων που αναφέρονται σε ζευγάρια άρθρων.

Η συλλογή *CRANFIELD* αποτελείται από κείμενα που έχουν ως θέμα την αεροναυτική. Το μέγεθος της συλλογής είναι 1400 κείμενα και οι επερωτήσεις που διαθέτει είναι 365, αλλά μόνο για τις 225 από αυτές υπάρχουν οι σχετικές απαντήσεις.

Τέλος, όσο αφορά την συλλογή *CISI* αποτελείται από 1460 κείμενα που επιλέχθηκαν από προηγούμενη συλλογή με όνομα Small [1] από το Ινστιτούτο Επιστημονικής Πληροφορίας. Τα κείμενα που επιλέχτηκαν είναι αυτά που είχαν τις περισσότερες αναφορές στην προηγούμενη συλλογή. Τα κείμενα της συλλογής περιλαμβάνουν πληροφορίες για τα ονόματα των συγγραφέων, για λέξεις που έχουν αποκοπεί οι καταλήξεις από τους τίτλους και τις περιλήψεις και το πλήθος κειμένων που αναφέρονται σε κάποια ζευγάρια άρθρων. Το πλήθος των επερωτήσεων που προσφέρει είναι 112.

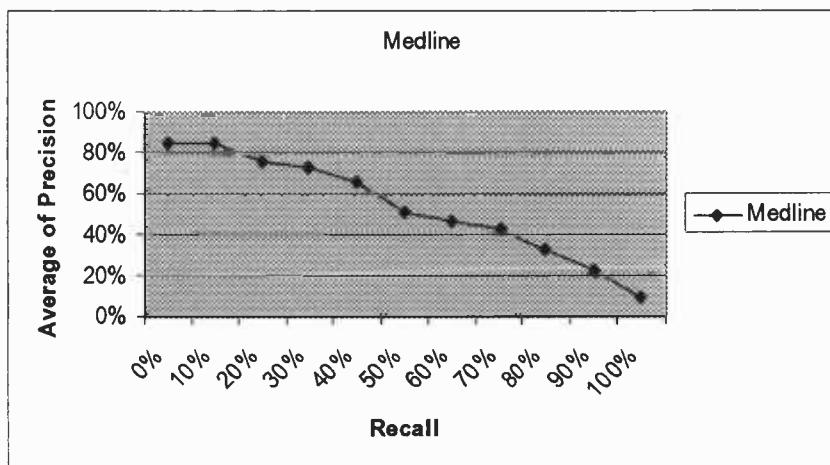
## 7.2 Περιγραφή Πειραμάτων - Εμφάνιση Αποτελεσμάτων

### 1<sup>o</sup> Πείραμα:

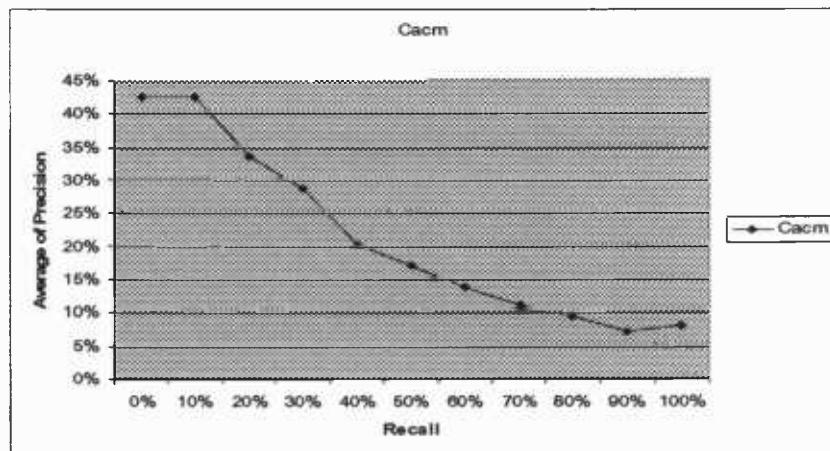
Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, γίνεται η ανάθεση των βαρών στις λέξεις (όρους) των κειμένων, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζονται το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω:

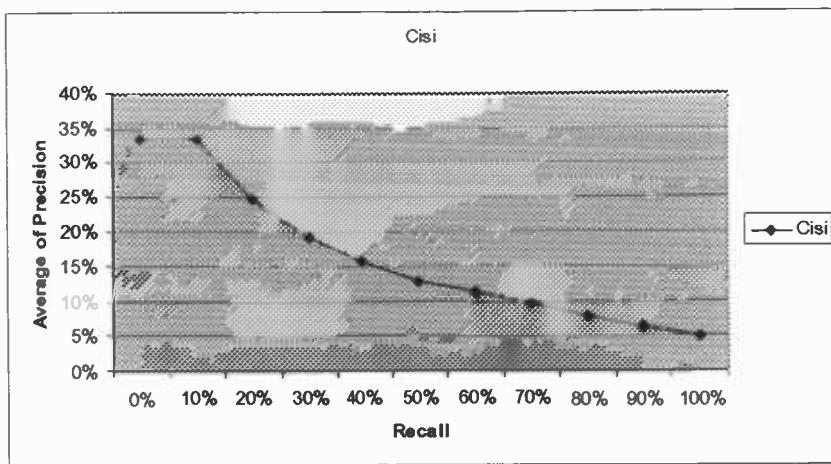
Recall	Precision
0%	84%
10%	84%
20%	76%
30%	72%
40%	65%
50%	51%
60%	46%
70%	43%
80%	33%
90%	22%
100%	9%



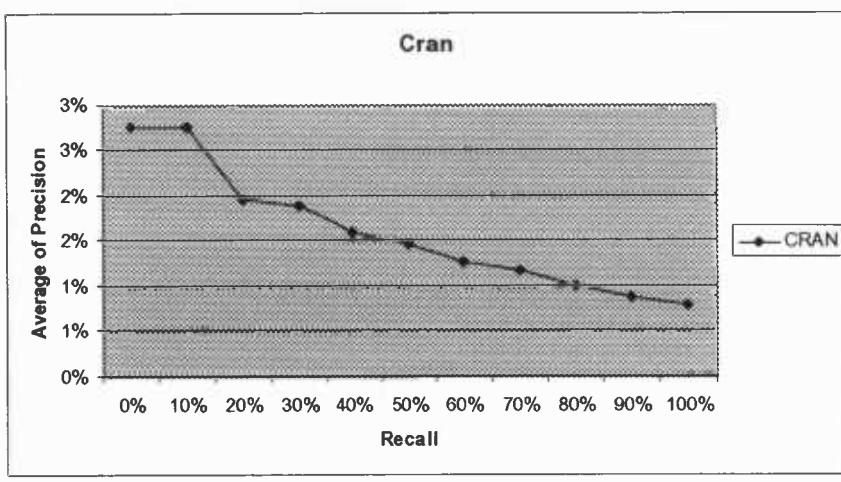
Recall	Precision
0%	42%
10%	42%
20%	33%
30%	29%
40%	20%
50%	17%
60%	14%
70%	11%
80%	9%
90%	7%
100%	8%



Recall	Precision
0%	33%
10%	33%
20%	25%
30%	19%
40%	16%
50%	13%
60%	11%
70%	10%
80%	8%
90%	6%
100%	5%



Recall	Precision
0%	3%
10%	3%
20%	2%
30%	2%
40%	2%
50%	1%
60%	1%
70%	1%
80%	1%
90%	1%
100%	1%

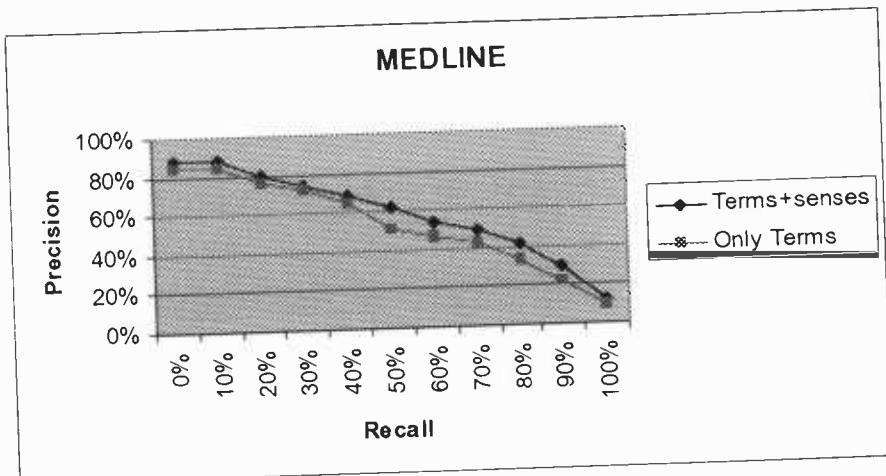


## 2<sup>ο</sup> Πείραμα:

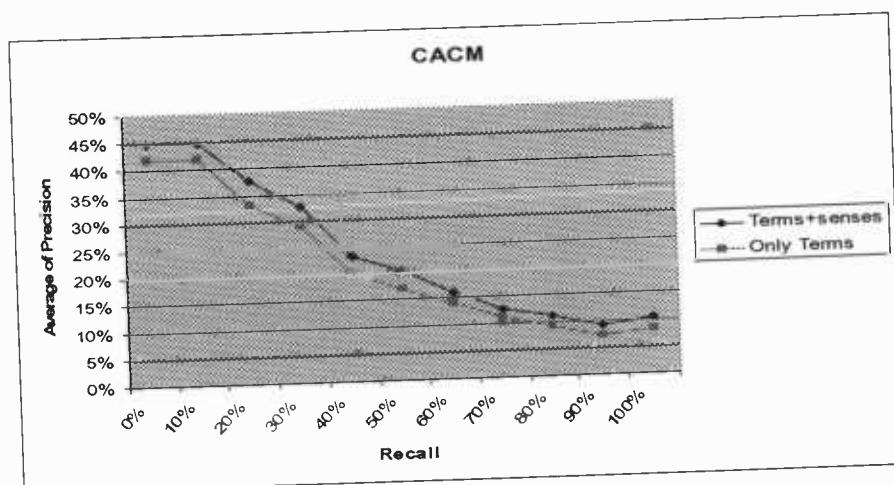
Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων και προκύπτει η αναπαράσταση των κειμένων με βάση τους όρους των κειμένων και τα senses τους, όπως αυτά προκύπτουν από την μέθοδο WSD της δημοσίευσης[3]. Στην συνέχεια, γίνεται η ανάθεση των βαρών στις λέξεις (όρους) των κειμένων και στα senses τους, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερωτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με τα αποτελέσματα του παραπάνω πειράματος:

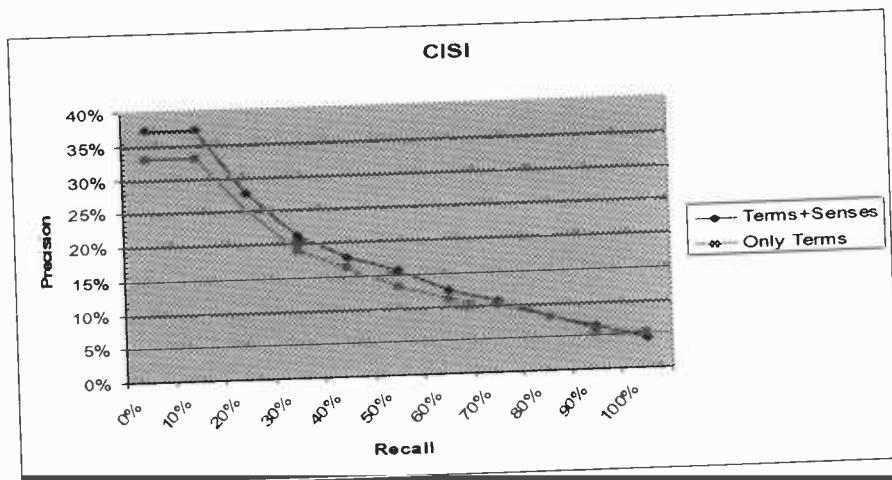
Recall	Precision
0%	88%
10%	88%
20%	80%
30%	74%
40%	69%
50%	62%
60%	54%
70%	49%
80%	41%
90%	30%
100%	12%



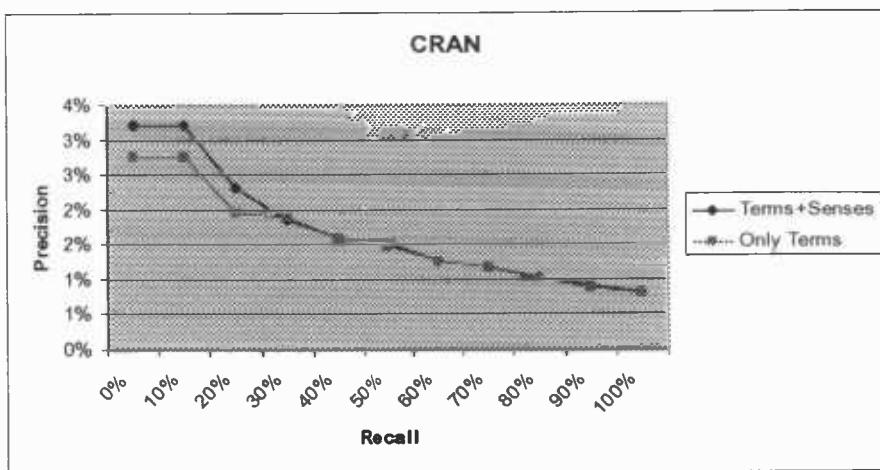
Recall	Precision
0%	45%
10%	45%
20%	38%
30%	33%
40%	23%
50%	20%
60%	16%
70%	13%
80%	11%
90%	9%
100%	10%



Recall	Precision
0%	37%
10%	37%
20%	28%
30%	21%
40%	18%
50%	15%
60%	12%
70%	10%
80%	8%
90%	6%
100%	4%



Recall	Precision
0%	3%
10%	3%
20%	2%
30%	2%
40%	2%
50%	2%
60%	1%
70%	1%
80%	1%
90%	1%
100%	1%

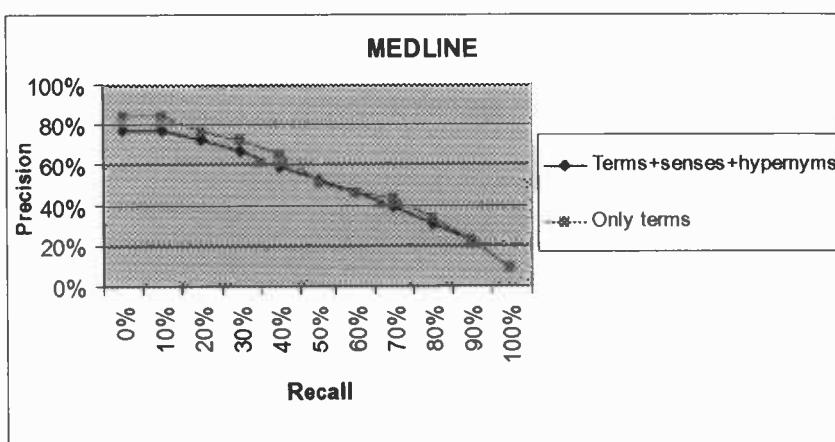


### 3<sup>ο</sup> Πείραμα:

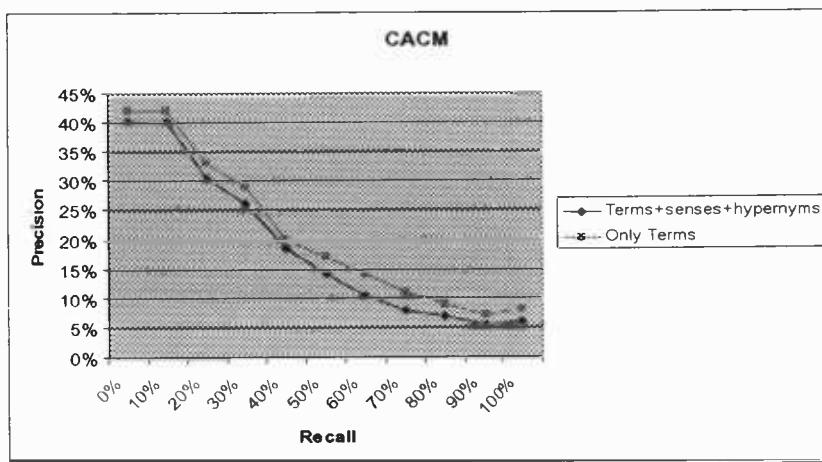
Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων και προκύπτει η αναπαράσταση των κειμένων με βάση τους όρους των κειμένων, τα senses τους και τα υπέρνυμα τους μέχρι την ρίζα, όπως αυτά προκύπτουν από το WordNet. Στην συνέχεια, γίνεται η ανάθεση των βαρών στις λέξεις (όρους) των κειμένων, στα senses τους και στα υπέρνυμα τους, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με τα αποτελέσματα του 1<sup>ο</sup> πειράματος:

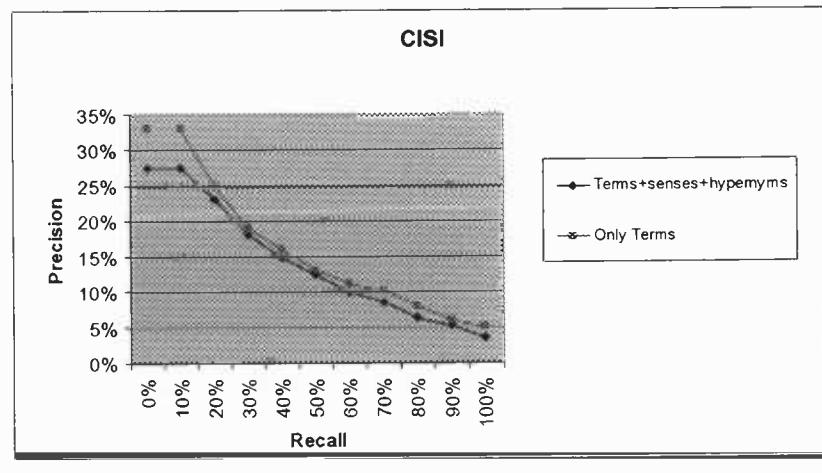
Recall	Precision
0%	77%
10%	77%
20%	73%
30%	67%
40%	59%
50%	52%
60%	46%
70%	39%
80%	30%
90%	22%
100%	9%



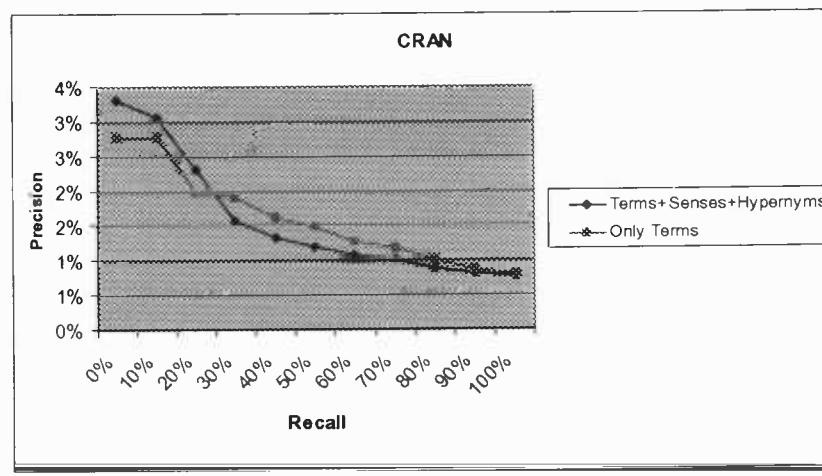
<b>Recall</b>	<b>Precision</b>
0%	40%
10%	40%
20%	30%
30%	26%
40%	18%
50%	14%
60%	10%
70%	8%
80%	7%
90%	5%
100%	6%



<b>Recall</b>	<b>Precision</b>
0%	28%
10%	28%
20%	23%
30%	18%
40%	15%
50%	12%
60%	10%
70%	9%
80%	6%
90%	5%
100%	3%



<b>Recall</b>	<b>Precision</b>
0%	3%
10%	3%
20%	2%
30%	2%
40%	1%
50%	1%
60%	1%
70%	1%
80%	1%
90%	1%
100%	1%



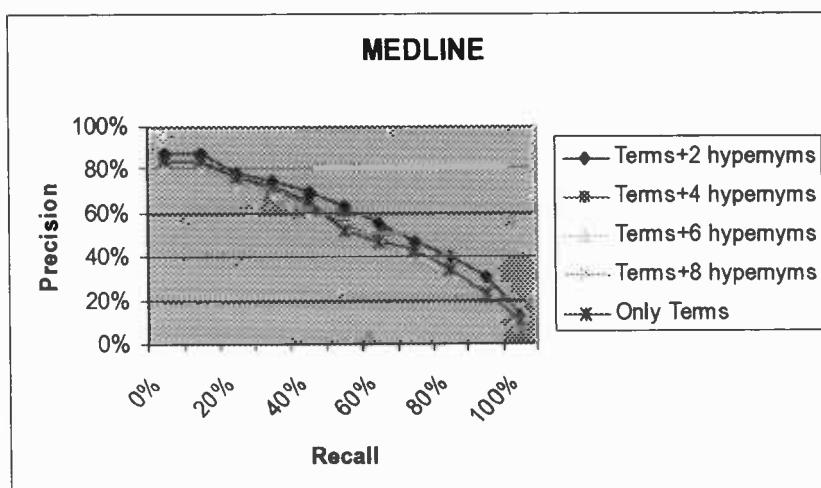
#### 4<sup>ο</sup> Πείραμα:

Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων και προκύπτει η αναπαράσταση των κειμένων με βάση τους όρους των κειμένων και 2, 4, 6 ή 8 από

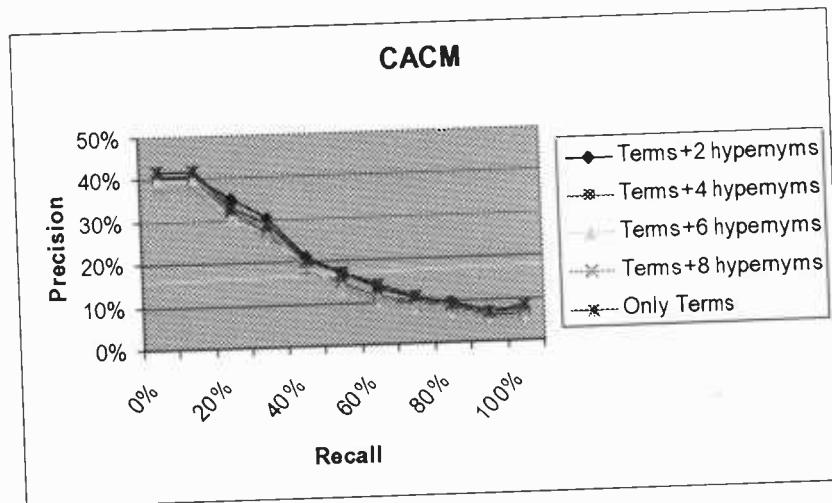
τα υπέρνυμα τους, όπως αυτά προκύπτουν από το WordNet. Στην συνέχεια, γίνεται η ανάθεση των βαρών στις λέξεις (όρους) των κειμένων, στα senses τους και στα υπέρνυμα τους, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με των μεταξύ τους αποτελεσμάτων και με τα αποτελέσματα του 1<sup>ο</sup> πειράματος:

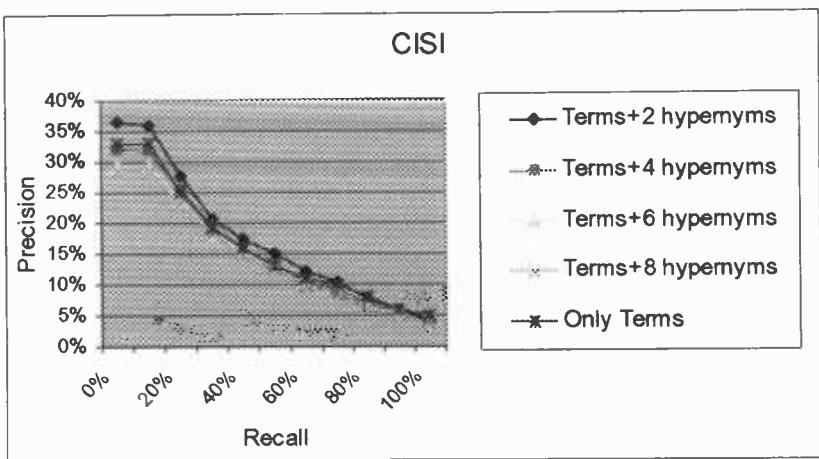
Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	87%	83%	81%	81%
10%	87%	83%	81%	81%
20%	78%	76%	73%	73%
30%	74%	70%	70%	69%
40%	69%	63%	61%	61%
50%	62%	54%	53%	53%
60%	55%	48%	47%	45%
70%	46%	40%	40%	38%
80%	39%	31%	31%	29%
90%	30%	23%	23%	23%
100%	12%	9%	9%	8%



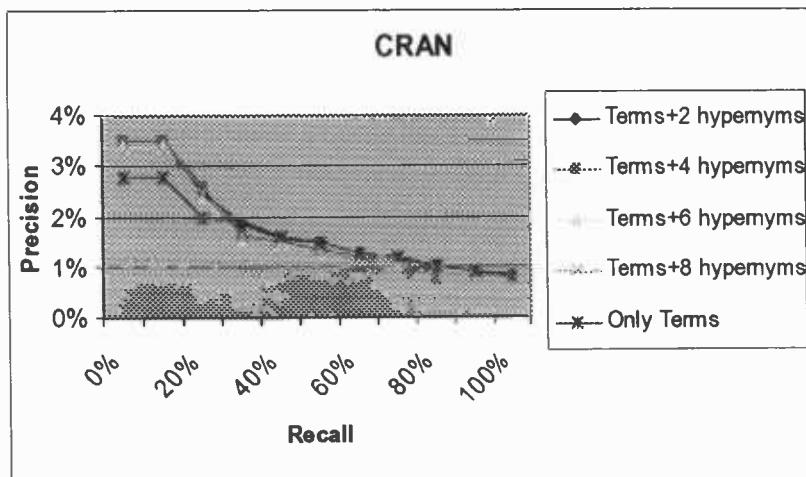
Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	41%	40%	40%	38%
10%	41%	40%	40%	38%
20%	35%	32%	30%	30%
30%	30%	27%	26%	25%
40%	21%	18%	19%	18%
50%	17%	14%	13%	13%
60%	13%	11%	10%	10%
70%	11%	10%	8%	8%
80%	9%	9%	7%	7%
90%	7%	6%	6%	6%
100%	7%	7%	6%	6%



Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	36%	32%	29%	28%
10%	36%	32%	29%	28%
20%	28%	25%	24%	23%
30%	21%	19%	18%	18%
40%	17%	16%	15%	15%
50%	15%	13%	13%	12%
60%	12%	11%	10%	10%
70%	10%	9%	9%	9%
80%	8%	7%	6%	6%
90%	6%	6%	5%	5%
100%	4%	4%	3%	3%



Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	3%	3%	3%	3%
10%	3%	3%	3%	3%
20%	2%	3%	2%	2%
30%	2%	2%	2%	2%
40%	2%	1%	1%	1%
50%	1%	1%	1%	1%
60%	1%	1%	1%	1%
70%	1%	1%	1%	1%
80%	1%	1%	1%	1%
90%	1%	1%	1%	1%
100%	1%	1%	1%	1%



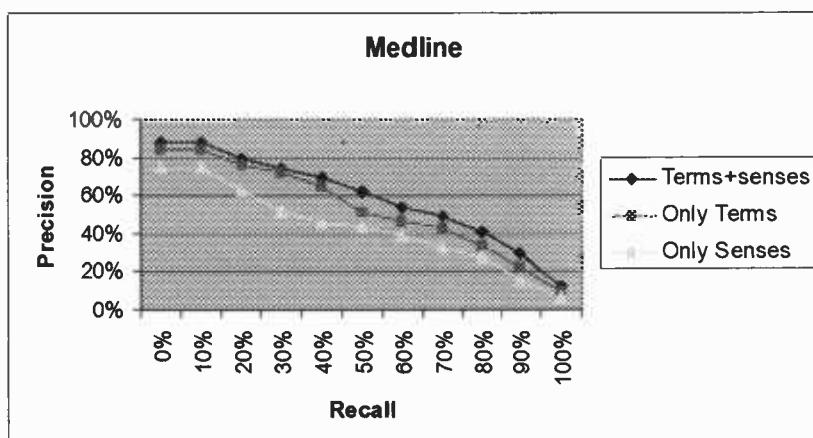
### 5ο Πείραμα:

Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων κατανόησης.

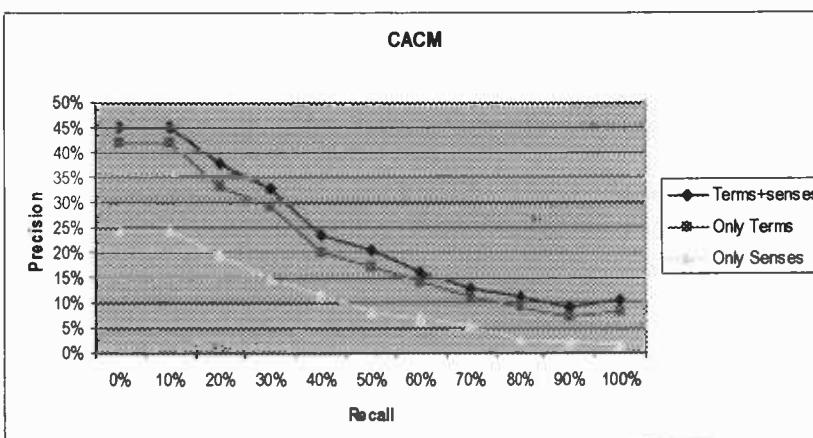
προκύπτει η αναπαράσταση των κειμένων μόνο με τα senses των όρων, όπως αυτά προκύπτουν από το WordNet. Στην συνέχεια, γίνεται η ανάθεση των βαρών στα senses των όρων, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με τα αποτελέσματα του 1<sup>ο</sup> και του 2<sup>ο</sup> πειράματος:

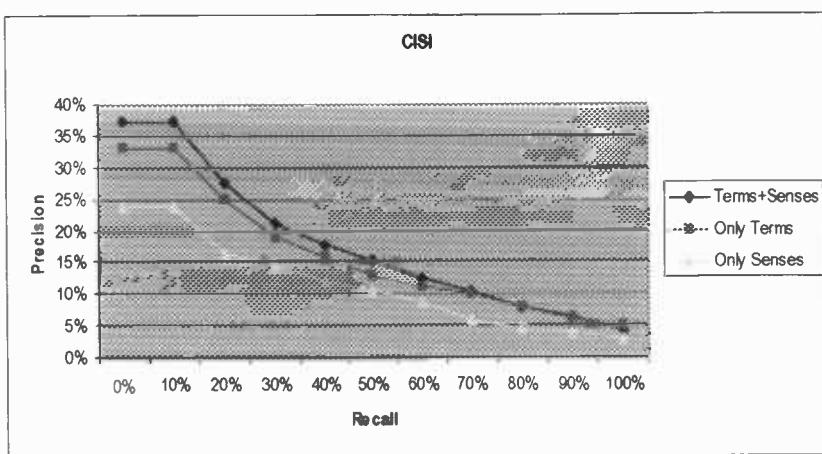
<b>Recall</b>	<b>Precision</b>
0%	75%
10%	75%
20%	63%
30%	51%
40%	45%
50%	43%
60%	39%
70%	32%
80%	28%
90%	16%
100%	6%



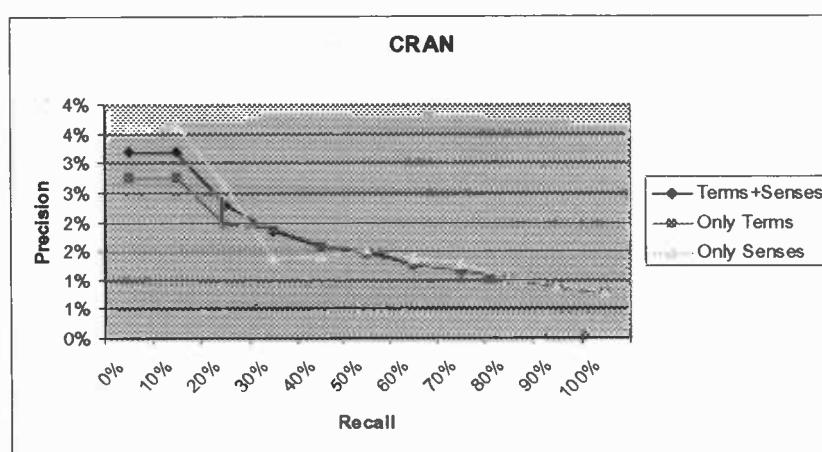
<b>Recall</b>	<b>Precision</b>
0%	24%
10%	24%
20%	19%
30%	15%
40%	11%
50%	8%
60%	7%
70%	5%
80%	2%
90%	2%
100%	1%



Recall	Precision
0%	24%
10%	24%
20%	16%
30%	15%
40%	12%
50%	11%
60%	9%
70%	6%
80%	4%
90%	4%
100%	3%



Recall	Precision
0%	4%
10%	4%
20%	3%
30%	1%
40%	1%
50%	2%
60%	1%
70%	1%
80%	1%
90%	1%
100%	1%

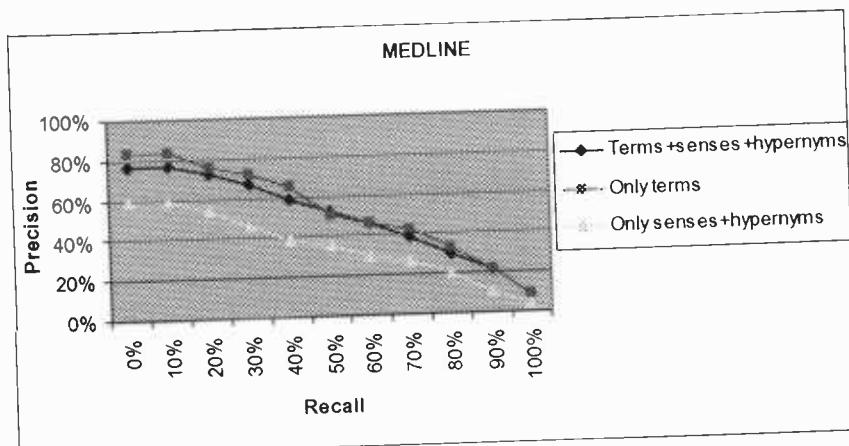


### 6<sup>ο</sup> Πείραμα:

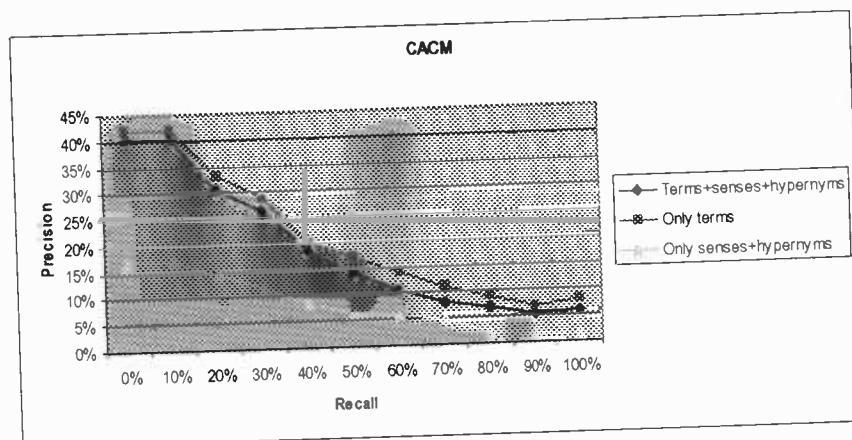
Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων και προκύπτει η αναπαράσταση των κειμένων με βάση τα senses των όρων του κειμένου και τα υπέρνυμα τους μέχρι την ρίζα, όπως αυτά προκύπτουν από το WordNet. Στην συνέχεια, γίνεται η ανάθεση των βαρών στα senses των όρων του κειμένου και στα υπέρνυμα τους, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με τα αποτελέσματα του 1<sup>ο</sup> και 3<sup>ο</sup> πειράματος:

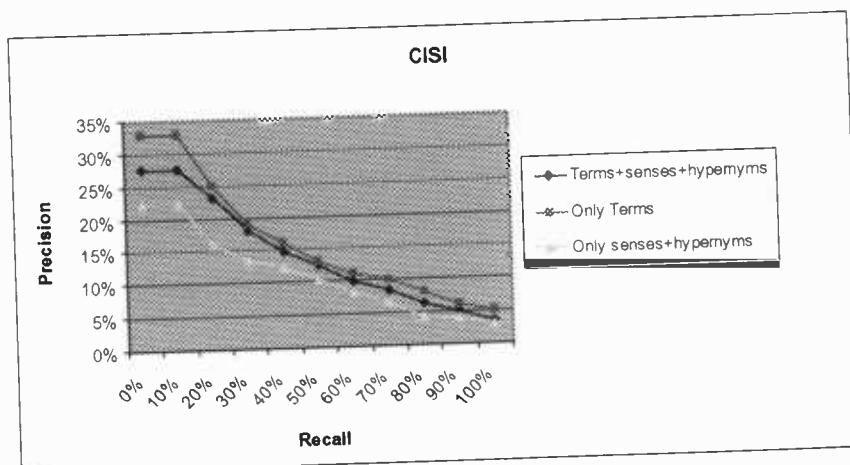
Recall	Precision
0%	59%
10%	59%
20%	54%
30%	46%
40%	39%
50%	35%
60%	30%
70%	26%
80%	20%
90%	11%
100%	4%



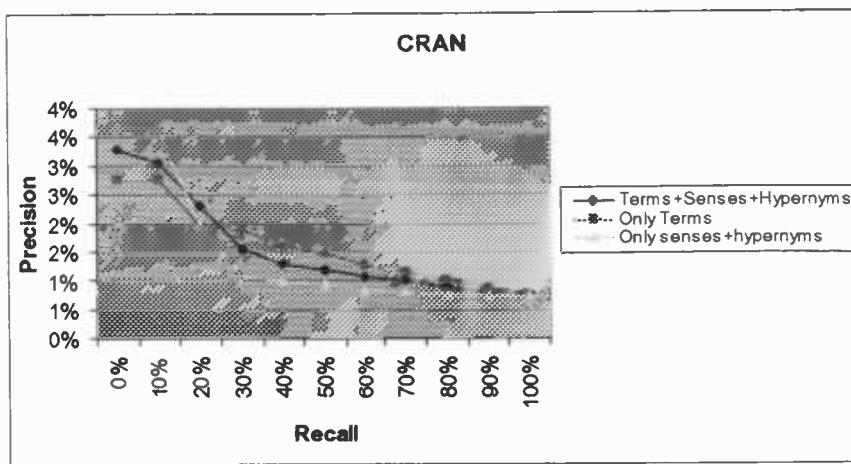
Recall	Precision
0%	17%
10%	17%
20%	14%
30%	12%
40%	8%
50%	6%
60%	6%
70%	5%
80%	2%
90%	1%
100%	1%



Recall	Precision
0%	22%
10%	22%
20%	16%
30%	14%
40%	12%
50%	10%
60%	8%
70%	6%
80%	4%
90%	4%
100%	3%



Recall	Precision
0%	4%
10%	4%
20%	2%
30%	1%
40%	1%
50%	1%
60%	1%
70%	1%
80%	1%
90%	1%
100%	1%

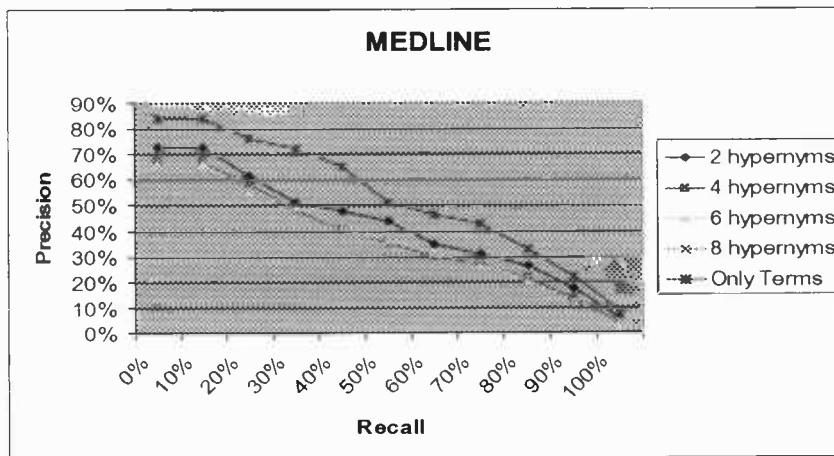


### 7<sup>ο</sup> Πείραμα:

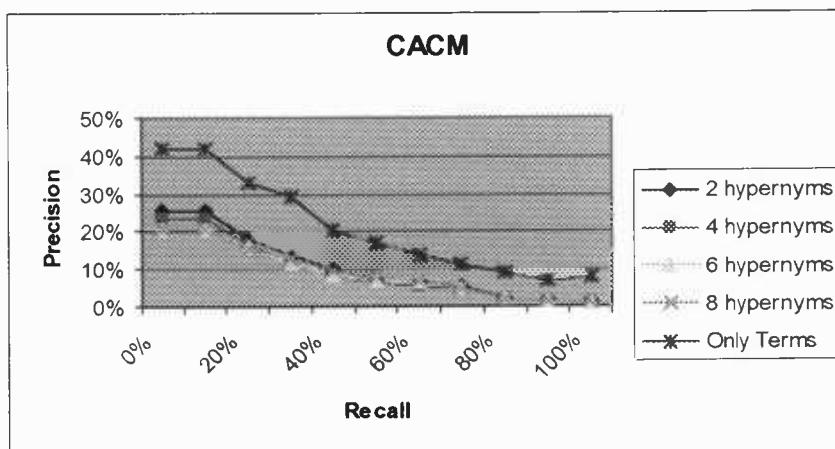
Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων και προκύπτει η αναπαράσταση των κειμένων με βάση 2, 4, 6 ή 8 από τα υπέρνυμα των όρων του κειμένου, όπως αυτά προκύπτουν από το WordNet. Στην συνέχεια, γίνεται η ανάθεση των βαρών στα υπέρνυμα των όρων του κειμένου, σύμφωνα με το TF-IDF. Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με των μεταξύ τους αποτελεσμάτων και με τα αποτελέσματα του 1<sup>ου</sup> πειράματος:

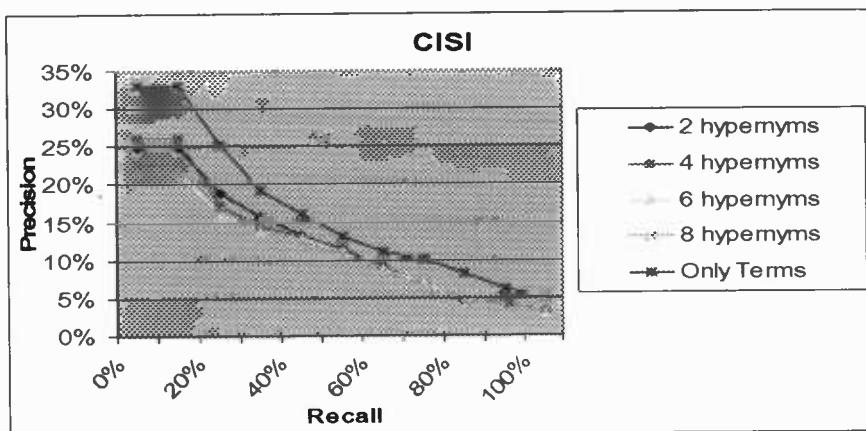
Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	72%	67%	66%	65%
10%	72%	67%	66%	65%
20%	61%	59%	56%	57%
30%	51%	48%	49%	47%
40%	48%	41%	41%	40%
50%	44%	36%	36%	36%
60%	35%	30%	32%	30%
70%	31%	27%	27%	27%
80%	27%	21%	20%	21%
90%	17%	14%	12%	12%
100%	6%	4%	4%	4%



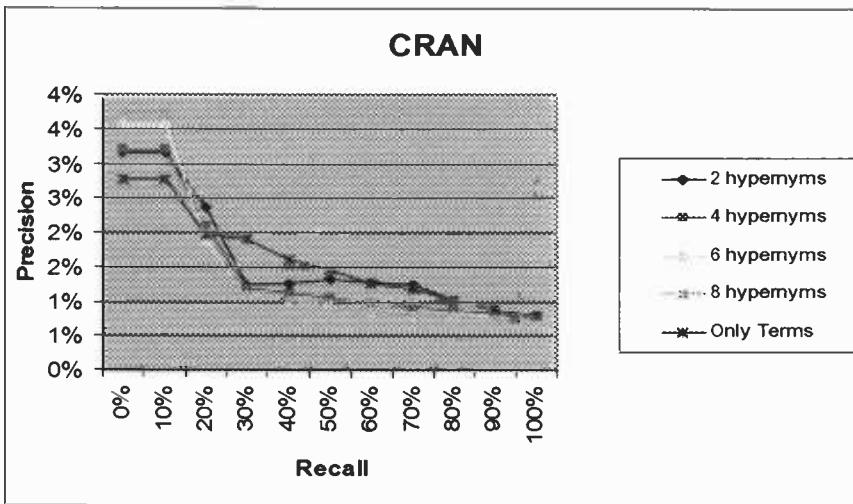
Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	25%	24%	21%	21%
10%	25%	24%	21%	21%
20%	18%	17%	16%	15%
30%	14%	11%	12%	13%
40%	10%	8%	8%	9%
50%	7%	6%	7%	8%
60%	6%	5%	6%	7%
70%	5%	5%	5%	5%
80%	2%	2%	2%	2%
90%	2%	2%	1%	2%
100%	1%	1%	1%	1%



Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	25%	26%	22%	23%
10%	25%	26%	22%	23%
20%	19%	17%	16%	16%
30%	16%	14%	14%	14%
40%	13%	13%	13%	12%
50%	11%	12%	11%	11%
60%	9%	9%	9%	8%
70%	7%	7%	7%	6%
80%	5%	5%	4%	4%
90%	4%	4%	4%	4%
100%	3%	3%	3%	3%



Recall	Precision (2 hypernyms)	Precision (4 hypernyms)	Precision (6 hypernyms)	Precision (8 hypernyms)
0%	3%	3%	4%	4%
10%	3%	3%	4%	4%
20%	2%	2%	2%	2%
30%	1%	1%	1%	1%
40%	1%	1%	1%	1%
50%	1%	1%	1%	1%
60%	1%	1%	1%	1%
70%	1%	1%	1%	1%
80%	1%	1%	1%	1%
90%	1%	1%	1%	1%
100%	1%	1%	1%	1%



### 8<sup>ο</sup> Πείραμα:

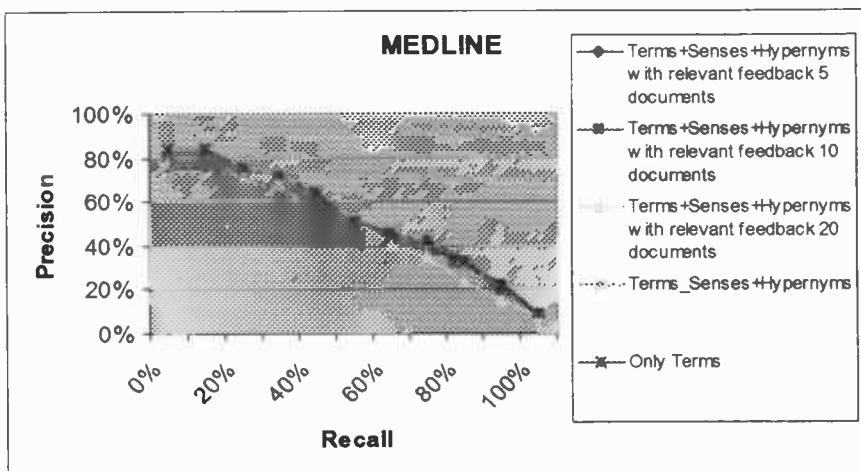
Αφού έχει εκτελεστεί το stemming των λέξεων και η αφαίρεση των τετριμμένων λέξεων (stopwords) από τα κείμενα, αυτή την φορά εκτελείται το disambiguation των όρων και προκύπτει η αναπαράσταση των κειμένων με βάση τους όρους των κειμένων, τα senses τους και τα υπέρονυμα τους μέχρι την ρίζα, όπως αυτά προκύπτουν από το WordNet.

Στην συνέχεια, γίνεται η ανάθεση των βαρών στις λέξεις (όρους) των κειμένων, σύμφωνα με το TF-IDF. Τα βάρη των senses των όρων και των υπέρονυμων τους καθορίζονται, εκτός από το γινόμενο TF\*IDF και με έναν ακόμα συντελεστή που προκύπτει σύμφωνα με την δημοσίευση[5].

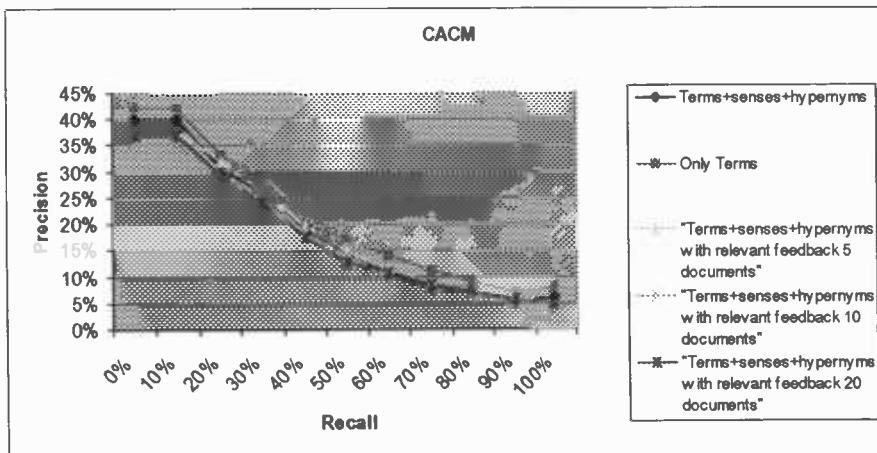
Οπότε, έχοντας την αναπαράσταση των κειμένων, και αναπαριστώντας με τον ίδιο τρόπο και τις επερωτήσεις που διαθέτει η κάθε συλλογή, εκτελούνται οι επερωτήσεις και τα αποτελέσματα κατατάσσονται με βάση το cosine similarity. Με βάση τα αποτελέσματα και τις «πραγματικές» απαντήσεις των επερωτήσεων, υπολογίζονται τα Precision και Recall για κάθε επερώτηση. Υπολογίζοντας το μέσο όρο των τιμών του Precision, για κάθε επερώτηση, για κάθε τιμή Recall, κατασκευάζονται οι καμπύλες των Precision-Recall.

Τα αποτελέσματα που προκύπτουν για κάθε συλλογή παρουσιάζονται παρακάτω και ταυτόχρονα γίνεται σύγκριση με τα αποτελέσματα του 1<sup>ο</sup> και του 3<sup>ο</sup> πειράματος:

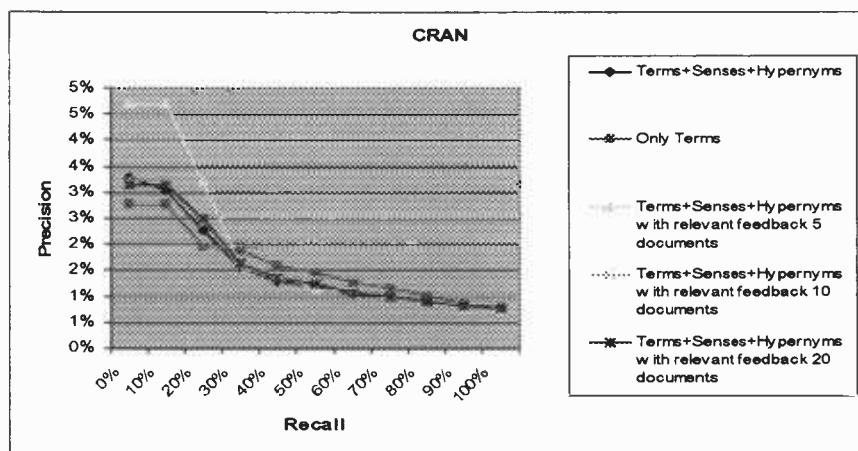
Recall	Precision (relevant feedback 5 documents)	Precision (relevant feedback 10 documents)	Precision (relevant feedback 20 documents)
0%	76%	80%	77%
10%	76%	80%	77%
20%	72%	72%	71%
30%	66%	66%	67%
40%	62%	60%	60%
50%	53%	53%	52%
60%	45%	43%	40%
70%	40%	38%	35%
80%	32%	31%	23%
90%	22%	19%	15%
100%	9%	8%	7%



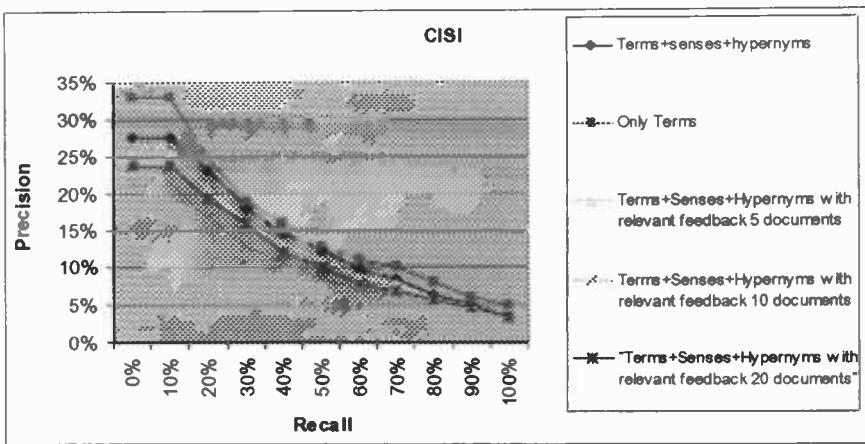
Recall	Precision (relevant feedback 5 documents)	Precision (relevant feedback 10 documents)	Precision (relevant feedback 20 documents)
0%	38%	35%	37%
10%	38%	35%	37%
20%	32%	30%	29%
30%	26%	26%	24%
40%	20%	18%	18%
50%	15%	14%	13%
60%	11%	11%	11%
70%	9%	8%	9%
80%	7%	7%	7%
90%	6%	6%	6%
100%	5%	5%	5%



Recall	Precision (relevant feedback 5 documents)	Precision (relevant feedback 10 documents)	Precision (relevant feedback 20 documents)
0%	5%	3%	3%
10%	5%	3%	3%
20%	3%	2%	2%
30%	2%	2%	2%
40%	1%	1%	1%
50%	1%	1%	1%
60%	1%	1%	1%
70%	1%	1%	1%
80%	1%	1%	1%
90%	1%	1%	1%
100%	1%	1%	1%



Recall	Precision (relevant feedback 5 documents)	Precision (relevant feedback 10 documents)	Precision (relevant feedback 20 documents)
0%	25%	21%	24%
10%	25%	21%	24%
20%	21%	17%	19%
30%	16%	13%	16%
40%	13%	11%	12%
50%	11%	9%	10%
60%	9%	7%	8%
70%	8%	6%	7%
80%	6%	5%	6%
90%	5%	5%	5%
100%	3%	3%	3%



### 7.3 Σχολιασμός Αποτελεσμάτων

Είναι φανερό από τα διαγράμματα του δεύτερου πειράματος, ότι η αναπαράσταση των κειμένων και των επερωτήσεων με τους όρους που περιέχουν και τα senses τους παρουσιάζουν αρκετά μεγάλη βελτίωση στα αποτελέσματα, σε σχέση με τα αποτελέσματα του πρώτου πειράματος που δεν έχει γίνει καμιά επεξεργασία. Αυτό το αποτέλεσμα είναι κατανοητό, αναλογίζοντας ότι με την προσθήκη των senses του WordNet είναι δυνατόν να ανακτώνται κείμενα που περιέχουν λέξεις που ανήκουν στο ίδιο synset με την λέξη της επερώτησης. Αναλυτικά, για την κάθε συλλογή παρατηρούνται τα εξής συμπεράσματα: για την medline αύξηση της απόδοσης έως και 4%, για την caci αύξηση της απόδοσης έως και 5%, για την cisι έως και 4% και για την cran αύξηση της απόδοσης μικρότερη από 1%.

Δυστυχώς όμως, τα αποτελέσματα του τρίτου πειράματος, όπου η αναπαράσταση των κειμένων και κατά επέκταση των επερωτήσεων γίνεται με βάση τους όρους, τα senses και όλα τα υπέρνυμα των senses τους, δεν είναι ενθαρρυντικά, αφού είναι χειρότερα από όταν δεν γίνεται καμιά επεξεργασία στις επερωτήσεις. Ίσως, και αυτό μπορεί να αιτιολογηθεί από το γεγονός ότι η προσθήκη όλων των υπέρνυμων των όρων της επερώτησης να γενικεύουν την επερώτηση και να αποπροσανατολίζουν την πληροφορία που χρειάζεται ο χρήστης και εκφράζει μέσω αυτής.

Το επόμενο πείραμα, το τέταρτο, έχει ως στόχος την διερεύνηση ενός αριθμού υπέρνυμων που είναι δυνατόν να λαμβάνεται, ώστε να μελετηθεί η βελτίωση των αποτελεσμάτων από αυτά που προκύπτουν χωρίς την επέκταση της επερώτησης. Το πλήθος των υπέρνυμων των όρων των κειμένων και της επερώτησης που εξετάζονται είναι το δύο, το τέσσερα, το έξι και το οκτώ. Αυτό που αποδεικνύεται, μέσω των παραπάνω διαγραμμάτων, είναι ότι τα καλύτερα αποτελέσματα μεταξύ των διαφορετικών πληθών υπέρνυμων, αλλά και από τα αρχικά αποτελέσματα τα έχει η αναπαράσταση με τα δύο υπέρνυμα, για όλες τις συλλογές. Αυτό σημαίνει ότι οι όροι των κειμένων και της επερώτησης, μαζί με την ερμηνεία αυτών και με ένα επιπλέον υπέρνυμο της ερμηνείας των όρων έχουν τα καλύτερα αποτελέσματα. Οπότε, εξάγεται το συμπέρασμα, ότι η επιπλέον πληροφορία που χρησιμοποιείται με την επέκταση της επερώτησης, όπως αποδεικνύεται, βοηθάει στην σωστή αποσαφήνιση των όρων της επερώτησης και των κειμένων. Η επιπλέον πληροφορία που εμφανίζεται στα κείμενα και στις επερωτήσεις των συλλογών χρησιμοποιώντας τα τέσσερα, τα έξι και τα οκτώ υπέρνυμα των όρων δεν αποφέρουν αισθητή διαφορά στα αποτελέσματα, σε σχέση με αυτά που υπάρχουν όταν χρησιμοποιούνται μόνο οι όροι τους, χωρίς περαιτέρω επεξεργασία.

Αναλυτικά, για κάθε συλλογή παρατηρούνται τα ακόλουθα συμπεράσματα: η συλλογή

Medline για δύο υπέρνυμα έχει ένα μέσο όρο αύξησης της αποδοτικότητας της ανάκτησης γύρω στο 5%, για τέσσερα υπέρνυμα έχει σχεδόν τα ίδια αποτελέσματα με αυτά χωρίς επεξεργασία, για έξι και οκτώ υπέρνυμα τα αποτελέσματα δεν είναι ενθαρρυντικά, αφού παρατηρείται μείωση της απόδοσης της ανάκτησης μέχρι και 3%. Η συλλογή Casm για δύο υπέρνυμα έχει παρόμοια αποτελέσματα με την χωρίς επεξεργασία, αλλά σε κάποια επίπεδα του recall υπάρχει αύξηση της αποδοτικότητας της ανάκτησης μέχρι και 2%, για τέσσερα υπέρνυμα έχει μείωση της απόδοσης γύρω στα 2%, για έξι και οκτώ υπέρνυμα τα αποτελέσματα, δεν είναι ενθαρρυντικά τα αποτελέσματα, αφού η μείωση της απόδοσης της ανάκτησης φτάνει μέχρι και 4%. Η συλλογή Cisi για δύο υπέρνυμα έχει ένα μέσο όρο αύξησης της αποδοτικότητας της ανάκτησης γύρω στο 3%, για τέσσερα υπέρνυμα έχει σχεδόν τα ίδια αποτελέσματα με αυτά χωρίς επεξεργασία, για έξι και οκτώ υπέρνυμα τα αποτελέσματα δεν είναι ενθαρρυντικά, αφού η μείωση της απόδοσης της ανάκτησης είναι μέχρι και 4%. Τέλος, για την συλλογή Cran παρατηρείται ελαφρώς αύξηση της απόδοσης για 2 υπέρνυμα (μικρότερη από 1%) και για τα υπόλοιπα τέσσερα, έξι ή οκτώ υπέρνυμα παρατηρούνται σχεδόν ίδια αποτελέσματα ή μείωση έως 1% με τα χωρίς επεξεργασία.

Στο πέμπτο, στο έκτο και στο έβδομο πείραμα, όπου η αναπαράσταση των κειμένων και κατά επέκταση των επερωτήσεων γίνεται χωρίς τους όρους, αλλά μόνο με βάση τα senses ή τα senses και όλα τα υπέρνυμα των senses τους ή δύο, τέσσερα, έξι ή οκτώ υπέρνυμα των όρων των κειμένων, αντίστοιχα, δεν είναι καθόλου ενθαρρυντικά. Η μείωση, τόσο από τα αρχικά αποτελέσματα, όσο και από τα αντίστοιχα με τους όρους είναι τεράστια. Τα αποτελέσματα αυτά μπορούν να αιτιολογηθούν, αναλογίζοντας ότι οι όροι της επερώτησης που καταφέρνει να προσδιορίσει την ερμηνεία του η διαδικασία του Word Sense Disambiguation, σύμφωνα με το WordNet είναι πολύ λίγοι και η αφαίρεση των όρων έχει μεγάλο αντίκτυπο στα αποτελέσματα.

Το όγδοο και τελευταίο πείραμα προσπαθεί να παρουσιάσει μια διαφορετική προσέγγιση στην επέκταση επερωτήσεων, όπου δεν γίνεται μόνο προσθήκη καινούριων όρων, αλλά και τροποποίηση των βαρών τους. Αναλυτικά, εκτός από την προσθήκη των ερμηνειών των όρων (senses) και των υπέρνυμων τους στις επερωτήσεις και στα κείμενα των συλλογών, διαφοροποιείται και το βάρος που τους αναθέτετε με το  $tf * idf * w$ , όπου  $w$  ο συντελεστής που προκύπτει με βάση την μελέτη [5] αντί για το απλό  $tf * idf$ . Δυστυχώς, τα αποτελέσματα δεν είναι τα αναμενόμενα, αφού είναι χειρότερα από το να μην είχαμε επεξεργαστεί καθόλου τους όρους της επερώτησης και των κειμένων, με μια εξαίρεση για την συλλογή cran που για τα σημεία 0%, 10% και 20% του recall παρατηρείται αύξηση της απόδοσης της ανάκτησης 3%, στην περίπτωση που τα βάρη προκύψουν από την εξέταση των 5 πρώτων αποτελεσμάτων της ανάκτησης. Όπως γίνεται σαφές από τα διαγράμματα υπάρχουν κάποιες διαφοροποιήσεις στην απόδοση που προκύπτει μεταξύ του διαφορετικού πλήθους των κειμένων που εξετάζονται κάθε φορά, για την παραγωγή των βαρών σύμφωνα με την μελέτη [5] που δεν θα ήταν αναμενόμενες εξαρχής τουλάχιστον. Πιο συγκεκριμένα, αντί να αυξάνεται η απόδοση, καθώς αυξάνεται το πλήθος των κειμένων που εξετάζονται παρατηρείται μείωση. Αυτή η μείωση μπορεί να αιτιολογηθεί αναλογίζοντας ότι μπορεί να υπάρχει αύξηση των κειμένων που εξετάζονται, αλλά οι περιορισμοί από τη σύγκριση των αποτελεσμάτων του συστήματος ανάκτησης και της ανάδρασης του χρήστη μπορεί να επικαλύπτονται από επερώτηση σε επερώτηση και συνολικά όχι μόνο να μην αυξάνεται η πληροφορία που προκύπτει, αλλά και να μειώνεται.

## 8. Γενικά Συμπεράσματα - Μελλοντική Εργασία

Από όλη την παραπάνω εργασία μπορεί εύκολα να κατανοηθεί ότι η επέκταση των επερωτήσεων, καθώς και των όρων των κειμένων με σχετικούς όρους μπορεί να βελτιώσει την αποτελεσματικότητα της ανάκτησης και ιδιαίτερα την ανάκληση (recall).

Είναι όμως εξίσου σημαντικό να επισημανθεί ότι η αλόγιστη επιλογή σχετικών όρων μπορεί να μειώσει την ακρίβεια (precision) και τα αποτελέσματα της ανάκτησης της πληροφορίας να είναι απογοητευτικά.

Γενικά, η επέκταση της επερώτησης με νέους όρους, που προέρχονται από τις σχέσεις των υπερνύμων των όρων της αρχικής επερώτησης, πρέπει να γίνεται με πολύ προσεκτικό τρόπο, αφού είναι πολύ εύκολο να γενικευτεί τόσο πολύ η επερώτηση που τα αποτελέσματα της ανάκτησης να μην είναι τα επιθυμητά.

Παράλληλα, κρίνεται απαραίτητη η τροποποίηση των βαρών των νέων σχετικών λέξεων. Διαφορετικά τα βάρη των νέων λέξεων μπορεί να είναι χαμηλότερα των βαρών των λέξεων της αρχικής επερώτησης και να μην αντικατοπτρίζεται η σημαντικότητα του κάθε όρου.

Υιοθετώντας τον παραπάνω τρόπο εύρεσης των βαρών των νέων όρων, η μελλοντική εργασία που μπορεί να γίνει για την βελτίωση της απόδοσης της ανάκτησης είναι η ανάγκη εύρεσης καλύτερου τρόπου αντιστοίχησης της ανάδρασης του χρήστη και των περιορισμών must/cannot link.

Διαφορετικά, είναι δυνατόν να βρεθούν εναλλακτικοί τρόποι υπολογισμού των βαρών των νέων όρων, που να μην χρησιμοποιούν την ανάδραση του χρήστη. Παράλληλα, είναι δυνατόν να εξεταστεί η απόδοση της ανάκτησης επεκτείνοντας την επερώτηση και με όρους που προέρχονται από άλλες σχέσεις του WordNet, εκτός των υπερνύμων.



## 9. Βιβλιογραφία

1. Modern Information Retrieval, Ricardo Baeza-Yates, Berthier Ribeiro-Neto
2. WordNet An Electronic Lexical Database, Christiane Fellbaum
3. Word Sense Disambiguation for Exploiting Hierarchical Thesauri in Text Classification, Dimitrios Mauroeidis, George Tsatsaronis, Michalis Vazirgiannis, Martin Theobald and Gerhard Weikum
4. Word Sense Disambiguation in Information Retrieval, Christopher Stokoe, Michael P.Oakes, John Tait
5. Distance metric learning with application to clustering with side-information, Eric P. Xing, Andrew Y.Ng, Michael I.Jordan and Stuart Russell
6. Integrating Constraints and Metric Learning in Semi-supervised Clustering, Mikhait Bilenko, Sugato Basu and Raymond J.Mooney
7. Y. Bar-Hillel. *Language and Information*. Addison-Wesley, 1964.
8. Weiss SF. Learning to disambiguate Information Storage and Retrieval 1973, 9:33-41
9. Krovetz, R. and Croft, W. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115-141
10. Voorhees, E. (1993) Using WordNet to Disambiguate Word Senses for Text Retrieval. SIGIR-93.
11. Wallis P. Information retrieval based on paraphrase Proceeding of PACLING Conference 1993
12. M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings, ACM Special Interest Group on Information Retrieval*, pages 142-151, 1994.
13. Stokoe, Oakes και Tait, Word sense disambiguation in information retrieval revisited.
14. J. Hutchins and H. Sommers. *Introduction to Machine Translation*. Academic Press, 1992.
15. P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL-91)*, pages 264-270, Berkley, C.A., 1991.
16. Dagan, Itai. "Word Sense Disambiguation using a second Language Monolingual Corpus". Computational Linguistics, vol 20, 1994 pp563-596
17. Kali, Morimoto. "Unsupervised Word Sense Disambiguation using a bilingual comparable corpora". Proceeding of the 19<sup>th</sup> international conference on Computational Linguistics, 2002, pp 411-417.

18. Oliveira, Wong, Li, Zheng. Unsupervised Word Sense Disambiguation and Rules Extraction using non-aligned bilingual corpus. Proceeding of NLP-KE'05
19. Y. Wilks and M. Stevenson. The Grammar of Sense: using part-of-speech tags as a first step in semantic disambiguation. To appear in *Journal of Natural Language Engineering*, 4(3).
20. A. Harley and D. Glennon. Sense tagging in action: Combining different tests with additive weights. In *Proceedings of the SIGLEX Workshop 'Tagging Text with Lexical Semantics'*, pages 74-78. Association for Computational Linguistics, Washington, D.C., 1997.
21. D. Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454-460, Nantes, France, 1992.
22. Mihalcea, Tarau, Figa "PageRank on Sematic Networks, with application to Word Sense Disambiguation" Coling 2004, Switzerland, Geneva, 2004
23. Patwardhan, Banerjeev, Pedersen " Using Measures of Semantic Relatedness for Word Sense Disambiguation, CICLing 2003:241-257
24. Shuang Lui, Clement Yu και Weiyi Meng Word Sense Disambiguation in Queries
25. Escudero, Marquez και Rigau "A comparison between Supervised Learning Algorithms for Word Sense Disambiguation" Proceeding of CoNLL-2000 και LLL-2000, pages 31-36, Lisbon, Portugal, 2000
26. D. Yarowsky. One sense per collocation. In *Proceedings ARPA Human Language Technology Workshop*, pages 266-271, Princeton, NJ, 1993
27. T. Pedersen and R. Bruce. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, August 1997.
28. R. Krovetz. Homonymy and polysemy in information retrieval. In *35th Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics (ACL/EACL-97)*, pages 72-78, Madrid, Spain, 1997.
29. A. Luk. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Meetings of the Association for Computational Linguistics (ACL-95)*, pages 181-188, Cambridge, M.A., 1995.
30. D. Yarowsky. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189-196, Cambridge, MA, 1995.
31. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze. "An Introduction to Information Retrieval", Cambridge University Press Cambridge, England, Press 2007



