



ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΕΙΔΟΥΣ ΗΧΗΚΗ
ειδ. 53694
Αρ. 005.γ
ΑΝΤ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

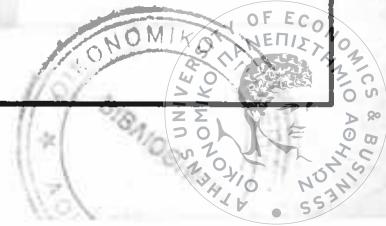
*Δημιουργία και Αξιολόγηση Συστήματος
Ανάκτησης Ελληνικών Κειμένων*

**Αντωνόπουλος Παναγιώτης
M3960001**

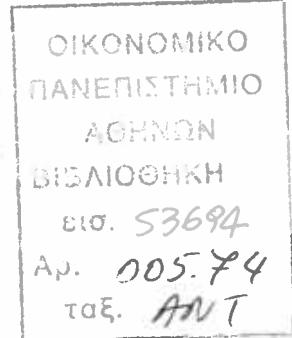
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ



ΑΘΗΝΑ, ΙΑΝΟΥΑΡΙΟΣ 1998



**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

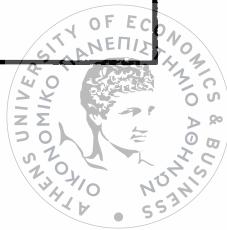


ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Δημιουργία και Αξιολόγηση Συστήματος
Ανάκτησης Ελληνικών Κειμένων**

**Αντωνόπουλος Παναγιώτης
M3960001**

Επιβλέπων Καθηγητής: Θεόδωρος Ζ. Καλαμπούκης



Πρόλογος

Θέμα της παρούσας διπλωματικής εργασίας είναι η δημιουργία και η αξιολόγηση ενός πρότυπου Συστήματος Ανάκτησης Ελληνικών Κειμένων. Πιο συγκεκριμένα, στα πλαίσια αυτής της εργασίας έγινε ο σχεδιασμός και η υλοποίηση μιας WWW μηχανή αναζήτησης στα πρότυπα των μεγάλων μηχανών αναζήτησης που υπάρχουν στο Internet (π.χ. Altavista, Infoseek, Lycos κ.λ.π.), η οποία μπορεί και υποστηρίζει πλήρως αναζητήσεις τόσο σε αγγλική όσο και σε ελληνική γλώσσα. Απότερος σκοπός ήταν η δημιουργία μιας μηχανής αναζήτησης που να ενσωματώνει τεχνικές και χαρακτηριστικά που συναντάμε σε παραδοσιακά συστήματα ανάκτησης πληροφοριών όπως αφαίρεση τετριψμένων λέξεων, αποκοπή καταλήξεων στους όρους, διαβάθμιση των αποτελεσμάτων κατά βαθμό σχετικότητας, εφαρμογή τελεστών στην ερώτηση του χρήστη κ.α. για την επίτευξη καλύτερων αποτελεσμάτων αλλά και για την πλήρη υποστήριξη των ελληνικών αναζητήσεων.

Αρχική φιλοδοξία της παρούσας διπλωματικής εργασίας ήταν η κατασκευή μιας μηχανής αναζήτησης για όλα τα αποθέματα πληροφοριών που υπήρχαν μέσα στο ελληνικό χώρο ή καλύτερα για όλα όσα υπήρχαν στην ελληνική γλώσσα. Κάπι τέτοιο όμως στην συνέχεια αποδείχτηκε πολύ δύσκολο λόγω του όγκου των πληροφοριών - HTML σελίδων - που υπάρχουν στο Internet αλλά και λόγου της δυσκολίας ανακάλυψης τους. Εξάλλου σκοπός της παρούσας διπλωματικής εργασίας ήταν η κατασκευή ενός πιλοτικού συστήματος ανάκτησης και όχι μια ολοκληρωμένη λύση. Ετσι προτιμήθηκε ο περιορισμός του συστήματος ανάκτησης σε αναζητήσεις ενός συγκεκριμένου χώρου τόσο θεματολογικά όσο και διευθυνσιακά οριοθετημένου.

Για την επίτευξη του παραπάνω στόχου αναπτύχθηκε ένα πρότυπο πρόγραμμα Web Robot το οποίο ήταν υπεύθυνο για την ανακάλυψη των αποθεμάτων - HTML σελίδων - που έπρεπε να δεικτοδοτηθούν από το σύστημα ανάκτησης και το κατέβασμα τους τοπικά στο δίσκο για την δημιουργία της συλλογής του συστήματος ανάκτησης. Το αποτέλεσμα της λειτουργίας του Web Robot ήταν η ανακάλυψη και το κατέβασμα 2500 χιλιάδων περίπου HTML σελίδων από 30 ελληνικά sites που ανήκουν στην κατηγορία "Health & Medicine".

Η δεικτοδότηση των σελίδων και η δημιουργία των ευρετηρίων καθώς και η λειτουργία του μηχανισμού αναζήτησης βασίστηκε στις υπηρεσίες ανάκτησης πληροφοριών που προσφέρει το Oracle ConText Option. Σε αυτό ενσωματώθηκε το πλήρες σύνολο των



τετριμμένων λέξεων της ελληνικής καθώς και συνάρτηση αποκοπής καταλήξεων της ελληνικής η οποία και ήταν υπεύθυνη για την υλοποίηση του ελληνικού stemming.

Τέλος, έγινε αξιολόγηση των αποτελεσμάτων του συστήματος ανάκτησης για τις περιπτώσεις χρήσης και μη χρήσης αποκοπής καταλήξεων στους όρους της ερώτησης του χρήστη. Η αξιολόγηση των αποτελεσμάτων έγινε με την υποβολή στο σύστημα ανάκτησης 20 ερωτήσεων σε ελληνική γλώσσα με και χωρίς αποκοπή καταλήξεων στους όρους αυτών και τον υπολογισμό και κατασκευή των αντίστοιχων γραφημάτων απόκρισης-ακρίβειας. Επίσης για λόγους πληρότητας έγινε και μια σύγκριση των αποτελεσμάτων του συστήματος ανάκτησης με τα αντίστοιχα που δίνουν οι μεγάλες μηχανές αναζήτησης του Internet που υποστηρίζουν ελληνικές αναζητήσεις και παρυυσιάστηκαν τα σχετικά συμπεράσματα.

Η δομή της διπλωματικής εργασίας χωρίζεται σε έξι κεφάλαια:

Στο πρώτο κεφάλαιο γίνεται αναφορά σε βασικές έννοιες της ανάκτησης πληροφοριών και παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασίστηκε η υλοποίηση του συστήματος ανάκτησης.

Στο δεύτερο κεφάλαιο γίνεται μια γενική περιγραφή των σημαντικότερων μηχανών αναζήτησης που υπάρχουν σήμερα στο Internet καθώς και των προβλημάτων που καλούνται να αντιμετωπίσουν και να ξεπεράσουν.

Στο τρίτο κεφάλαιο αναλύεται η σχέση μεταξύ των συστημάτων ανάκτησης πληροφοριών και των συστημάτων διαχείρισης βάσεων δεδομένων, δεδομένου ότι το σύστημα ανάκτησης υλοποιήθηκε με την βοήθεια του Oracle RDBMS. Στη συνέχεια γίνεται σύντομη παρουσίαση ενός τέτοιου συστήματος (σύστημα βιβλιογραφικών εγγραφών retriev) και περιγράφονται οι υπηρεσίες ανάκτησης πληροφοριών που παρέχει το Oracle ConText 7.3 στις οποίες βασίστηκε η υλοποίηση του συστήματος ανάκτησης.

Στο τέταρτο κεφάλαιο γίνεται πλήρης περιγραφή της αρχιτεκτονικής και λειτουργίας τόσο του προγράμματος του Web Crawler όσο και του συστήματος ανάκτησης. Και στις δύο περιπτώσεις παρουσιάζονται με εικόνες οι λειτουργίες τους και τα σενάρια χρήσης τους.

Στο πέμπτο κεφάλαιο γίνεται αναφορά στις παραμέτρους που συμμετέχουν και την μεθοδολογία που ακολουθείται για την αξιολόγηση των αποτελεσμάτων των συστημάτων ανάκτησης και στη συνέχεια παρουσιάζονται τα αποτελέσματα και συμπεράσματα που προέκυψαν από πειράματα πάνω στο σύστημα ανάκτησης.



Τέλος, στο έκτο κεφάλαιο γίνεται μια κριτική αξιολόγηση του εργαλείου Oracle ConText που χρησιμοποιήθηκε για την υλοποίηση του συστήματος ανάκτησης και εστιάζονται θέματα βελτιώσεων και περαιτέρω έρευνας που μπορούν να γίνουν τόσο στο σύστημα ανάκτησης που υλοποιήθηκε όσο και γενικότερα στο χώρο της ανάκτησης πληροφοριών για την υποστήριξη των ελληνικών αναζητήσεων.

Κλείνοντας τον πρόλογο αυτό, αισθάνομαι την ανάγκη να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Θεόδωρο Ζ. Καλαμπούκη για την μεθοδική καθοδήγηση και τις εποικοδομητικές παρατηρήσεις που μου παρείχε σε όλη τη διάρκεια εκπόνησης της παρούσας εργασίας καθώς επίσης και τον συμφοιτητή μου Γιώργο Χαλκιαδάκη για την αμέριστη βοήθεια - τόσο γνωστική όσο και πρακτική - που μου παρείχε.

Αντωνόπουλος Δ. Παναγιώτης
Αθήνα, Φεβρουάριος 1998

ΠΡΟΛΟΓΟΣ	23
Επίσημη Μητρώος Έργα, Ημέρα Διάταξης	23
Ι	23
1. Το πανεπιστημιακό έργο με τη Βορειανή ΔΕΗΑΣ	24
1.1. Η επίσημη ημέρα της εγγράτης απόστασης	25
1.2. Ημέρα της μάθησης	26
1.3. Επίλεγοντας απόσταση	27
1.4. Επίλεγοντας	28
2. Η παραπομπή προτάσεων	29
2.1. Άριστη πρότερη προτάσεις Oracle Ορθοί 7.3	31
2.2. Σημεία Αναζήτησης	31
2.3. Τιμολόγηση αποτελεσμάτων αναζήτησης	35
2.4. Επιλογές Αναζητήσεων	36
2.5. Επίλεγοντας προτάσεις	40
3. ΕΠΙΒΛΔΙΟΣ	44
4. Σύστημα Ανάπτυξης Ελληνικών Κερδίου	44
4.1. Εγκαί	44
4.2. Δημοποιούμενη έργατη	45
4.2.1. Αρχικοποίηση αποτελεσμάτων των Web Crawler	45
4.2.2. Αρχικοποίηση των διανομητών Ανέπτυξης	49
4.2.3. Αρχικοποίηση αποτελεσμάτων Ανέπτυξης	59
5. ΕΠΕΙΔΗΣ	64
6. Λέξιολογος Ανεπτυξτικής Απόστασης	64
6.1. Αποτελεσματούργια Στοιχεία Ανέπτυξης	64
6.2. Αποτελεσματούργια Στοιχεία Ανέπτυξης	64
6.3. Αποτελεσματούργια Στοιχεία Ανέπτυξης	64
6.4. Αποτελεσματούργια Στοιχεία Ανέπτυξης	64



Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1	4
Ανάκτηση Πληροφοριών	4
1.1 Αντικείμενο της Ανάκτησης Πληροφοριών	4
1.2 Δομή ενός IRS	4
1.3 Λειτουργία ενός IRS	7
ΚΕΦΑΛΑΙΟ 2	10
Μηχανές Αναζήτησης στο Διαδίκτυο και Ανάκτηση Πληροφοριών	10
2.1 Γενικά	10
2.2 Θεματολογικά Δέντρα και Ιεραρχίες Πληροφοριών	11
2.3 Μηχανές Αναζήτησης.....	12
2.3.1 Βασικές Διαφορές των Μηχανών Αναζήτησης	13
2.3.2 Πληρότητα και Προσεγγιστικότητα των Μηχανών Αναζήτησης	14
2.3.3 Κύρια χαρακτηριστικά των σημαντικότερων Μηχανών Αναζήτησης....	15
ΚΕΦΑΛΑΙΟ 3	23
Ανάκτηση Πληροφοριών και Βάσεις Δεδομένων	23
3.1 Γενικά	23
3.2 Μια τρίτη προσέγγιση: Υλοποίηση IRS με τη βοήθεια DBMS	24
3.2.1 Το Σύστημα βιβλιογραφικών εγγραφών retriev	25
3.2.1.1 Εισαγωγή βιβλίου	26
3.2.1.2 Επιλογές αναζήτησης.....	27
3.2.2 Συμπεράσματα	30
3.3 Περίπτωση της Oracle	30
3.3.1 Χαρακτηριστικά του Oracle ConText Option 7.3	31
3.3.1.1 Στήλες Κειμένου	31
3.3.1.2 Βαθμολογία ανακτηθέντων κειμένων.....	35
3.3.1.3 Επιλογές Αναζήτησης.....	36
3.3.1.4 Γλωσσολογικές υπηρεσίες	40
ΚΕΦΑΛΑΙΟ 4	44
Σύστημα Ανάκτησης Ελληνικών Κειμένων	44
4.1 Γενικά	44
4.2 Δημιουργία της Βάσης.....	45
4.2.1 Αρχιτεκτονική και Λειτουργία του Web Crawler	45
4.3 Αρχιτεκτονική του Συστήματος Ανάκτησης	53
4.4 Λειτουργία του Συστήματος Ανάκτησης.....	59
ΚΕΦΑΛΑΙΟ 5	64
Αξιολόγηση Αποτελεσμάτων	64
5.1 Αποτελεσματικότητα του Συστήματος Ανάκτησης.....	64
5.1.1 Αποτελέσματα Πειράματος	68
5.1.1.1 Αποτελέσματα με αποκοπή καταλήξεων στους όρους	70
5.1.1.2 Αποτελέσματα χωρίς αποκοπή καταλήξεων στους όρους	73

ΚΕΦΑΛΑΙΟ 1

Ανάκτηση Πληροφοριών

1.1 Αντικείμενο της Ανάκτησης Πληροφοριών

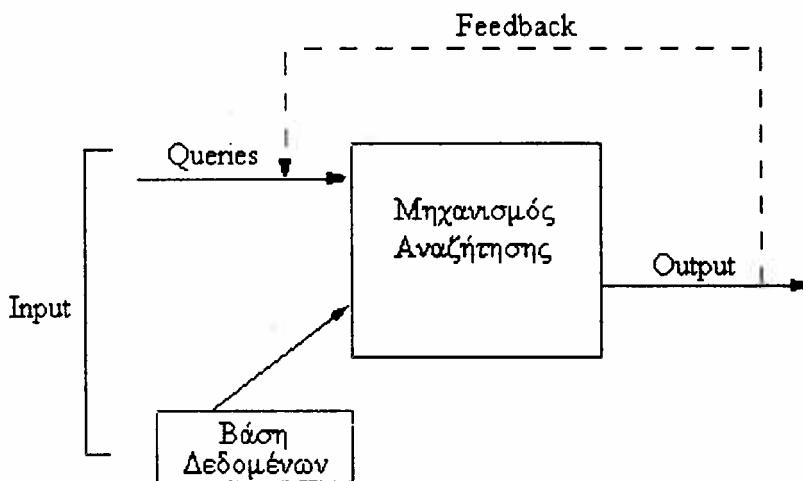
Το πρόβλημα της ανάκτησης πληροφοριών (Information Retrieval) υπάρχει ως ερευνητικό πεδίο απασχολώντας την επιστημονική κοινότητα εδώ και τρεις δεκαετίες, ενώ στις μέρες μας η ανάγκη για αποδοτική και αποτελεσματική ανάκτηση πληροφοριών αποκτά ιδιαίτερη βαρύτητα λόγω του όγκου και της ποικιλίας των πηγών της. Παραδοσιακές εφαρμογές της ανάκτησης πληροφοριών συναντάμε σε οργανισμούς, υπηρεσίες, βιβλιοθήκες πανεπιστήμια κ.α. όπου υπάρχει συνήθως μεγάλος όγκος πληροφορίας σε μορφή κειμένου.

Η ανάκτηση πληροφοριών γνωστή στη διεθνή βιβλιογραφία ως Text Retrieval αφορά την αναζήτηση και ανάκτηση της σχετικής πληροφορίας μέσα από μία συλλογή κειμένων που είναι σε μορφή ελεύθερου κειμένου φυσικής γλώσσας. Με άλλα λόγια, ο χρήστης θέτει τις πληροφοριακές του ανάγκες υπό την μορφή μιας ερώτησης (query), ενδεχομένως σε φυσική γλώσσα, και το σύστημα ανάκτησης πληροφοριών (Information Retrieval System - IRS ή Text Retrieval System - TRS) αναλαμβάνει να επιστρέψει στον χρήστη εκείνα τα κείμενα που θεωρεί ότι είναι σχετικά (relevant) σε σχέση πάντα με τις πληροφορίες που ζητάει. Το πόσο καλό είναι ένα σύστημα ανάκτησης φαίνεται από τα αποτελέσματα που δίνει δηλαδή από τα κείμενα που επιστρέφει. Μερικά από αυτά μπορεί τελικά να είναι σχετικά με την ερώτηση του χρήστη ενώ άλλα μπορεί να μην είναι σχετικά. Κεντρικό σημείο δηλαδή στη όλη στρατηγική της ανάκτησης πληροφοριών είναι το κατά πόσο σχετικά είναι ή όχι τα κείμενα που επιστρέφονται σε σχέση με την ερώτηση του χρήστη και γίνεται συνεχής προσπάθεια να περιοριστεί ο αριθμός των μη-σχετικών κειμένων και να αυξηθεί ο αριθμός των σχετικών κειμένων στα αποτελέσματα.

1.2 Δομή ενός IRS

Η δομή και η λειτουργία ενός IRS φαίνεται στο σχήμα 1.1. Σε γενικές γραμμές ένα σύστημα ανάκτησης πληροφοριών αποτελείται από δύο βασικά μέρη: Τη βάση

δεδομένων και τον μηχανισμό αναζήτησης. Η βάση δεδομένων περιλαμβάνει τα κείμενα σε μορφή ελεύθερου κειμένου και κάποιες εσωτερικές δομές δεδομένων που χρησιμοποιούνται για την αναπαράσταση και αναζήτηση των κειμένων. Η μηχανή αναζήτησης είναι υπεύθυνη για την ανάκτηση των σχετικών κειμένων από την βάση και την επιστροφή τους ως απάντηση στην ερώτηση του χρήστη. Η αναζήτηση των σχετικών κειμένων βασίζεται στην εσωτερική αναπαράστασή τους και αυτή με την σειρά της προκύπτει από την ανάλυση των περιεχομένων των κειμένων.

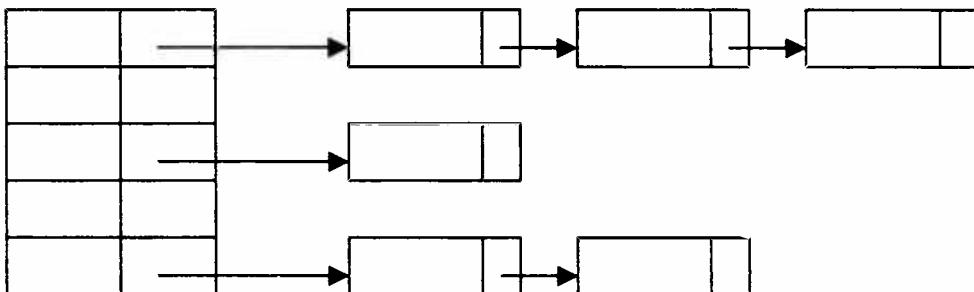
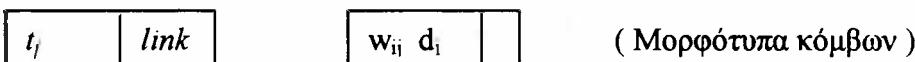


Σχήμα 1.1: Δομή και λειτουργία ενός IRS

Ορισμένες λέξεις σε κάθε κείμενο περιέχουν χρήσιμη πληροφορία και ο συνδυασμός τους είναι ικανός να αποδώσει εν μέρει το "νόημα" του κειμένου. Αυτές οι λέξεις αποτελούν τους όρους (terms) του κειμένου και σε αυτούς βασίζεται η εσωτερική αναπαράσταση των κειμένων και η λειτουργία της αναζήτησης. Είναι η στοιχειώδης πληροφορία που μπορεί να χειριστεί το σύστημα ανάκτησης. Η συχνότητα εμφάνισης ενός όρου σε ένα κείμενο εκφράζει και την σπουδαιότητά του στο αντίστοιχο κείμενο, πράγμα που επιτρέπει την αναπαράσταση του κάθε κειμένου ως ένα διάνυσμα όρων με διαφορετικό βάρος (weight) - σπουδαιότητα - ο καθένας. Ορισμένες λέξεις των κειμένων θεωρούνται ότι δεν συνεισφέρουν στο νόημα του κειμένου και μπορούν να αγνοηθούν και κατά την αναζήτηση και κατά την κατασκευή των εσωτερικών δομών δεδομένων. Αυτές οι λέξεις είναι συνήθως τα άρθρα, οι προθέσεις, οι σύνδεσμοι κ.α. και αποτελούν τις λεγόμενες τετριμμένες λέξεις (stopwords).

Η πιο συχνά χρησιμοποιούμενη εσωτερική δομή για την αποθήκευση και αναζήτηση των κειμένων είναι αυτή των ανεστραμμένων αρχείων (inverted files). Υπάρχουν και άλλες δομές που μπορούν χρησιμοποιηθούν για αυτό το σκοπό αλλά αξίζει λίγο να επιμείνουμε στα ανεστραμμένα αρχεία λόγω των πλεονεκτημάτων που παρουσιάζουν αλλά και λόγω του ότι το σύστημα ανάκτησης που υλοποιήθηκε στα πλαίσια της

παρούσας διπλωματικής εργασίας χρησιμοποιεί ανεστραμμένα αρχεία. Η μορφή των ανεστραμμένων αρχείων φαίνεται στο σχήμα 1.2 όπου υπάρχει μία λίστα - συνήθως ταξινομημένη - με όλους του όρους που υπάρχουν στα κείμενα και κάθε όρος περιέχει ένα δείκτη σε μία λίστα που περιέχει όλα τα κείμενα στα οποία συναντάται αυτός ο όρος. Κάθε ανεστραμμένη λίστα αποτελείται από ζεύγη της μορφής (αναγνωριστικό κειμένου, βάρος του όρου στο κείμενο). Τα πλεονεκτήματα των ανεστραμμένων λιστών είναι ότι υλοποιούνται εύκολα και είναι πολύ γρήγορες. Τα μειονεκτήματά τους είναι ο επιπλέον χώρος που καταλαμβάνουν και το κόστος ενημέρωσης και αναδιοργάνωσής τους συν το κόστος συγχώνευσης των λιστών των κειμένων ειδικά όταν αυτές είναι μεγάλες.



t_j : Ο j-οστός όρος

d_i : Το i-στο κείμενο

w_{ij} : Το βάρος του όρου j στο κείμενο i

Σχήμα 1.2: Ανεστραμμένο αρχείο όρων

Στη διαδικασία κατασκευής της εσωτερικής αναπαράστασης των περιεχομένων των κειμένων στη μορφή των ανεστραμμένων αρχείων εκτός από την διαδικασία απόρριψης των τετριψμένων λέξεων χρησιμοποιείται και η τεχνική της αποκοπής καταλήξεων (suffix stripping) στους όρους των κειμένων. Είναι επιθυμητό οι όροι που έχουν κοινή ρίζα να μειωθούν σε αυτή την κοινή ρίζα (stem). Για παράδειγμα οι όροι

Ειδικός

Ειδικότητα

Ειδίκευση

Ειδικευμένη

προέρχονται από το ίδιο θέμα (*Eidik*) και μπορούν να μειωθούν στο κοινό αυτό θέμα. Η διαδικασία αυτή γίνεται - παρά το γεγονός ότι είναι υπολογιστικά δαπανηρή - για δύο κυρίως λόγους:

1. Η λεκτική συγγένεια των λέξεων έχει αποδειχθεί ότι στην πλειοψηφία των περιπτώσεων συνεπάγεται και την σημασιολογική συγγένεια τους. Ετσι εφόσον η αναζήτηση των σχετικών κειμένων για μία ερώτηση βασίζεται σε λεκτικά κριτήρια είναι προς όφελος των αποτελεσμάτων η αναζήτηση να βασίζεται στην ρίζα των όρων και όχι σε ολόκληρο των όρο.
2. Η αποκοπή των καταλήξεων στους όρους δεν αποτελεί ξεκάθαρα μία γραμματολογική διαδικασία. Σκοπός της είναι να φέρει όσο το δυνατόν περισσότερες λέξεις με κοινό νόημα στην ίδια ρίζα, με απώτερο στόχο την οικονομία στο χώρο αποθήκευσης των ευρετηρίων και την βελτίωση του χρόνου αναζήτησης.^[1,4]

Χαρακτηριστικά αρκεί να αναφέρουμε ότι πειραματικά για τα Ελληνικά κείμενα έχει υπολογιστεί^[4] ότι το 47% περίπου των λέξεων σε μία συλλογή κειμένων είναι τετριμμένες λέξεις, ενώ από το 53 % που απομένει, το 45% μπορεί να μειωθεί σε κοινές ρίζες μέσο της διαδικασίας αποκοπής καταλήξεων (stemming).

1.3 Λειτουργία ενός IRS

Όπως φαίνεται και στο σχήμα 1.1, ο μηχανισμός αναζήτησης του συστήματος ανάκτησης χρησιμοποιεί σαν δεδομένα εισόδου την ερώτηση του χρήστη και την εσωτερική αναπαράσταση των κειμένων προκειμένου να εκτελέσει την αναζήτηση.

Για να είναι δυνατή η ανάκτηση των κειμένων απαιτείται η αναπαράσταση του ερωτήματος του χρήστη σε μορφή συμβατή με την εσωτερική αναπαράσταση των κειμένων ώστε να μπορεί να πραγματοποιηθεί κάποιουν είδους σύγκριση για να επιστραφούν τα σχετικά κείμενα. Συνήθως το μοντέλο στο οποίο βασίζεται αυτή η σύγκριση είναι αυτό του Διανυσματικού Χώρου (Vector Space Model)^[1] στο οποίο κάθε κείμενο και ερώτηση του χρήστη αναπαρίστανται με ένα διάνυσμα δεικτοδοτημένων όρων όπου ο κάθε όρος έχει διαφορετική βαρύτητα. Ο τρόπος με τον οποίο υπολογίζονται τα βάρη των όρων στα διάφορα κείμενα ποικίλει και βασίζεται στην στατιστική ανάλυση της συχνότητας εμφάνισης των όρων μέσα στα κείμενα μαζί πολλές φορές με την βοήθεια γραμματολογικών κανόνων που αναδεικνύουν την σχέση μεταξύ των διαφόρων όρων.

Το τελικό ζητούμενο στη διαδικασία αναζήτησης είναι η εξεύρεση εκείνων των κειμένων των οποίων το διάνυσμα αναπαράστασης τους είναι κοντά στο διάνυσμα της ερώτησης του χρήστη. Το πόσο κοντά είναι τα δύο διανύσματα εκφράζεται από την γωνία Φ που σχηματίζεται μεταξύ τους ή πιο καλά από το συνημίτονο αυτής της

γωνίας και αποτελεί το λεγόμενο μέτρο ομοιότητας (*similarity measure*) του διανύσματος της ερώτησης q και των κειμένων της βάσης d_i

$$sim(q, d_i) = \cos \Phi$$

Το συνημίτονο της γωνίας Φ εκφράζει την ομοιότητα της ερώτησης του χρήστη με το κάθε κείμενο και μπορεί να πάρει τιμές στο διάστημα $[0,1]$ όπου υψηλότερη τιμή φανερώνει και μεγαλύτερο βαθμό σχετικότητας του αντίστοιχου κειμένου με την ερώτηση.

Η λειτουργία του μηχανισμού αναζήτησης κατά τον τρόπο που αναφέραμε βασίζεται σε δύο βασικές παραδοχές :

1. **Θεωρούμε** ότι εάν ένα κείμενο είναι σχετικό με μία έννοια, τότε θα περιέχει λέξεις ή τμήματα λέξεων σχετικά με τη έννοια αυτή.
2. **Θεωρούμε** ότι η λεκτική συγγένεια των λέξεων προδίδει και εννοιολογική συγγένεια

Παρά το γεγονός ότι τις περισσότερες φορές, οι λέξεις ενός κειμένου ή πιο καλά οι όροι του μπορούν να προσδιορίσουν σε μεγάλο βαθμό τα περιεχόμενα του, υπάρχουν περιπτώσεις στις οποίες να μην ισχύουν απόλυτα τα παραπάνω. Τα αποτελέσματα ενός συστήματος ανάκτησης που βασίζεται μόνο στους όρους των κειμένων για την εκτέλεση της αναζήτησης, εξαρτώνται άμεσα από τα χαρακτηριστικά και τις ιδιαιτερότητες που παρουσιάζουν αυτοί οι όροι των κειμένων. Έτσι, η ύπαρξη πολύ γενικών ή αντίθετα πολύ ειδικών όρων στη συλλογή, λέξεις συνώνυμες ή λέξεις με περισσότερα από ένα νοήματα αποτελούν τους βασικούς παράγοντες που επηρεάζουν της επίδοση ενός IRS που βασίζεται μόνο στους όρους των κειμένων.

Για το ξεπέρασμα των παραπάνω προβλημάτων και την βελτίωση της επίδοσης των IRS, έχει γίνει προσπάθεια η αναζήτηση να μην βασίζεται εξ ολοκλήρου σε ανεξάρτητους όρους αλλά να λαμβάνεται υπόψη η συσχέτιση μεταξύ των όρων των κειμένων διαμέσου σχηματισμού φράσεων από όρους των κειμένων ή τη χρήση θησαυρού στην ερώτηση του χρήστη. Στην πρώτη περίπτωση προσπαθούμε τους όρους που είναι πολύ γενικοί (term exhaustivity) - έχουν μεγάλη συχνότητα μέσα στην συλλογή - να τους κάνουμε πιο ειδικούς (term specificity) συνδυάζοντας τους για τον σχηματισμό φράσης με άλλους όρους που έχουν συνήθως μικρότερη συχνότητα. Για παράδειγμα, η φράση computer graphics είναι πιο εξειδικευμένη από τις λέξεις computer και graphics ξεχωριστά. Η δημιουργία φράσεων σημαίνει "στένεμα" των όρων και αυξάνει της ακρίβεια (precision) των αποτελεσμάτων. Το αντίθετο κάνει ο θησαυρός. Ο θησαυρός ομαδοποιεί ειδικούς όρους κάτω από πο γενικούς δηλαδή ομαδοποιεί όρους χαμηλής συχνότητας σε κλάσεις υψηλότερης συχνότητας. Με αυτό τον τρόπο "πλατειάζει" την ερώτηση του χρήστη αυξάνοντας την πιθανότητα ταιριάσματος σε σύγκριση με τον αρχικό όρο και μεγαλώνει την απόκριση (recall) του συστήματος ανάκτησης.

Θα πρέπει να πούμε ότι τα παραπάνω αποτελούν ένα βασικό μοντέλο ενός συστήματος ανάκτησης καθώς υπάρχουν και έχουν προταθεί κατά καιρούς και άλλοι

τρόποι και τεχνικές αναζήτησης^[1,3,5]. Μερικοί από τους πιο γνωστούς περιλαμβάνουν την εισαγωγή προχωρημένων τεχνικών επεξεργασίας φυσικής γλώσσας (Natural Language Processing) κατά την δεικτοδότηση και αναζήτηση των κειμένων, την εφαρμογή τεχνικών έμπειρων συστημάτων (Expert Systems) κατά την φάση της αναζήτησης σχετικά με την επιλογή της βέλτιστης στρατηγικής αναζήτησης ανάλογα με το είδος της ερώτησης, την υιοθέτηση και χρήση διαφορετικών δομών για την εσωτερική αναπαράσταση των κειμένων όπως αρχεία υπογραφών, την εφαρμογή επεκταμένης λογικής BOOL (Extended Boolean Retrieval) στην διατύπωση της ερώτησης του χρήστη με αυτόματη ή χειρονακτική απόδοση βαρών στους λογικούς τελεστές κ.α.

Οπως φαίνεται και στο σχήμα 1.1, σε ορισμένα συστήματα ανάκτησης πραγματικού χρόνου (on-line) είναι εφικτό για τον χρήστη να τροποποιήσει τα κριτήρια αναζήτησής του κατά την διάρκεια μια συνόδου (session) αναζήτησης προκειμένου να πετύχει καλύτερα αποτελέσματα από την αρχική ερώτηση που έκανε. Αυτή η διαδικασία επαναδιατύπωσης και καλυτέρευσης του ερωτήματος είναι γνωστή ως επανατροφοδότηση (feedback).

Τέλος, το σύστημα ανάκτησης αφού πραγματοποιήσει την αναζήτηση βασισμένη σε κάποια από τις τεχνικές που περιγράψαμε παραπάνω, επιστρέφει τα αποτελέσματα στο χρήστη υπό την μορφή λίστας με τα αναγνωριστικά των κειμένων που βρέθηκαν να ικανοποιούν τα κριτήρια αναζήτησης μαζί με το αντίστοιχο για κάθε κείμενο βαθμό σχετικότητας που δείχνει το μέτρο στο οποίο το επιτυγχάνουν αυτό. Συνήθως τα κείμενα επιστρέφονται κατά φθίνουσα σειρά σχετικότητας (relevance ranking) ώστε να διευκολύνουν το χρήστη στο διάβασμα των αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 2

Μηχανές Αναζήτησης στο Διαδίκτυο και Ανάκτηση Πληροφοριών

2.1 Γενικά

Το World Wide Web αποτελεί σήμερα το κυριότερο μέσο δημοσίευσης πληροφοριών στο διαδίκτυο (internet) με άμεσο αποτέλεσμα ο όγκος των πληροφοριών που υπάρχει να είναι ήδη τεράστιος ενώ αναμένεται να συνεχίζει να αυξάνεται με ανεξέλεγκτους ρυθμούς και στα επόμενα χρόνια. Με την μεγάλη ανάπτυξη του WWW, τόσο σε αριθμό περιοχών (sites) όσο και σε αριθμό σελίδων που αυτά περιέχουν, έγινε προφανές ότι ο παραδοσιακός μέχρι πρότινος τρόπος αναζήτησης πληροφοριών στο διαδίκτυο μέσο της τυχαίας περιήγησης (surfing) του χρήστη ανάμεσα στις Web σελίδες ακολουθώντας τυχαία συνδέσμους (links) και μεταπηδώντας από την μία στην άλλη, δεν ήταν πλέον ικανός να οδηγήσει τον ενδιαφερόμενο στη επιθυμητή πληροφορία άμεσα και γρήγορα, ενώ σήμερα πάνει να είναι και πρακτικός σε σχέση πάντα με τον όγκο και τις πηγές των πληροφοριών που υπάρχουν.

Για το σκοπό αυτό, τα τελευταία χρόνια ένας αριθμός από καινούργια εργαλεία έχουν αναπτυχθεί προκειμένου να γίνει πιο αποδοτική και αποτελεσματική η αναζήτηση των επιθυμητών πληροφοριών στο WWW. Ανεξάρτητα από το είδος των εργαλείων που χρησιμοποιούμε για την ανάκτηση των πληροφοριών είτε αυτές βρίσκονται σε διάφορες πηγές στο WWW είτε όχι, η ίδια η ανάκτηση έχει ως αποτέλεσμα τη διαίρεση του χώρου των κειμένων σε δύο μέρη: Στα κείμενα-σελίδες που ανακτήθηκαν και σε αυτές που δεν ανακτήθηκαν. Αυτά τα δύο μέρη περιέχουν σελίδες σχετικές ή όχι σε σχέση με την πληροφορία που αναζητούνται, η καθεμία με διαφορετικό βαθμό σχετικότητας ως προς την ερώτηση. Η *Ακρίβεια* μετρά πόσο καλά τα ανακτήθέντα κείμενα-σελίδες ανταποκρίνονται σε εκείνο που αναζητεί ο τελικός χρήστης π.χ. το ποσοστό των σχετικών κειμένων επί του συνόλου των ανακτήθέντων. Η *Απόκριση* μετρά πόσα σχετικά κείμενα από το σύνολο των σχετικών τελικά ανακτήθηκαν. Προκειμένου να βελτιωθεί η ακρίβεια που προσφέρουν αυτά τα εργαλεία, τα ανακτήθέντα κείμενα συχνά κατατάσσονται με ένα υποτιθέμενο βαθμό σχετικότητας ως προς την ερώτηση, που υπολογίζεται με διάφορες τεχνικές ανάλογα με το εργαλείο αναζήτησης που χρησιμοποιούμε. Πάντως για την ακριβή μέτρηση

της ακρίβειας και της απόκρισης απαιτείται η ακριβής γνώση όλων των περιεχομένων των κειμένων-σελίδων σε σχέση πάντα με τις πληροφοριακές ανάγκες του χρήστη που κάνει την αναζήτηση. Τα υπάρχοντα συστήματα μπορούν μονάχα να προσεγγίσουν και όχι να υπολογίσουν την ακρίβεια και τη απόκριση λόγω του είδους της συλλογής στη οποία γίνεται η αναζήτηση (εκατομμύρια Web σελίδες με ευμεταβλητό περιεχόμενο και διεύθυνση) αλλά και λόγω του μεγέθους της (μεγάλος αριθμός σελίδων που συνεχώς αυξάνεται). Ο κύριος στόχος κάθε συστήματος ανάκτησης πληροφοριών, είτε η συλλογή του βασίζεται σε κείμενα που προέρχονται από το WWW είτε όχι, είναι η μεγιστοποίηση της ακρίβειας και της απόκρισης με ταυτόχρονη ελάττωση του υπολογιστικού κόστους και της δυσκολίας χρήσης του από το τελικό χρήστη.

Τα εργαλεία που υπάρχουν σήμερα για την αναζήτηση πληροφοριών στο WWW, ακολουθούν δύο διαφορετικές προσεγγίσεις:

1. Αναζήτηση πληροφοριών μέσο περιήγησης του χρήστη σε θεματολογικά δέντρα και ιεραρχίες-κατηγορίες πληροφοριών, και
2. Αναζήτηση με τη χρήση κριτηρίων αναζήτησης συνήθως όρων, λέξεων, φράσεων κ.λ.π. χρησιμοποιώντας μηχανές αναζήτησης (search engines).

Θα μπορούσαμε εδώ να αναφέρουμε και μία τρίτη υβριδική προσέγγιση, τις λεγόμενες μετά-μηχανές αναζήτησης (meta search engines), οι οποίες συγχωνεύουν και φίλτράρουν κατά κάποιο τρόπο τα αποτελέσματα αναζήτησης των μηχανών των δύο πρώτων κατηγοριών και επιστρέφουν τα "βελτιστοποιημένα" αποτελέσματα στο τελικό χρήστη.

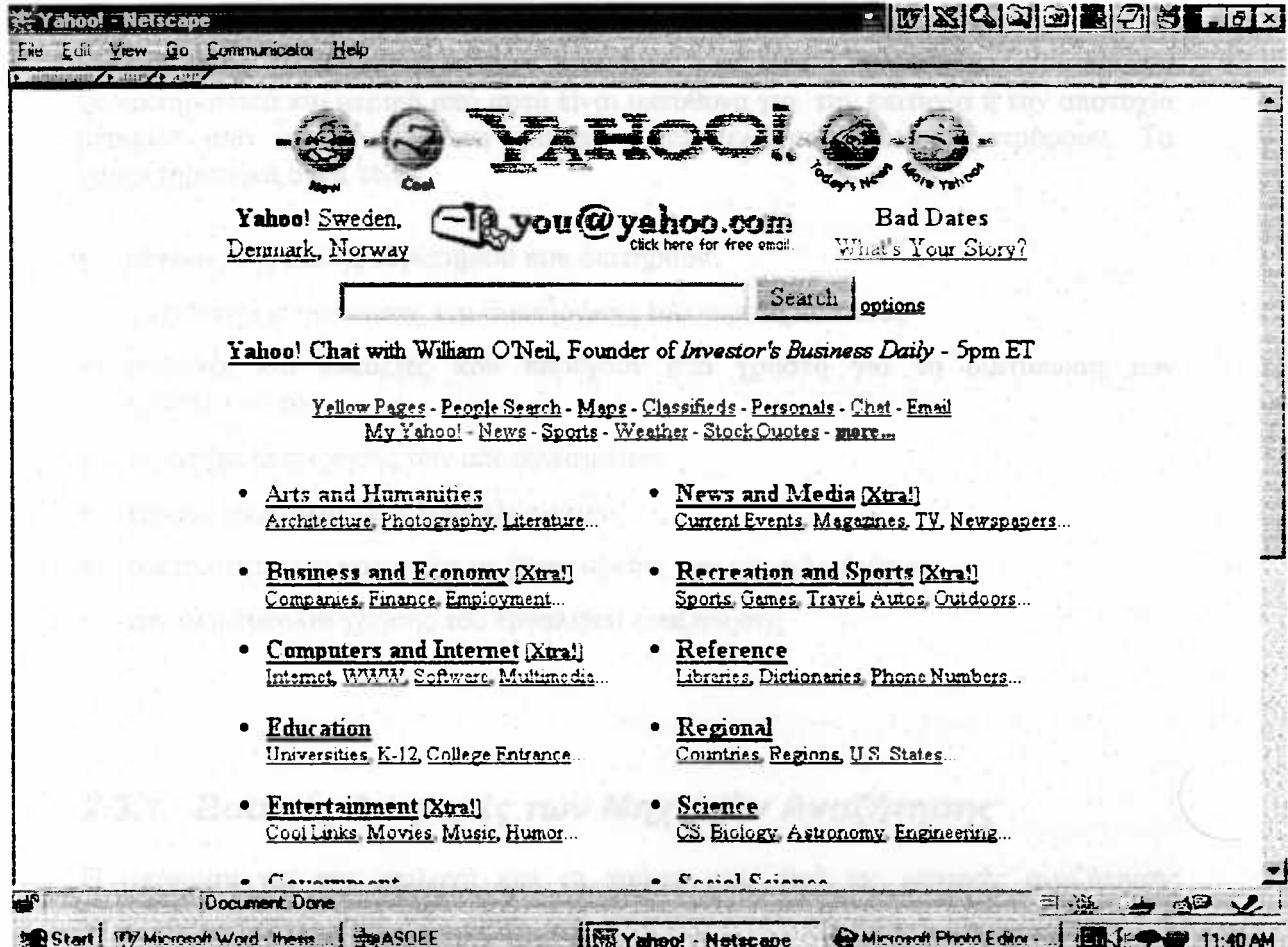
2.2 Θεματολογικά Δέντρα και Ιεραρχίες Πληροφοριών

Τα θεματολογικά δέντρα είναι εργαλεία που παρέχουν μια δομημένη και οργανωμένη ιεραρχία από κατηγορίες πληροφοριών κατά θεματικές ενότητες. Κάτω από κάθε κατηγορία ή υποκατηγορία υπάρχουν οι λίστες με τους συνδέσμους προς τις αντίστοιχες σελίδες που περιέχουν την πληροφορία. Η κάθε σελίδα κατατάσσεται στη ανάλογη κατηγορία είτε ορίζοντας το θέμα της ο δημιουργός της σελίδας είτε χειρονακτικά από τους διαχειριστές και τους υπεύθυνους για τη συντήρηση αυτών των θεματολογικών δέντρων. Επιτλέον μερικά από αυτά τα εργαλεία παρέχουν και τη δυνατότητα αναζήτησης πληροφοριών στο θεματολογικό δέντρο με βάση όρους και λέξεις κατά τρόπο ανάλογο με αυτόν των μηχανών αναζήτησης.

Τα πιο διάσημα εργαλεία που ανήκουν σε αυτή την κατηγορία είναι το Yahoo, Galaxy και η WWW Virtual Library. Αυτά τα εργαλεία έχουν ένα μεγάλο αριθμό



από θεματικές ενότητες που οι κάθε μία περιλαμβάνει πολλές υποκατηγορίες και ένα μεγάλο αριθμό τελικών κειμένων-σελίδων (βλ. εικόνα 2.1).



Εικόνα 2.1: Το θεματολογικό δέντρο Yahoo. Διακρίνονται οι κατηγορίες πληροφοριών και η δυνατότητα αναζήτησης μέσα σε αυτές.

2.3 Μηχανές Αναζήτησης

Η λειτουργία των μηχανών αναζήτησης βασίζεται στην διατήρηση και ενημέρωση ευρετηρίων (indexes) που περιέχουν τις λέξεις-όρους των κείμενων-σελίδων που υπάρχουν στο WWW. Για τη δημιουργία και την ενημέρωση αυτών των ευρετηρίων χρησιμοποιούνται ειδικού σκοπού προγράμματα γνωστά ως "Web Robots", "Web Wanderers", "Web Worms", "Spiders", ή "Web Crawlers"^[6]. Αυτά τα προγράμματα ταξιδεύουν στο διαδίκτυο αρχίζοντας με την ανάκτηση ενός αρχικού κειμένου-σελίδας και αναδρομικά κάνουν το ίδιο πράγμα για όλους τους συνδέσμους που υπάρχουν σε αυτή τη σελίδα κ.ο.κ. και με αυτό το τρόπο ανακαλύπτονται καινούργια, κείμενα-σελίδες που θα πρέπει να ανακτηθούν αλλά και ενημερώνονται τα ήδη

δεικτοδοτημένα για τυχόν αλλαγές. Τα κείμενα-σελίδες που ανακαλύπτει το Web Robot στην συνέχεια δίνονται στο τμήμα εκείνο του λογισμικού που είναι υπεύθυνο για τη δεικτοδότηση των κειμένων και την ενημέρωση της βάσης-ευρετηρίου και κατ' αυτό τον τρόπο χτίζεται και συντηρείται το ευρετήριο πάνω στο οποίο κάνει τις αναζητήσεις η μηχανή αναζήτησης.

Οι μηχανές αναζήτησης που υπάρχουν σήμερα διαφέρουν σε αρκετά χαρακτηριστικά και μερικά από αυτά είναι υπεύθυνα για την επιτυχία ή την αποτυχία μερικών από αυτές με βάση πάντα τα αποτελέσματα που επιστρέφουν. Τα χαρακτηριστικά αυτά είναι:

- μέγεθος της βάσης-ευρετηρίου που διατηρούν
- συχνότητα ενημέρωσης και συντήρησης του ευρετηρίου τους
- επιλογές και ευκολίες που παρέχουν στο χρήστη για τη διατύπωση των ερωτήσεων του
- ταχύτητα επιστροφής των αποτελεσμάτων
- τρόπος εμφάνισης των αποτελεσμάτων
- σχετικότητα και κατάταξη με βάση αυτήν, των αποτελεσμάτων
- την όλη ευκολία χρήσης του εργαλείου αναζήτησης

2.3.1 Βασικές Διαφορές των Μηχανών Αναζήτησης

Η απόφαση για την επίλογή και τη χρήση μιας από τις μηχανές αναζήτησης εξαρτάται μερικώς από τον τρόπο που δεικτοδοτεί τις σελίδες η κάθε μηχανή. Για παράδειγμα, η μηχανή αναζήτησης WebCrawler δεικτοδοτεί κάθε λέξη των Web σελίδων, ενώ η μηχανή Lycos κατασκευάζει το ευρετήριο από ειδικά επιλεγμένες λέξεις όπως αυτές που βρίσκονται στον τίτλο, στις επικεφαλίδες και από τις 100 πιο σημαντικές λέξεις του κάθε κειμένου^[7,9]. Αυτή η διαφορά στο τρόπο δεικτοδότησης έχει συχνά ως αποτέλεσμα την επιστροφή εντελώς διαφορετικών αποτελεσμάτων για την ίδια ερώτηση του χρήστη.

Επίσης, παρατηρούνται διαφορές στις επιλογές που δίνονται στο χρήστη για τη διατύπωση της ερώτησης του. Μερικές μηχανές χρησιμοποιούν τον τελεστή "or" ως τον εξ ορισμού τελεστή μεταξύ των όρων της ερώτησης του χρήστη και βασίζονται σε ειδικούς αλγορίθμους για την εξεύρεση και κατάταξη των αποτελεσμάτων. Άλλες μηχανές αναζήτησης παρέχουν τους εναλλακτικούς τελεστές "and", "adjacent", "near", "nol" κ.α. ενώ μερικές από αυτές παρέχουν πιο προχωρημένες τεχνικές όπως αποκοπής καταλήξεων, αναζήτηση με βάση μια φράση (phrase searching) ή μια έννοια (concept searching) κ.α.

Η ακρίβεια των αποτελεσμάτων που επιστρέφουν οι μηχανές αναζήτησης εξακολουθεί να αποτελεί πρόβλημα και αντικείμενο μελέτης. Οι μηχανές που προσφέρουν πιο προχωρημένες επιλογές διατύπωσης του ερωτήματος επιστρέφουν συνήθως και καλύτερα αποτελέσματα. Μερικές μηχανές αναζήτησης είναι πιο καλές για ορισμένου είδους ερωτήσεις και ορισμένες πιο καλές για άλλου είδους. Για παράδειγμα, η έλλειψη του τελεστή "adjacency" για φράσεις ή για λέξεις στη μηχανή WebCrawler την κάνει ακατάλληλη για ερώτηση σχετικά με ένα συγγραφέα ή ένα ονοματεπώνυμο. Σε μια τέτοια ερώτηση, εφόσον τον όνομα και το επώνυμο θεωρούνται ότι είναι δύο ξεχωριστές και ασυσχέτιστες λέξεις μεταξύ τους, η πιθανότητα για καλά αποτελέσματα είναι πολύ μικρή.

Σε αντίθεση, η δεικτοδότηση όλων των λέξεων των σελίδων που πραγματοποιεί ο WebCrawler, μπορεί να είναι πολύ αποτελεσματική στην αναζήτηση ενός συγκεχυμένου (obscure) όρου. Επίσης και το μέγεθος της βάσης-ευρετηρίου παίζει μεγάλο ρόλο στη ποιότητα των αποτελεσμάτων. Για παράδειγμα, η μηχανή Lycos παρόλο που προσφέρει ελάχιστες επιλογές στην διατύπωση του κριτηρίου αναζήτησης, έχει συνήθως καλά αποτελέσματα εκεί που οι άλλες μηχανές αποτινγχάνουν, λόγω του ότι έχει δεικτοδοτήσει ένα τεράστιο αριθμό από κείμενα-σελίδες.

Κάθε μηχανή αναζήτησης διαφέρει σημαντικά στο τρόπο με τον οποίο κατατάσσει και παροιμιάζει τα αποτελέσματα. Τα κείμενα που βρίσκονται πιο ψηλά στη κατάταξη από άποψη σχετικότητας δε σημαίνει ότι κατ' ανάγκην θα μας δώσουν και την επιθυμητή πληροφορία, ενώ πολλές φορές συμβαίνει ο πιο καλός σύνδεσμος να βρίσκεται στις πιο κάτω θέσεις στη λίστα των αποτελεσμάτων. Οι πληροφορίες που επιστρέφονται μαζί με κάθε σύνδεσμο συνεισφέρει στην απόφαση μας στο εάν θα ακολουθήσουμε το σχετικό σύνδεσμο ή όχι. Για παράδειγμα, η μηχανή WebCrawler επιστρέφει τα αποτελέσματα στη μορφή μια λίστας από συνδέσμους χωρίς πρόσθετη πληροφορία. Από την άλλη μεριά, η μηχανή Lycos παράγει μια - δημιουργημένη αυτόματα από υπολογιστή - σύντομη περιγραφή για κάθε σύνδεσμο πράγμα που διευκολύνει την απόφαση του χρήστη στο εάν θα πρέπει να τον ακολουθήσει ή όχι. Η μηχανή Opentext προχωρεί ακόμα περισσότερο και προτείνει στον χρήστη συνδέσμους προς ορισμένες περιοχές που θεωρεί ότι περιέχουν σελίδες σχετικές με το ερώτημα.

2.3.2 Πληρότητα και Προσεγγιστικότητα των Μηχανών Αναζήτησης

Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων που συνθέτουν ένα "κλειστό κόσμο", στον οποίο μία αρνητική απάντηση σε ένα ερώτημα συνεπάγεται τη μη εξεύρεση απάντησης στην βάση που να ικανοποιεί το κριτήριο αναζήτησης, το WWW αποτελεί ένα ανοικτό και διαρκώς μεταβαλλόμενο κόσμο όπου μία αρνητική απάντηση σε ένα ερώτημα σημαίνει ότι οι πληροφορίες που ικανοποιούν αυτό το ερώτημα δε μπορούν να προσεγγιστούν από τη μηχανή αναζήτησης και όχι κατ' ανάγκη ότι δεν υπάρχουν. Μπορούμε δηλαδή να πούμε ότι κάθε αναζήτηση στα

WWW θεωρείται ατελής. Κάθε μηχανή αναζήτησης χαρακτηρίζεται από το ποσοστό της πληρότητας δηλαδή από το σύνολο των κειμένων-σελίδων που έχει δεικτοδοτήσει και μπορεί να προσεγγίσει. Η προσέγγιση όμως ενός κειμένου δεν εξασφαλίζεται ούτε επιτυχάνεται κατ' ανάγκη με την πληρότητα της μηχανής αναζήτησης. Ετσι παρόλο που ένα κείμενο-σελίδα μπορεί να βρίσκεται μέσα στο πληροφοριακό χώρο που έχει δεικτοδοτήσει η μηχανή αναζήτησης, η αφαίρεση από αυτό λέξεων ή και όρων που θεωρήθηκαν περιττοί κατά τη διαδικασία της κατηγοριοποίησης ή της δεικτοδότησης, μπορεί να κάνουν το συγκεκριμένο κείμενο μη προσεγγίσιμο ή ορισμένους όρους του μη χρησιμοποιήσιμους. Συμπεραίνουμε δηλαδή ότι η πληρότητα και η προσέγγιστικότητα μιας μηχανής αναζήτησης αποτελεί τη βάση και την απαραίτητη προϋπόθεση για την εφαρμογή των επιλογών αναζήτησης του χρήστη και επιστροφής των σωστών αποτελεσμάτων. Υπάρχουν διάφοροι λόγοι που ευθύνονται για το επίπεδο της πληρότητας και προσέγγιστικότητας μιας μηχανής αναζήτησης^[8].

- Το κείμενο μπορεί να είναι σε μορφή που δεν μπορεί να κατανοηθεί από τη μηχανή αναζήτησης. Για παράδειγμα, το κείμενο-σελίδα μπορεί να περιέχει εντολές έξω από αυτές που ορίζει το πρότυπο standard HTML (HyperText Markup Language) και να είναι μη αναγνώσιμο.
- Το κείμενο μπορεί να προστατεύεται με συνθηματικό ή να βρίσκεται κάτω από την προστασία FireWall. Τέτοιου είδους κείμενα μπορεί να είναι προσεγγίσιμα από τον εξουσιοδοτημένο χρήστη όχι όμως και από τη μηχανή αναζήτησης και πιο συγκεκριμένα το Web Robot.
- Εάν ο μηχανισμός δεικτοδότησης μιας μηχανής αναζήτησης αποφασίσει ότι ένας όρος του κειμένου δεν είναι αρκετά σημαντικός ώστε να δεικτοδοτηθεί, τότε το κείμενο αυτό γίνεται μη προσεγγίσιμο με κριτήριο αναζήτησης αυτόν τον όρο. Αυτό το φαινόμενο συναντάται σε iεραρχικές μηχανές αναζήτησης όπου στα διάφορα επίπεδα θεματικής κατηγοριοποίησης του κειμένου απορρίπτονται αρκετοί όροι του κειμένου. Το ίδιο συμβαίνει και όταν δεικτοδοτούνται μόνο οι λέξεις των τίτλων των κειμένων και όχι όλες οι λέξεις.
- Κάθε μέρα δημιουργούνται και δημοσιεύονται καινούργια κείμενα τα οποία δεν έχουν ακόμα ανακαλυφθεί από τα Web Robots και δεικτοδοτηθεί από τις μηχανές αναζήτησης.
- Κείμενα τα οποία παράγονται δυναμικά π.χ. από CGI scripts, δεν είναι δυνατό να δεικτοδοτηθούν. Επίσης σύνδεσμοι που υπάρχουν σε Frames ή σε Image maps των Web σελίδων δεν είναι εύκολο να ανακαλυφθούν από τα Web Robots και κατ' επέκταση να προσπελαστούν.

2.3.3 Κύρια χαρακτηριστικά των σημαντικότερων Μηχανών Αναζήτησης

Παρακάτω (βλ. πίνακες 2.2-2.6) παραθέτουμε τα κυριότερα χαρακτηριστικά των έξι πιο σημαντικών μηχανών αναζήτησης που υπάρχουν σήμερα στο Internet^[10].

ΓΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Search Engine	AltaVista	Excite	HotBot	InfoSeek	Lycos	Web Crawler
Size (pages in mills)	Big (100)	Big (55)	Big (80)	Medium (30)	Medium (30)	Small (2)
Pages crawled per day	10 million	3 million	Up to 10 million	-	6 to 10 million	-
Freshness	1 day to 3 months	1 to 3 weeks	1 day to 2 weeks	Minutes to 2 months	1 to 2 weeks	Updated weekly
Date	Yes	No	File Date	No	Yes (via detailed display)	No

Πίνακας 2.2: Γενικά χαρακτηριστικά

Size: Το πλήθος των σελίδων που έχουν δεικτοδοτηθεί.

Pages Crawled Per Day: Το πλήθος των σελίδων που επισκέπτεται το Web robot ανά ημέρα.

Freshness: Η συχνότητα με την οποία γίνεται η ενημέρωση του ευρετηρίου.

Date: Ημερομηνία δεικτοδότησης της σελίδας. Το File Date σημαίνει ότι καταχωρείται η ημερομηνία τροποποίησης και όχι δεικτοδότησης της αντίστοιχης σελίδας.

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ CRAWLER						
--- Παράγοντες από τους οποίους εξαρτάται το αν και το πότε θα δεικτοδοτηθεί μία σελίδα ---						
Search Engine	AltaVista	Excite	HotBot	InfoSeek	Lycos	Web Crawler
Submitted Pages	1 day	3 weeks	1 to 2 days	Within minutes	1 to 2 weeks	1 - 3 weeks
Non-submitted pages	1 to 3 months	3 weeks	2 weeks	1 - 2 months	1 to 2 weeks	Not added, in most cases
Depth	No limit	No limit	No limit	Sample	Sample	Sample
Frames Support	No	No	No	Yes	Yes	No
Image Maps	Yes	No	No	Yes	No	Yes
Password Protected Sites	No	Yes	No	Yes	Yes	No
Link Popularity	No	No	Yes	No	Yes	Yes
Learns Frequency	Yes	No	Yes	Yes	No	No
Keep Out	Robots.txt	robots.txt, both in future	Both	robots.txt	robots.txt	Both
Redirection	Redirected URL used	Redirected URL used	-	Redirected URL used	-	Redirected URL used

Πίνακας 2.3: Χαρακτηριστικά Crawler

Submitted Pages: Πόσο γρήγορα θα δεικτοδοτηθεί μια σελίδα που υποβλήθηκε για δεικτοδότηση από τον ίδιο το χρήστη. Είναι συνήθως η επιλογή Add URL που υπάρχει στην κεντρική σελίδα των περισσοτέρων μηχανών αναζήτησης.

Non-Submitted Pages : Πόσο γρήγορα θα δεικτοδοτηθούν οι σελίδες που υπάρχουν σαν σύνδεσμοι σε μία σελίδα που υποβλήθηκε για δεικτοδότηση από τον ίδιο το χρήστη.

Depth: Σε τι βάθος θα φτάσει το Web Robot ξεκινώντας από μία σελίδα που υποβλήθηκε για δεικτοδότηση από ένα χρήστη. Μπορεί να πάρει δύο τιμές :

- **No Limit:** Το Web Robot προχωρά ωσότου συγκεντρώσει όλες τις σελίδες από την αντίστοιχη Web περιοχή στην οποία ανήκει η αρχική σελίδα.
- **Sample:** Συγκεντρώνει μόνο ένα δείγμα από τις σελίδες που υπάρχουν σε αυτή τη περιοχή

Frames Support: Αν μπορεί το Web Robot να ακολουθήσει συνδέσμους που παράγονται μέσα από Frames.

Image Maps: Αν μπορεί το Web Robot να ακολουθήσει συνδέσμους που υπάρχουν μέσα σε Image maps.

Password Protected Sites: Αν μπορεί το Web Robot να "ταξιδέψει" σε περιοχές που προστατεύονται με συνθηματικό.

Link Popularity: Αν μπορεί η μηχανή αναζήτησης να βρει πόσο δημοφιλής είναι μία σελίδα μετρώντας τους συνδέσμους-αναφορές που υπάρχουν από άλλες σελίδες προς αυτήν. Πολλές μηχανές χρησιμοποιούν το παραπάνω κριτήριο για να αποφασίσουν αν θα πρέπει ή όχι να δεικτοδοτήσουν μια σελίδα.

Learns Frequency: Αν μπορεί η μηχανή αναζήτησης να μαθαίνει πόσο συχνά αλλάζουν οι σελίδες μιας περιοχής ώστε να αποφασίζεται η συχνότητα επίσκεψης αυτής της περιοχής από το αντίστοιχο Web Robot.

Keep Out: Αν το Web Robot που χρησιμοποιεί η μηχανή αναζήτησης υπακούει στις εντολές που υπάρχουν στο αρχείο robots.txt ή στα meta robots tags των web σελίδων. Το αρχείο robots.txt είναι ένα ειδικό αρχείο που υπάρχει σε Web εξυπηρετητές (servers) και αναφέρει πια μέρη της Web περιοχής δεν πρέπει να προσπελαστούν/δεικτοδοτηθούν από το Web Robot. Τα meta robots tags είναι απλά tags που υπάρχουν σε σελίδες HTML και υποδεικνύουν το αν θα πρέπει να δεικτοδοτηθεί η συγκεκριμένη σελίδα ή όχι.

Redirection: Μερικές περιοχές επανακαθοδηγούν (redirect) τους επισκέπτες σε άλλες διευθύνσεις συνήθως εκτός της περιοχής. Για παράδειγμα, κάποιος που πηγαίνει στη διεύθυνση <http://maxonline.com/webmasters/> θα σταλεί αυτόματα στην διεύθυνση <http://searchenginewatch.com/>. Αυτή η ρύθμιση δείχνει ποια από τις δύο διευθύνσεις επιλέγουν για δεικτοδότηση οι μηχανές αναζήτησης.

ΚΑΤΑΤΑΞΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ						
--- Παράγοντες που επηρεάζουν την κατάταξη των αποτελεσμάτων---						
Search Engine	AltaVista	Excite	HotBot	InfoSeek	Lycos	Web Crawler
Stop Words	Yes	Yes	Yes	No	Yes	No
Relevancy Boosters	None	3 or 4 star review	Keywords in meta tag	Keywords in meta tag	None	Keywords in titles, Link Popularity
Spam Penalty	Yes	Yes	Yes	Yes	Yes	Yes

Πίνακας 2.4: Κατάταξη των αποτελεσμάτων

Stop Words: Αν η μηχανή αναζήτησης αγνοεί τις τετριμμένες λέξεις από το περιεχόμενο των σελίδων και την συμβολοσειρά αναζήτησης του χρήστη.

Relevancy Boosters: Συνήθως οι μηχανές αναζήτησης κατατάσσουν τα αποτελέσματα κατά αύξουσα σειρά σχετικότητας με βάση την τοποθέτηση και την συχνότητα των όρων μέσα στη Web σελίδα. Αυτή η επιλογή δείχνει αν υπάρχουν και άλλοι παράγοντες που λαμβάνονται υπόψη για την αύξηση του βαθμού σχετικότητας της αντίστοιχης σελίδας όπως π.χ τα περιεχόμενα των meta-tags.

Spam Penalty: Αν η μηχανή αναζήτησης καταλαβαίνει ότι μία σελίδα προσπαθεί με δόλιο τρόπο να βρεθεί στην κορυφή τις λίστας των αποτελεσμάτων. Για παράδειγμα, η μεγάλη επανάληψη της ίδιας λέξης-όρου σε μια σελίδα κατά τρόπο που αυτό να μην είναι ορατό στον τελικό χρήστη, έχει ως αποτέλεσμα αυτή η σελίδα να έχει μεγάλο ποσοστό σχετικότητας όταν αναζητηθεί με αυτήν τη λέξη-όρο.

ΕΜΦΑΝΙΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ						
--- Παράγοντες που επηρεάζουν την εμφάνιση των αποτελεσμάτων---						
Search Engine	AltaVista	Excite	HotBot	InfoSeek	Lycos	Web Crawler
Meta Tag Support	Yes	No	Yes	Yes	Partial	Yes
Title	Page title, otherwise, "No title"	Page title, otherwise, "Untitled"	Page title, otherwise, URL	Page title, otherwise, first line on page	Page title, otherwise, first line on page	Page title, otherwise, URL
Description	Meta tag, or first few lines on page	Sentences grouped by concept; most dominant sentences extracted	Meta tag, or first few lines on page	Meta tag, or first 200 characters after <body> tag	Created based on content	Meta tag, or first 275 characters after <body> tag
Results at a time	10	10, 20, 30, 40, 50	10, 25, 50, 75, 100	10, 20 (titles only)	5, 10, 15, 20, 30, 40, 50	10, 25, 100
Display Options	Standard, Compact, Text-Only	Summaries, Titles only, Sort by site	Full (4 lines), Brief (1 line), Titles only	Summaries, Titles Only	Standard, Summary, Detailed	Titles only, Summaries

Πίνακας 2.5: Εμφάνιση των αποτελεσμάτων

Meta Tag Support: Αν η μηχανή αναζήτησης λαμβάνει υπόψη της τις πληροφορίες που υπάρχουν στα meta tags των web σελίδων.

Titles: Αναφέρει με ποιο τρόπο η μηχανή αναζήτησης δημιουργεί τους τίτλους των σελίδων στη λίστα των αποτελεσμάτων.

Descriptions: Αναφέρει με ποιο τρόπο η μηχανή αναζήτησης δημιουργεί την περιγραφή των σελίδων στη λίστα των αποτελεσμάτων

Results At A Time: Πόσα αποτελέσματα-σύνδεσμοι εμφανίζονται σε κάθε σελίδα αποτελεσμάτων.

Display Options: Αναφέρει με ποια μορφή παρουσιάζονται τα αποτελέσματα αναζήτησης.

ΛΟΙΠΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ						
Search Engine	AltaVista	Excite	HotBot	InfoSeek	Lycos	Web Crawler
URL Status Check	Displays listing	None	Semi-displays listing	Displays listing	Reports if indexed	Reports if indexed
Site Removal	Remove pages and resubmit	Remove site or install robots.txt	Install robots.txt	Remove and resubmit site or install robots.txt	-	Remove page, resubmit using Dead URL form
Crawler Name	Scooter	Architext Spider	Slurp the Web Hound	Side winder	T-Rex	Spidey
Indexes ALT text	Yes	No	No	Yes	Yes	Yes
Indexes comments	No	No	Yes	Yes	No	No
Stemming	No	No	No	Yes	Yes	No

Πίνακας 2.6: Λοιπά χαρακτηριστικά

URL Status Check: Αναφέρει αν μπορούμε να ρωτήσουμε τη μηχανή αναζήτησης σχετικά με το αν έχει δεικτοδοτηθεί μια συγκεκριμένη σελίδα ή όχι. Οι τιμές που παίρνει είναι :

- "Displays listing" σημαίνει ότι η μηχανή αναζήτησης μας επιτρέπει να δούμε τη δεικτοδότηση που υπάρχει για μια συγκεκριμένη σελίδα. Το ίδιο σημαίνει και το "semi" για το HotBot θεωρώντας ότι δεν παρέχεται το ίδιο εύκολα και απλά αυτή η δυνατότητα.
- "Reports if indexed" σημαίνει ότι παρέχεται μια φόρμα που υποδεικνύει αν η αντίστοιχη σελίδα είναι δεικτοδοτημένη αλλά δεν μπορούμε να δούμε την καταχώρηση που έχει δημιουργηθεί στο ευρετήριο.

Site Removal: Δείχνει τη συμπεριφορά των μηχανών αναζήτησης όταν σελίδες έχουν διαγραφεί από μία περιοχή ή έχουν μεταφερθεί σε μία άλλη.

- **Removal:** Οταν η μηχανή αναζήτησης ανακαλύψει ότι οι διευθύνσεις URLs (Uniform Resource Locator) των σελίδων που διαγράφηκαν ή μεταφέρθηκαν δεν είναι έγκυρα θα σβήσει από το ευρετήριο τις καταχωρήσεις που αναφέρονται σε αυτές τις σελίδες.
- **Robots.txt:** Δημιουργείται ένα καινούργιο αρχείο robots.txt όπως αναφέρθηκε παραπάνω που προειδοποιεί για αυτές τις σελίδες ή τη συγκεκριμένη περιοχή και αφαιρούνται αυτές οι σελίδες από το ευρετήριο.

ΚΕΦΑΛΑΙΟ 3

Ανάκτηση Πληροφοριών και Βάσεις Δεδομένων

3.1 Γενικά

Ο τεράστιος όγκος της πληροφορίας που είναι ευρέως διαθέσιμος σήμερα και η ποικιλία και ανομοιογένεια των πηγών της, επιβάλει την εξεύρεση ενός αποτελεσματικού τρόπου διαχείρισης της. Η αποθήκευση, ανάλυση, ανάκτηση και διανομή της πληροφορίας αποτελούν σήμερα μία από τις βασικότερες προκλήσεις που καλούνται να αντιμετωπίσουν οι σύγχρονες επιχειρήσεις. Τα σημερινά συστήματα διαχείρισης βάσεων δεδομένων (DBMS) επιτρέπουν στις επιχειρήσεις να διαχειρίζονται τα δομημένα δεδομένα (structured data) αποθηκευμένα σε πίνακες (tables) αρκετά καλά και αποτελεσματικά, ωστόσο ποσοστό μεγαλύτερου του 90% της ψηφιακής πληροφορίας που υπάρχει σήμερα αποτελείται από μη δομημένα κείμενα και έγγραφα τα οποία παράγονται από σύγχρονες εφαρμογές βάσεων δεδομένων ή και άλλες εφαρμογές όπως ηλεκτρονικό ταχυδρομείο (E-mail), σελίδες στο WWW (HTML pages), On-Line αναφορές κ.λ.π.

Κάτω υπό αυτές τις συνθήκες που επιβάλει το ίδιο το είδος και η μορφή της πληροφορίας - δομημένης πληροφορίας που μπορεί να παρασταθεί με ακρίβεια με τους υπάρχοντες τύπους δεδομένων (integer, boolean, real κ.α.) και αδόμητης στην μορφή ελεύθερου κειμένου με άγνωστο νοηματικό περιεχόμενο - δύο ήταν οι μέχρι τώρα προσεγγίσεις που μπορούσαν να ακολουθηθούν και οι οποίες ήταν ασύμβατες μεταξύ τους και δεν επέτρεπαν την ολοκλήρωσή τους σε ένα ενιαίο περιβάλλον:

Η πρώτη υποδεικνύει τη συγκέντρωση της πληροφορίας σε απλά φυσικά αρχεία και η χρησιμοποίηση ενός συστήματος ανάκτησης πληροφοριών τρίτου-κατασκευαστή για την διαχείριση και περαιτέρω επεξεργασία της. Αυτά τα συστήματα συνήθως είναι αποτελεσματικά για μικρού ή μεσαίου μεγέθους συλλογές και αποδεικνύονται μη ικανά να ικανοποιήσουν τις απαιτήσεις για αποθήκευση και ανάκτηση της πληροφορίας σε επίπεδο ολόκληρου του οργανισμού (enterprise-wide) όπου το μεγάλο μέγεθος και η συχνότητα ενημέρωσής της είναι τα κύρια χαρακτηριστικά.

Επιπλέον είναι εντελώς άχρηστα όταν πρόκειται να χειριστούν τα δομημένα δεδομένα σε επίτεδο οργανισμού.

Η δεύτερη προσέγγιση υποδεικνύει την εισαγωγή της αδόμητης πληροφορίας σε μορφή κειμένου σε στήλες των πινάκων της βάσης δεδομένων της επιχείρησης υπό τη μορφή συμβολοσειράς. Αυτή η προσέγγιση λύνει το πρόβλημα της συνύπαρξης δομημένης και αδόμητης πληροφορίας αλλά περιορίζει κατά πολύ τις δυνατότητες σχετικά με το ψάξιμο και την ανάκτηση της αδόμητης πληροφορίας. Επιπλέον αυτή η λύση επιβάλει η αδόμητη πληροφορία να είναι σε απλή μορφή κειμένου για να μπορεί να εισαχθεί σε στήλες πίνακα και όχι να έχει κάποιο συγκεκριμένο μορφότυπο όπως αυτό των σύγχρονων εφαρμογών π.χ. μορφή HTML σελίδων, MS Word έγγραφο κ.α.

Λόγω της ανάγκης να ικανοποιηθούν οι απαιτήσεις που επιβάλει η επίτευξη ολοκληρωμένης και αποτελεσματικής διαχείρισης της πληροφορίας, πολλές επιχειρήσεις έχουν καταφύγει στη ταυτόχρονη υιοθέτηση και των δύο αυτών προσεγγίσεων προκειμένου να πετύχουν τα πλεονεκτήματα που προσφέρει η κάθε μια από αυτές. Αυτό όμως το πράγμα απαιτεί μεγάλη προσπάθεια και προσοχή για τη διατήρηση της ακεραιότητας και συνέπειας των δεδομένων που υπάρχουν σε αυτά τα δύο ξεχωριστά συστήματα ανάκτησης των δεδομένων, ενώ εξακολουθεί να παραμένει το πρόβλημα της ολοκλήρωσής τους (integration) και λειτουργίας τους σε ένα ενιαίο περιβάλλον.

3.2 Μια τρίτη προσέγγιση: Υλοποίηση IRS με τη βοήθεια DBMS

Μια τρίτη υβριδική προσέγγιση που έχει προταθεί για την αποτελεσματική συνύπαρξη και διαχείριση δομημένης και αδόμητης πληροφορίας είναι η υλοποίηση ενός συστήματος ανάκτησης πληροφοριών χρησιμοποιώντας την μηχανή ενός σχεσιακού συστήματος διαχείρισης βάσεων δεδομένων. Η υλοποίηση δηλαδή του IRS θα βασίζεται σε σχεσιακούς πίνακες και στις δυνατότητες διαχείρισης και ανάκτησης που προσφέρει το DBMS για τα δομημένα δεδομένα. Μια τέτοια υλοποίηση προσφέρει αρκετά πλεονεκτήματα αφού τα DBMS μπορούν πλέον να διαχειριστούν αποδοτικά τα δεδομένα ενώ προσφέρουν και δυνατότητες που είναι ιδιαίτερα χρήσιμες για την απόδοση των IRS όπως η δυνατότητα δημιουργίας δεικτών (indexes) για ορισμένες στήλες πινάκων πράγμα που αυξάνει την ταχύτητα προσπέλασης στα δεδομένα αυτών των στηλών. Επιπλέον εκτός από την καλή απόδοση, τα DBMS προσφέρουν εύκολη διαχείριση (administration) και έλεγχο πάνω στα δεδομένα, ταυτόχρονη προσπέλαση σε αυτά πολλών χρηστών, ασφάλεια και ακεραιότητα και τέλος συνεπή τρόπο εσωτερικής αποθήκευσης των δεδομένων ανεξάρτητα του τύπου της μηχανής ή λειτουργικού συστήματος.

3.2.1 Το Σύστημα βιβλιογραφικών εγγραφών retriev

Στις αμέσως επόμενες παραγράφους θα κάνουμε μια σύντομη παρουσίαση του συστήματος retriev, ενός πειραματικού/ερευνητικού συστήματος που αναπτύχθηκε από το School of Computer Applications του Dublin City University^[11] και που προσφέρει υπηρεσίες ανάκτησης πληροφοριών βασισμένο στις λειτουργίες και δυνατότητες που παρέχει ένα DBMS. Πιο συγκεκριμένα, το retriev είναι ένα σύστημα που υλοποιήθηκε για την αποθήκευση και την ανάκτηση βιβλιογραφικών εγγραφών που αφορούν τεχνικά εγχειρίδια/αναφορές αλλά η διαφορά του με άλλα παρόμοια είναι ότι προσφέρει εκτεταμένες δυνατότητες αναζήτησης του χρήστη σχετικά με τον τίτλο ενός βιβλίου και τους συγγραφείς του ενώ παρέχει και δυνατότητες επανατροφοδότησης και επανακατάταξης (re-rank) των αποτελεσμάτων αν ο χρήστης επιθυμεί κάτι τέτοιο.

Το σύστημα retriev για την λειτουργία του χρησιμοποιεί τους παρακάτω σχεσιακούς πίνακες :

	t reps		stems
x1 uniq	<i>r num</i> serial		<i>Stems</i> char(20)
x2 dupl	<i>title</i> char(100)		<i>o set</i> smallint
	<i>au1</i> char(20)	x3 dupl	<i>Fld</i> char(1)
	<i>au2</i> char(20)		<i>d type</i> char(1)
	<i>origin</i> char(40)	x4 dupl	<i>r num</i> integer
	<i>number</i> char(20)		
	<i>p date</i> char(4)		
	<i>a date</i> date		
x6 uniq	stopwords		
	<i>stopword</i> Char(20)	x5 uniq	freqs
			<i>Stem</i> char(20)
			<i>Freq</i> integer
	persistent		
	<i>d type</i> Char(1)	x7 dupl	au codes
	<i>numb</i> Integer		<i>r num</i> integer
		x8 dupl	<i>au1code</i> char(6)
	t stems		<i>au2code</i> char(6)
	<i>stem</i> Char(20)		<i>d type</i> char(1)
	reldocs		
	<i>stem</i> Char(20)	x9 dupl	aunames
	<i>r num</i> Integer		<i>Surname</i> char(20)
			<i>r num</i> integer
			<i>d type</i> char(1)

Πίνακας 3.1: Σχήμα της βάσης του συστήματος retriev

Σε μερικούς από τους αυτούς πίνακες έχουν δημιουργηθεί δείκτες για ορισμένες στήλες ($x1, x2, x3, x4, x5, x6, x7, x8, x9$ όπου $\text{uniq}=\mu$ μοναδικές τιμές, $\text{dupl}=\mu$ διπλοεμφανιζόμενες τιμές) ώστε να αυξάνεται η ταχύτητα πρόσβασης του DBMS στα δεδομένα αυτών των στηλών. Παρόλο που οι συντήρηση και αναδιοργάνωση των δεικτών απαιτεί επιπλέον κόστος για το DBMS, η ταχύτητα ενός συστήματος ανάκτησης πληροφοριών είναι πρωταρχικής σημασίας και γίνεται προσπάθεια να επιτευχθεί ισορροπία ανάμεσα στη ταχύτητα που προσφέρουν και στο κόστος συντήρησή τους.

Ο κύριος πίνακας του συστήματος *retriev* είναι ο πίνακας *t_reps* ο οποίος κρατά την πληροφορία για κάθε βιβλίο-εγγραφή που εισάγεται στη βάση. Ο πίνακας αυτός περιλαμβάνει τα πεδία *r_num* που είναι ο κωδικός του βιβλίου, το πεδίο *title* μεγέθους 100 χαρακτήρων που κρατά τον τίτλο του βιβλίου, τα πεδία *au1* και *au2* μεγέθους 20 χαρακτήρων στα οποία μπορούν να αποθηκευτούν τα ονόματα δύο συγγραφέων, το πεδίο *origin* μεγέθους 40 χαρακτήρων που κρατά την προέλευση του βιβλίου, το πεδίο *number* μεγέθους 20 χαρακτήρων το οποίο είναι ο σειριακός αριθμός του βιβλίου, το πεδίο *p_date* το οποίο κρατά το έτος έκδοσης του βιβλίου και τέλος το πεδίο *a_date* το οποίο κρατά την ημερομηνία απόκτησης και εισαγωγής του βιβλίου στη βιβλιοθήκη. Οι υπηρεσίες αναζήτησης που προσφέρει το *retriev* αφορούν τον τίτλο και τους συγγραφείς των βιβλίων.

3.2.1.1 Εισαγωγή βιβλίου

Κάθε φορά που εισάγεται μια καινούργια εγγραφή στον πίνακα *t_reps* δηλαδή έχουμε εισαγωγή ενός καινούργιου βιβλίου γίνεται μια σειρά από πράξεις στα πεδία *title* και *au1* και *au2* ώστε να μπορούν μετέπειτα να προσφερθούν εκτεταμένες υπηρεσίες ανάκτησης πληροφοριών σε αυτά τα πεδία. Πιο συγκεκριμένα, από τον τίτλο κάθε βιβλίου αφαιρούνται οι τετριμμένες λέξεις και αυτό γίνεται με την χρήση του πίνακα *stopwords*. Η ύπαρξη του δείκτη *xb* για το πεδίο *stopword* του πίνακα *stopwords* επιταχύνει την όλη διαδικασία. Στις λέξεις του τίτλου που απομένουν εφαρμόζεται ο αλγόριθμος αποκοπής καταλήξεων του Porter^[12] και κάθε ρίζα που προκύπτει εισάγεται ως εγγραφή στον πίνακα *stems* μαζί με τον κωδικό *r_num* του βιβλίου από τον τίτλο του οποίου προήλθε αυτή η ρίζα. Επίσης η ρίζα εισάγεται ταυτόχρονα και στον πίνακα *t_stems*. Οταν τελειώσει η επεξεργασία κάθε λέξης του τίτλου με τον τρόπο που περιγράψαμε, ενημερώνεται ο πίνακας *freqs* ο οποίος κρατά την συχνότητα εμφάνισης κάθε ρίζας στην βάση. Αν η ρίζα είναι καινούργια στη βάση, τότε η συχνότητα της είναι 1 και οι SQL εντολές που ενημερώνουν τον πίνακα *freqs* είναι

```
INSERT INTO FREQS (STEM,FREQS)
VALUES ("stem",1)
```

ειδάλλως απλά αυξάνεται η συχνότητα εμφάνισης αυτής της ρίζας κατά 1 με τις εντολές



```
UPDATE FREQS SET FREQ=FREQ+1
WHERE STEM="stem"
```

Όσον αφορά τα ονόματα των συγγραφέων ακολουθείται διαφορετική διαδικασία. Για κάθε όνομα συγγραφέα παράγεται ένας 6-ψήφιος SOUNDEX-88 κωδικός που επιτρέπει την αναζήτηση ενός βιβλίου κατά συγγραφέα με την χρήση ονομάτων συγγραφέων που έχουν τον ίδιο ήχο άσχετα αν γράφονται διαφορετικά (sound-alike matching). Με κάθε εισαγωγή μιας βιβλιογραφικής εγγραφής γίνεται ταυτόχρονη εισαγωγή στον πίνακα *au_codes*, ο οποίος περιέχει τον κωδικό του βιβλίου και τους κωδικούς των συγγραφέων που παράγει ο SOUNDEX αλγόριθμος. Επιπλέον, κάθε φορά που δημιουργείται SOUNDEX κωδικός για κάποιον συγγραφέα γίνεται εισαγωγή στον πίνακα *au_names* το πλήρες όνομα του συγγραφέα και ο αντίστοιχος κωδικός του βιβλίου. Εάν υπάρχει μόνο ένας συγγραφέας για ένα βιβλίο, μόνο μία εισαγωγή γίνεται στον πίνακα *au_names* ενώ ο SOUNDEX κωδικός για τον δεύτερο συγγραφέα στον πίνακα *au_codes* είναι 0. Εάν υπάρχουν περισσότεροι από δύο συγγραφείς για κάποιο βιβλίο τότε η συμβολοσειρά "et al." καταχωρείται σαν όνομα δεύτερου συγγραφέα στον πίνακα *t_reps* και παράγεται ο SOUNDEX κωδικός 0 για τον δεύτερο συγγραφέα.

Τέλος, ο πίνακας *persisten* περιέχει ορισμένες πληροφορίες που βοηθούν την λειτουργία της ανάκτησης και την κατάταξη των επιστρεφόμενων αποτελεσμάτων. Τέτοιες πληροφορίες είναι ο τρέχον αριθμός βιβλίων που υπάρχουν στη βάση, ο αριθμός των ριζών, ο αριθμός εμφάνισης των ριζών σε όλη την βάση και τέλος ο αριθμός εμφάνισης της πιο συχνά εμφανιζόμενης ρίζας στη βάση. Αυτές οι πληροφορίες αποθηκεύονται στο πεδίο *numb* μαζί με κάποιο αντίστοιχο αναγνωριστικό στο πεδίο *d_type* που ερμηνεύει την τιμή του πεδίου *numb*. Θα πρέπει να πούμε ότι η ενημέρωση του πίνακα *persisten* γίνεται αμέσως μετά από κάθε εισαγωγή βιβλίου ώστε τα δεδομένα που περιέχει να είναι επίκαιρα και να ανταποκρίνονται στα πραγματικά δεδομένα που περιέχει η βάση.

Εάν κάποιες από τις παραπάνω ενέργειες δεν ολοκληρωθεί με επιτυχία τότε ολόκληρη η συναλλαγή (transaction) αποσύρεται και καμία από τις αλλαγές που είχαν γίνει δεν ισχύει. Μόνο όταν όλες οι παραπάνω πράξεις εκτελεστούν με επιτυχία, οι αλλαγές κατοχυρώνονται, πράγμα που εξασφαλίζει την ακεραιότητα των δεδομένων στη βάση.

3.2.1.2 Επιλογές αναζήτησης

Το σύστημα retriev παρέχει τις ακόλουθες δυνατότητες αναζήτησης στον χρήστη: Ακριβής αναζήτηση κατά τίτλο ή συγγραφέα, αναζήτηση κατά συγγραφέα με χρήση SOUNDEX και τέλος αναζήτηση κατά τίτλο με την χρήση φράσης.



Ακριβής αναζήτηση κατά τίτλο ή συγγραφέα

Αυτή η δυνατότητα αναζήτησης βασίζεται εξ ολοκλήρου σε ταίριασμα συμβολοσειράς (exact pattern matching) και παρέχεται από το ίδιο το DBMS χωρίς την μεσολάβηση κάποιου ενδιάμεσου σταδίου επεξεργασίας της ερώτησης του χρήστη όπως συμβαίνει στα IRS. Η αναζήτηση που πραγματοποιείται λαμβάνει υπόψη της τις διαφορές κεφαλαίων-μικρών (case-sensitive) ενώ παρέχεται από το ίδιο το DBMS η δυνατότητα χρήσης χαρακτήρων μπαλαντέρ (Wildcards). Για παράδειγμα η αναζήτηση ως προς κάποιο τίτλο θα μπορούσε να πραγματοποιηθεί με τις εντολές

```
SELECT * FROM T_REPS
WHERE TITLE MATCHES "input title"
```

Με ανάλογο τρόπο η αναζήτηση ως προς κάποιο συγγραφέα θα μπορούσε να γίνει με τις παρακάτω SQL εντολές

```
SELECT R_NUM FROM T_AUNAMES
WHERE SURNAME ="input surname"
```

Και στις δύο περιπτώσεις οι δείκτες που υπάρχουν στα αντίστοιχα πεδία των πινάκων επιταχύνουν την αναζήτηση που πραγματοποιεί το DBMS.

Αναζήτηση κατά συγγραφέα με χρήση SOUNDEX

Αυτή η δυνατότητα αναζήτησης υλοποιείται με την χρήση του αλγορίθμου SOUNDEX-88^[13,14] για sound-alike ταίριασμα ονομάτων. Ο αλγόριθμος SOUNDEX-88 παράγει ένα 6-ψήφιο κωδικό για κάθε όνομα που τον δίνεται ως παράμετρος ανεξαρτήτου μήκους. Διαφέρει από τον κλασσικούς SOUNDEX και Davidson^[15] αλγορίθμους στο ότι δεν απαιτείται ακριβές ταίριασμα μεταξύ των κωδικών ώστε να έχουμε sound-alike ταίριασμα μεταξύ των ονομάτων. Εάν δεν βρεθεί ακριβές ταίριασμα μεταξύ των SOUNDEX-88 κωδικών, ακολουθεί μια διαδοχική αναζήτηση για ολοένα και λιγότερο επιτυχημένο ταίριασμα μεταξύ των κωδικών ωστότου επιτευχθεί κάποιου είδους ταίριασμα

Στην περίπτωση που η αναζήτηση του χρήστη αφορά sound-alike ταίριασμα, η ακόλουθες SQL εντολές εκτελούνται επανευλημένα με τιμή κωδικού "123456", "12345?", "1234?6" κ.ο.κ. ωστότου επιτευχθεί κάποιου είδους ταίριασμα

```
SELECT R_NUM FROM AUCODES
WHERE AU1CODE MATCHES "code"
OR AU2CODE MATCHES "code"
```

Με αυτόν τον τρόπο εξασφαλίζεται ότι θα επιστραφούν οι εγγραφές που έχουν το πιο καλό ταίριασμα στα ονόματα των συγγραφέων ενώ η ύπαρξη δεικτών επιταχύνει την όλη αναζήτηση. Τα αποτελέσματα αυτής της αναζήτησης παρουσιάζονται κατά φθίνουσα σειρά ταιριάσματος των ονομάτων των συγγραφέων.

3.2.2 Διατύπωση

Αναζήτηση κατά τίτλο με την χρήση φράσης

Σε αυτού του είδους την αναζήτηση, ο χρήστης διατυπώνει την ερώτηση του σε φυσική γλώσσα η οποία τυγχάνει επεξεργασίας προτού γίνει η οποιαδήποτε αναζήτηση. Από την ερώτηση του χρήστη αρχικά αφαιρούνται οι τετριμμένες λέξεις και στη συνέχεια γίνεται αποκοπή καταλήξεων στις λέξεις που απομένουν. Στην συνέχεια αποδίδονται βάρη στους όρους της ερώτησης, πραγματοποιείται η αναζήτηση και τα κείμενα βαθμολογούνται σε σχέση με την ερώτηση και τα αποτελέσματα επιστρέφονται ταξινομημένα κατά φθίνουσα σειρά σχετικότητας. Στους όρους της ερώτησης αποδίδονται βάρη αντιστρόφως ανάλογα της συχνότητας του όρου στα κείμενα της συλλογής (Inverse Document Frequency) τα οποία υπολογίζονται από τον τύπο

$$W_i = \log \frac{\max freq}{freq_i}$$

όπου $freq_i$ είναι η συχνότητα εμφάνισης του όρου i στα κείμενα της συλλογής και $\max freq$ είναι ο αριθμός των κειμένων που περιέχουν τον όρο που έχει την υψηλότερη συνολική συχνότητα στη βάση. Η βαθμολογία που παίρνει ένα κείμενο υπολογίζεται ως το άθροισμα των βαρών των όρων της ερώτησης που υπάρχουν σε αυτό. Η παράμετρος $freq_i$ που απαιτείται για τον υπολογισμό των βαρών υπάρχει στο πίνακα $freqs$ ενώ η παράμετρος $\max freq$ υπάρχει στο πίνακα *persistency*.

Θα πρέπει να πούμε ότι η όλη διαδικασία απόδοσης βαρών στους όρους της ερώτησης και βαθμολογίας των κειμένων γίνεται κατά τρόπο διαφανές προς τον χρήστη. Τα αποτελέσματα που βλέπει ο χρήστης είναι μονάχα μία ταξινομημένη λίστα κειμένων κατά φθίνουσα σειρά σχετικότητας με έντονα φωτισμένους του όρους της ερώτησης στο περιεχόμενο των κειμένων.

Θα πρέπει να πούμε ότι ο χρήστης με την εμφάνιση των αποτελεσμάτων και τον χαρακτηρισμό ορισμένων κειμένων ως σχετικών με την ερώτηση, μπορεί να απαιτήσει από το σύστημα ένα δεύτερο υπολογισμό των βαρών των όρων της ερώτησης και στη συνέχεια μια επαναβαθμολόγηση των κειμένων ώστε να προκύψει μια καινούργια βελτιστοποιημένη κατάταξη των αποτελεσμάτων. Πολλές φορές όμως συμβαίνει και το ίδιο το σύστημα να μπορεί να κρίνει από μόνο του αν θα πρέπει να προβεί σε μία ανακατάταξη των αποτελεσμάτων αν ισχύουν ορισμένες συνθήκες οι οποίες όμως έχουν επαληθευτεί μόνο πειραματικά. Προκειμένου να είναι εφικτές τέτοιου είδους επεξεργασίες και δυνατότητες επαναδιαβάθμισης, τα αποτελέσματα κάθε αναζήτησης αποθηκεύονται προσωρινά στον πίνακα *reldocs* ο οποίος περιέχει τους όρους των κειμένων που επεστράφησαν στα πεδία *stem* και *r_num* αντίστοιχα. Τέλος σημειώνουμε ότι το σύστημα *retriev* παρέχει και την δυνατότητα επανατροφοδότησης και επιτρέπει στον χρήστη να επεκτείνει την ερώτηση του συμπεριλαμβάνοντας όρους από κείμενα που γνωρίζει ο χρήστης ότι είναι σχετικά με την ερώτηση.



3.2.2 Συμπεράσματα

Οπως φάνηκε από την παραπάνω σύντομη παρουσίαση, ο προσανατολισμός των συστημάτων διαχείρισης βάσεων δεδομένων είναι αρκετά διαφορετικός από αυτόν των συστημάτων ανάκτησης πληροφοριών. Τα μεν πρώτα συστήματα επικεντρώνουν την προσοχή τους στις λειτουργίες αναζήτησης δομημένων δεδομένων μέσα από την μορφή σχεσιακών πινάκων και σχέσεις πρωτευόντων-δευτερευόντων κλειδιών, τα δε IRS είναι προσανατολισμένα στη ανάκτηση των κατάλληλων πληροφοριών από πλήθος κειμένων με άγνωστο νοηματικό περιεχόμενο χωρίς να υπάρχει σαφή μέτρο που να χαρακτηρίζει τα αποτελέσματα της αναζήτησης σωστά ή λάθος. Παρόλα αυτά οι δυνατότητες χειρισμού και αναζήτησης των δεδομένων που προσφέρουν τα DBMS μπορούν να χρησιμοποιηθούν έμμεσα μέσα από σχέσεις πινάκων για την εφαρμογή τεχνικών αναζήτησης και βαθμολόγησης των αποτελεσμάτων όπως συμβαίνει στα συστήματα ανάκτησης. Παράδειγμα εφαρμογής των παραπάνω αποτελεί το σύστημα retriev που περιγράψαμε το οποίο υλοποιεί μέσα από σχέσεις πινάκων λειτουργίες του IR όπως αποκοπή καταλήξεων, απόδοση βαρών στους όρους της ερώτησης, επανατροφοδότηση κ.λ.π.

Το μέλλον δεν είναι η προσομοίωση της λειτουργίας των IRS πάνω σε συστήματα DBMS αλλά η ολοκλήρωσή τους σε ένα ενιαίο περιβάλλον λειτουργίας και διαχείρισής τους διατηρώντας το καθένα τα ιδιαίτερα χαρακτηριστικά που το κάνουν να υπερέχει σε σχέση πάντα με το είδος και την φύση της πληροφορίας που καλείται να επεξεργαστεί. Μια τέτοια προσέγγιση θα αναλύσουμε στη συνέχεια περιγράφοντας την περίπτωση της Oracle.

3.3 Περίπτωση της Oracle

Η Oracle έρχεται να γεφυρώσει το χάσμα που υπάρχει ανάμεσα στις δομημένες και αδόμητες πληροφορίες ενσωματώνοντας και τον δύο αυτούς τύπους σε στήλες πινάκων όπως γινόταν μέχρι τώρα μόνο με τα δομημένα δεδομένα. Επιπλέον, ο χειρισμός και η ανάκτηση και των δύο αυτών τύπων δεδομένων εξακολουθεί να γίνεται διαμέσου της γνωστής γλώσσας επερωτήσεων SQL (Structured Query Language) χωρίς να απαιτείται ιδιαίτερη αντιμετώπιση για το κάθε τύπο δεδομένων ξεχωριστά.

Πιο συγκεκριμένα, το Oracle ConText Option αποτελεί μία επέκταση του συστήματος διαχείρισης βάσεων δεδομένων που παρέχει η Oracle7 έκδοση 7.3 προσφέροντας βασικές αλλά και προχωρημένες υπηρεσίες ανάκτησης πληροφοριών από ορισμένες στήλες πινάκων. Μέχρι τώρα τα πεδία των πινάκων που μπορούσαν να αποθηκεύσουν αδόμητη πληροφορία σε μορφή ελεύθερου κειμένου αντιμετωπίζονταν σαν τύπος συμβολοσειράς και το γάξιμο και η αναζήτηση πληροφορίας μέσα σε αυτά περιοριζόταν σε μια απλή αναζήτηση ενός υποδείγματος (pattern matching) μέσα στο περιεχόμενα της αντίστοιχης στήλης. Στην περίπτωση του ConText Option, σε μια στήλη ενός πίνακα μπορεί να αποδοθεί μία συγκεκριμένη πολιτική (policy) που την ορίζει ως μια στήλη ειδικής μορφής στην οποία θα μπορούν

να εφαρμοστούν όλες οι δυνατότητες ανάκτησης κειμένων που παρέχει το ConText στα περιεχόμενα της. Επιπλέον, όλες οι λειτουργίες σχετικά με τη διαχείριση των ειδικών αυτών στηλών αλλά και της ανάκτησης κειμένων γίνονται διαμέσου ειδικών για αυτό το σκοπό εντολών της SQL πράγμα που επιτρέπει σε ήδη υπάρχουσες εφαρμογές να αποκτήσουν την δυνατότητα ανάκτησης κειμένων για στήλες οι οποίες μέχρι τώρα κρατούσαν αδόμητη πληροφορία σε μορφή συμβολοσειράς.

3.3.1 Χαρακτηριστικά του Oracle ConText Option 7.3

3.3.1.1 Στήλες Κειμένου

Σε κάθε στήλη ενός Oracle πίνακα που μπορεί να κρατήσει αδόμητη πληροφορία δηλαδή στήλες με τύπο δεδομένων συμβολοσειράς όπως TEXT, CHAR, VARCHAR κ.α. μπορεί να τις αποδοθεί μια συγκεκριμένη πολιτική που την καθιστά δυνατή για ανάκτηση κειμένου ενώ η ίδια η στήλη μετατρέπεται με αυτόν τον τρόπο σε στήλη κειμένου (text column) κατά την ορολογία της Oracle. Η πολιτική που αποδίδεται σε τέτοιες στήλες καθορίζει ένα σύνολο από δυνατές παραμέτρους (preferences) που υποδεικνύουν τα χαρακτηριστικά και την φύση αυτών των στηλών. Πιο συγκεκριμένα, μια πολιτική παρέχει πληροφορίες που αφορούν την δεικτοδότηση των περιεχομένων των στηλών κειμένου και ανάλογα με την είδος τους ανήκουν στις παρακάτω επτά κατηγορίες^[17]:

- Αποθήκευση δεδομένων (Data store)
- Φίλτρα (Filters)
- Συμπιεστής (Compressor)*
- *στη παρούσα έκδοση 7.3 δεν έχει υλοποιηθεί
- Λεξικός αναλυτής (Lexer)
- Μηχανή δεικτοδότησης (Engine)
- Διαχείριση λέξεων (Wordlist)
- Διαχείριση τετριμένων λέξεων (Stoplist)

Αποθήκευση δεδομένων

Οι παράμετροι που ανήκουν σε αυτή τη κατηγορία ορίζουν τον τρόπο με τον οποίο αποθηκεύονται τα δεδομένα δηλαδή το περιεχόμενο των κειμένων στις στήλες κειμένου. Υπάρχουν τρεις δυνατοί τρόποι αποθήκευσης:

- **Απευθείας (Direct)**, όπου τα περιεχόμενα των κειμένων στα οποία θέλουμε να κάνουμε αναζητήσεις αποθηκεύονται απευθείας μέσα στις στήλες κειμένου των πινάκων. Θα πρέπει να πούμε ότι δεν υπάρχει περιορισμός στο μέγεθος των κειμένων όπως θα γινόταν αν αποθηκεύονταν σε μια κοινή στήλη τύπου συμβολοσειράς.
- **Εξωτερικός (External)**, όπου τα περιεχόμενα των κειμένων βρίσκονται σε φυσικά αρχεία στο δίσκο και τα κελιά των στηλών κειμένου περιέχουν δείκτες

προς τα αντίστοιχα αρχεία. Οι δείκτες υλοποιούνται με το να αποθηκεύεται το όνομα του αντίστοιχου αρχείου στη στήλη του πίνακα.

- **Λεπτομερής (Master/Detail)**, που χρησιμοποιείται κυρίως όταν τα περιεχόμενα ενός κειμένου δεν αποθηκεύονται όλα σε μία στήλη κειμένου μιας εγγραφής του πίνακα αλλά επεκτείνονται σε περισσότερες από μία εγγραφές. Αυτός ο τύπος αποθήκευσης θα μπορούσε να χρησιμοποιηθεί στην περίπτωση που χωρίζαμε τα περιεχόμενα ενός κειμένου σε ξεχωριστές παραγράφους και κατ' επέκταση εγγραφές.

Φίλτρα

Οι παράμετροι που υπάρχουν σε αυτή τη κατηγορία καθορίζουν τον τύπο των περιεχόμενων που αποθηκεύονται σε στήλες κειμένου. Τα περιεχόμενα αυτών των στηλών μπορούν να βρίσκονται στην ακόλουθη μορφή:

- ASCII
- HTML
- MS Word for Windows 6.0, 6.1
- WordPerfect 5.0, 5.1, 6.x
- Autorecognize*

*δεν παρέχεται στη παρούσα έκδοση 7.3

Το Oracle Context Option αποθηκεύει τα περιεχόμενα των κειμένων στην αρχική τους μορφή και χρησιμοποιεί τα παραπάνω φίλτρα για να μπορεί να δημιουργήσει προσωρινές ASCII εκδόσεις των περιεχομένων τους ώστε να μπορεί να γίνει η ανακάλυψη και δεικτοδότηση των όρων τους. Θα πρέπει να πούμε ότι τα παραπάνω φίλτρα λειτουργούν μόνο όταν ο τρόπος αποθήκευσης των κειμένων είναι απευθείας (βλ. κατηγορία Αποθήκευση δεδομένων).

Λεκτικός αναλυτής

Οι παράμετροι σε αυτή την κατηγορία καθορίζουν τον τρόπο με τον οποίο διαβάζεται (parsing) το περιεχόμενο των κειμένων και αναγνωρίζονται οι λεκτικές μονάδες (tokens) που θα δεικτοδοτηθούν. Τα αγγλικά και οι περισσότερες ευρωπαϊκές γλώσσες χρησιμοποιούν τις ίδιες ρυθμίσεις για την κατηγορία Λεκτικός αναλυτής. Οι λεκτικές μονάδες σε αυτές τις γλώσσες χωρίζονται μεταξύ τους με κενά και άλλα ειδικής σημασίας σύμβολα όπως κόμματα, τελείες, ερωτηματικά κ.α.. Συνήθως, στην κατηγορία αυτή γίνονται επιπλέον ρυθμίσεις για τις περιπτώσεις γλωσσών που δεν ακολουθούν τους συνηθισμένους κανόνες γραμματικής ανάλυσης όπως είναι τα Ιαπωνικά, Κινέζικα και ορισμένες άλλες Ασιατικές γλώσσες.

Μηχανή δεικτοδότησης

Οι ρυθμίσεις σε αυτή τη κατηγορία έχουν να κάνουν με τον τρόπο με τον οποίο κάνει την δεικτοδότηση των κειμένων η μηχανή του ConText. Οι παράμετροι που υπάρχουν σε αυτή την κατηγορία αφορούν το μέγεθος της μνήμης που δεσμεύεται για

την δεικτοδότηση των κειμένων, το χώρο στο οποίο θα αποθηκευτούν οι πίνακες δεικτοδότησης κ.α

Διαχείριση λέξεων

Οι παράμετροι αυτής της κατηγορίας χρησιμοποιούνται για να καθορίσουν την λειτουργία δύο τεχνικών που λαμβάνουν χώρα κατά την δεικτοδότηση και την ανάκτηση των κειμένων :

- Αποκοπή καταλήξεων
- Λειτουργία Soundex

Με την αποκοπή καταλήξεων, οι λέξεις που περιέχονται στην ερώτηση του χρήστη επεκτείνονται ώστε κατά την αναζήτηση να επιστρέφονται όχι μόνο τα κείμενα που περιέχουν την ακριβή λέξη που δόθηκε στην ερώτηση αλλά και όλα τα κείμενα που περιέχουν λέξεις που έχουν την ίδια ρίζα με την αντίστοιχη που δόθηκε στην ερώτηση. Ετσι για παράδειγμα μία αναζήτηση που είχε σαν κριτήριο την λέξη "καταναλωτής" με την αποκοπή καταλήξεων θα επιστρέφονταν και τα κείμενα που περιείχαν τις λέξεις "καταναλωτής", "καταναλωτές", "καταναλωτικός" κ.α.

Λόγω του ότι διαφορετικές γλώσσες έχουν και διαφορετικούς γραμματικούς κανόνες, υπάρχουν και διαφορές στον τρόπο με τον οποίο υλοποιείται η αποκοπή καταλήξεων στη κάθε γλώσσα. Το ConText περιέχει stemmer που ανήκει στη Xerox Corporation's Xsoft Division και υποστηρίζει τις ακόλουθες γλώσσες: Αγγλικά, Γερμανικά, Ισπανικά, Γαλλικά, Ιταλικά και Ολλανδικά.

Θα πρέπει να τονίσουμε ότι το ConText υποστηρίζει αποκοπή καταλήξεων σε επίπεδο ερώτησης του χρήστη και όχι σε επίπεδο δεικτοδότησης των λέξεων των κειμένων. Με άλλα λόγια τα ανεστραμμένα αρχεία που κατασκευάζει το ConText περιέχουν τους όρους στη αρχική τους μορφή και όχι το θέμα του κάθε όρου, ενώ η διαδικασία αποκοπής της κατάληξης στους όρους της ερώτησης του χρήστη απαιτεί την χρήση του τελεστή \$ πριν από τον αντίστοιχο όρο. Με αυτό τον τρόπο το ConText δίνει την επιλογή στον χρήστη για το αν θέλει να χρησιμοποιήσει αποκοπή καταλήξεων ή όχι στην αναζήτησή του, αλλά επιβαρύνει πολύ το ανεστραμμένο αρχείο και την απόδοση της μηχανής αναζήτησης.

Με την χρήση του Soundex, το ConText κατά την δεικτοδότηση των περιεχομένων των κειμένων δημιουργεί μια λίστα με όλες τις λέξεις που έχουν το ίδιο ήχο με τη κάθε λέξη που υπάρχει στα κείμενα. Σε κάθε τέτοια λέξη αποδίδεται ένα μοναδικό κλειδί και αποθηκεύεται σε ξεχωριστό ευρετήριο από το κυρίως ευρετήριο. Ετσι όταν ο χρήστης ενεργοποιήσει τη επιλογή soundex σε μία ερώτησή του και αυτό γίνεται με την προσθήκη του τελεστή ! πριν από τον αντίστοιχο όρο, η ερώτηση θα επιστρέψει όλα τα κείμενα που περιέχουν όρους που έχουν την ίδια ηχώ με τον συγκεκριμένο όρο της ερώτησης άσχετα αν γράφονται διαφορετικά. Για παράδειγμα μια αναζήτηση με τον όνομα "Smith" και την χρήση του Soundex θα επιστρέψει και κείμενα που

περιέχουν τις λέξεις "Smith", "Smythe", "Smit". Η λειτουργία Soundex στην παρούσα έκδοση 7.3 του ConText παρέχεται μόνο για την Αγγλική γλώσσα.

Διαχείριση τετριμμένων λέξεων

Οι παράμετροι αυτής της κατηγορίας καθορίζουν τις τετριμμένες λέξεις οι οποίες δεν θα ληφθούν υπόψη κατά τη δεικτοδότηση των λέξεων των κειμένων. Ο αριθμός των τετριμμένων λέξεων στην παρούσα έκδοση μπορεί να είναι μέχρι 255 ενώ υπάρχει έτοιμη λίστα με τετριμμένες λέξεις της αγγλικής γλώσσας.

Βλέπουμε λοιπόν ότι για να μπορέσουμε να πραγματοποιήσουμε ανάκτηση κειμένου από μία στήλη ενός πίνακα, θα πρέπει πρώτα να της έχει αποδοθεί μια πολιτική που να την μετατρέπει σε στήλη κειμένου. Το ConText παρέχει ορισμένες σταθερές για την κάθε μία κατηγορία που επιτρέπουν την δημιουργία της επιθυμητής πολιτικής με το ανάλογο σύνολο ρυθμίσεων. Αυτές οι σταθερές είναι :

Κατηγορία Data store

- DIRECT : Τα κείμενα αποθηκεύονται στον πίνακα ένα ανά εγγραφή
- MASTER DETAIL : Το κάθε κείμενο που αποθηκεύεται στον πίνακα επεκτείνεται σε περισσότερες από μία εγγραφές
- OSFILE : Τα κείμενα αποθηκεύονται σε φυσικά αρχεία και μόνο το όνομα τους αποθηκεύεται στον πίνακα

Κατηγορία Filter

- FILTER NOP : Απλό κείμενο ASCII
- HTML FILTER : Κείμενο με HTML εντολές μορφοποίησης (tags)
- BLASTER FILTER : Κείμενο σε ειδική μορφή

Κατηγορία Lexer

- JAPANESE V-GRAM LEXER : Χρησιμοποιείται για την σάρωση (parsing) και λεκτική ανάλυση Ιαπωνικών κειμένων
- BASIC LEXER : Για τα αγγλικά και όλες τις υποστηριζόμενες γλώσσες

Κατηγορία Engine

- GENERIC ENGINE : Εξ ορισμού οι πίνακες δεικτοδότησης αποθηκεύονται στον χώρο (tablespace) του χρήστη ενώ η μνήμη που δεσμεύεται για δεικτοδότηση είναι 12MB

Κατηγορία Wordlist

- GENERIC WORDLIST : Εξ ορισμού δεν είναι ενεργοποιημένη η επιλογή soundex.



Κατηγορία Stoplist

- **GENERIC STOPLIST** : Εξ ορισμού χρησιμοποιούνται οι τετριμμένες λέξεις της Αγγλικής γλώσσας, 99 σε αριθμό.

Για την κάθε μία κατηγορία μπορούμε να κάνουμε και άλλες επιπλέον ρυθμίσεις ώστε να προσαρμόσουμε την πολιτική ακριβώς στις απαιτήσεις μας. Αυτές οι ρυθμίσεις γίνονται μέσο της εκχώρησης των κατάλληλων τιμών στις αντίστοιχες παραμέτρους της κάθε κατηγορίας. Επι στην κατηγορία FILTER με παράμετρο BLASTER FILTER μπορούμε να ορίσουμε ακριβώς το φίλτρο που θέλουμε να χρησιμοποιήσουμε π.χ. να εκχωρήσουμε στην παράμετρο FORMAT την τιμή 1 για WordPerfect 5.0, 11 για MS WORD for Windows κ.α.. Με τον ίδιο τρόπο στην κατηγορία OSFILE, η παράμετρος PATH μπορεί να περιέχει τους καταλόγους στους οποίους βρίσκονται αποθηκευμένα τα κείμενα υπό την μορφή αρχείων ή στην κατηγορία GENERIC STOPLIST να εκχωρήσουμε στην παράμετρο STOP_WORD τις τετριμμένες λέξεις που θέλουμε να χρησιμοποιήσουμε.

3.3.1.2 Βαθμολογία ανακτηθέντων κειμένων

Το Oracle ConText αφού πραγματοποιήσει μία αναζήτηση στα περιεχόμενα των στηλών κειμένου, επιστρέφει μαζί με το κάθε κείμενο που ικανοποιεί τα κριτήρια αναζήτησης και έναν αριθμό ο οποίος δείχνει το μέτρο στο οποίο το κάθε κείμενο ικανοποιεί αυτά τα κριτήρια. Αυτός ο αριθμός είναι γνωστός ως βαθμός σχετικότητας του κειμένου ως προς την ερώτηση και τα αποτελέσματα επιστρέφονται σε φθίνουσα σειρά ως προς αυτό τον αριθμό (relevancy ranking). Συνήθως, ο βαθμός σχετικότητας βασίζεται σε στατιστική ανάλυση του ποσοστού εμφάνισης του κριτηρίου αναζήτησης μέσα στο κάθε κείμενο. Στο ConText, ένα κείμενο το οποίο περιέχει τους όρους του κριτηρίου αναζήτησης 10 φορές θεωρείται πιο σχετικό από ένα άλλο το οποίο περιέχει τους όρους 5 μόνο φορές, αν και με αυτόν τον τρόπο πριμοδοτούνται τα μεγάλου μεγέθους κείμενα αφού θα περιέχουν πιο πολλές φορές του όρους της ερώτησης. Σε βασικές ερωτήσεις (basic queries), ο βαθμός σχετικότητας του κάθε κειμένου υπολογίζεται με βάση τον αριθμό των εμφανίσεων του κάθε όρου της ερώτησης στο αντίστοιχο κείμενο. Σε πιο προχωρημένες ερωτήσεις του χρήστη, ο βαθμός σχετικότητας επηρεάζεται από διάφορες συσχετίσεις μεταξύ των λέξεων και φράσεων των κειμένων. Επιπλέον, τα βάρη των όρων στην ερώτηση του χρήστη - αν χρησιμοποιείται αυτή η επιλογή διατύπωσης της ερώτησης - επηρεάζουν το βαθμό σχετικότητας του κάθε κειμένου με το να δίνουν περισσότερο ή λιγότερο έμφαση στην εμφάνιση αυτών των όρων μέσα στα περιεχόμενα των κειμένων.

Ο βαθμός σχετικότητας του κάθε κειμένου δημιουργείται αυτόματα από την μηχανή του ConText κατά την φάση της αναζήτησης. Η μηχανή αναζήτησης του ConText παράγει ένα βαθμό σχετικότητας για κάθε κελί της στήλης κειμένου που ικανοποιεί τα κριτήρια αναζήτησης. Το ανώτατο όριο της βαθμολογίας ενός κειμένου είναι 100 και κάθε κείμενο που ικανοποιεί τα κριτήρια αναζήτησης του αποδίδεται ένας βαθμός σχετικότητας μεταξύ του 1 και 100.

3.3.1.3 Επιλογές Αναζήτησης

Το Oracle ConText διαθέτει μία ποικιλία τελεστών που μπορούν να εφαρμοστούν στους όρους της ερώτησης του χρήστη ώστε να πετύχουμε τα επιθυμητά αποτελέσματα. Οι τελεστές αυτοί μπορούν να ταξινομηθούν στις παρακάτω τέσσερις κατηγορίες^[16]:

- Λογικοί τελεστές (Logical operators)
- Τελεστές γειτνίασης των όρων (Proximity operators)
- Τελεστές επέκτασης των όρων (Expansion operators) *
- Τελεστές ελέγχου των αποτελεσμάτων (Control operators)

*Λειτουργούν μόνο για την αγγλική γλώσσα

Λογικοί τελεστές

Οι λογικοί τελεστές συνδέουν με διάφορους λογικούς τρόπους τους όρους ή τις φράσεις της ερώτησης του χρήστη. Οι σειρά με την οποία εφαρμόζονται δεν έχει σημασία εκτός από τον τελεστή MINUS. Για παράδειγμα, το "A minus B" δίνει διαφορετικό αποτέλεσμα από το "B minus A". Οι λογικοί τελεστές συνδυάζουν τη βαθμολογία των τελεστών τους δίνοντας ένα τελικό βαθμό σχετικότητας μέχρι 100. Οι διαθέσιμοι λογικοί τελεστές είναι :

Τελεστής	Τρόπος αναπαράστασης	Περιγραφή
AND	& ή and	Και οι δύο τελεστέοι πρέπει να υπάρχουν. Επιστρέφει τη μικρότερη από τη βαθμολογία των τελεστών του.
OR	ή or	Πρέπει να υπάρχει τουλάχιστον ο ένας από τους δύο τελεστέους. Επιστρέφει τη μεγαλύτερη από τη βαθμολογία των τελεστών του.
ACCUMULATE	, ή accum	Παρόμοια λειτουργία με τον τελεστή OR αλλά όταν υπάρχουν και οι δύο τελεστέοι επιστρέφεται το άθροισμα της βαθμολογίας τους.
MINUS	- ή minus	Ψάχνει για κείμενα που περιέχουν τον αριστερό τελεστέο και ταυτόχρονα δεν περιέχουν τον δεξιό τελεστέο. Επιστρέφεται η βαθμολογία του αριστερού τελεστέου μείον τη βαθμολογία του δεξιού τελεστέου.

Παράδειγμα

Κείμενο	ΕΡΩΤΗΣΕΙΣ ΜΕ ΧΡΗΣΗ ΛΟΓΙΚΩΝ ΤΕΛΕΣΤΩΝ				
	Υπολογιστής & Πληροφορία	Υπολογιστής Πληροφορία	Υπολογιστής , Πληροφορία	Πληροφορία - Υπολογιστής	Υπολογιστής - Πληροφορία
Υπολογιστής Πληροφορία Υπολογιστής Πληροφορία Υπολογιστής Πληροφορία Υπολογιστής Πληροφορία	20	40	60	20	0
Υπολογιστής Πληροφορία Υπολογιστής Πληροφορία	10	20	30	10	0
Υπολογιστής Υπολογιστής Πληροφορία	10	20	30	0	10
Υπολογιστής Πληροφορία	10	10	20	0	0
Υπολογιστής	0	10	10	0	10
Πληροφορία	0	10	10	10	0

* Κάθε εμφανιστή όρου ή φράσης παίρνει 10 βαθμούς.

Θα πρέπει να πούμε ότι αν θέλουμε να χρησιμοποιήσουμε στο κριτήριο αναζήτησης λέξεις οι οποίες έχουν ειδική σημασία για το ConText όπως οι λέξεις "and", "or" κ.λ.π. ή θέλουμε να συμπεριλάβουμε στην ερώτησή μας και τετριμμένες λέξεις, τότε θα πρέπει να τις περικλείσουμε μέσα σε αγκύλες. Για παράδειγμα για να ψάξουμε για κείμενα που περιέχουν το λογότυπο "AT&T" θα πρέπει η ερώτηση να έχει την μορφή "{AT&T}" ειδάλλως το Context θα εκλάβει τον χαρακτήρα "&" ως τον λογικό τελεστή AND. Με τον ίδιο τρόπο για να ψάξουμε για κείμενα που περιέχουν την φράση "The Firm" θα πρέπει να περικλείσουμε όλη την φράση σε αγκύλες ειδάλλως η τετριμμένη αγγλική λέξη "The" θα αγνοηθεί από το ConText και η αναζήτηση θα γίνει μόνο με τον όρο "Firm".

Τελεστές γειτνίασης των όρων

Οι λέξεις ή οι φράσεις που βρίσκονται κοντά η μία στην άλλη θεωρούνται ότι σχετίζονται περισσότερο μεταξύ τους από άλλες που βρίσκονται μακριά η μία από την άλλη. Ο τελεστής γειτνίασης των όρων υπολογίζει τη βαθμολογία με βάση το πόσο κοντά βρίσκονται οι όροι μεταξύ τους και όχι με βάση την συχνότητα εμφάνισης του όρου ή της φράσης στο κείμενο. Η βαθμολογία που επιστρέφεται για

ένα κείμενο είναι η υψηλότερη βαθμολογία που παράγεται για τους όρους της ερώτησης που βρίσκονται κοντά ο ένας στον άλλον σε αυτό το κείμενο. Η βαθμολογία ενός κειμένου είναι 100 όταν οι δύο τελεστέοι είναι γειτονικοί μέσα στο κείμενο.

Τελεστής	Τρόπος αναπαράστασης	Περιγραφή
NEAR	; ή near	Επιστρέφει υψηλότερη βαθμολογία όταν οι τελεστέοι βρίσκονται κοντά μεταξύ τους στο κείμενο

Παράδειγμα

Εάν το κριτήριο αναζήτησης ήταν "ice ; cream" τότε η φράση "I love ice cream" παίρνει μεγαλύτερη βαθμολογία από τη φράση "ice is colder than cream".

Τελεστές επέκτασης των όρων

Οι τελεστές επέκτασης των όρων είναι μοναδιαίοι τελεστές και παίρνουν ως παράμετρο έναν όρο και επιστρέφουν έναν αριθμό από όρους που παράγονται με βάση αυτόν. Υπάρχουν τρεις τύποι τελεστών επέκτασης :

Τελεστής	Τρόπος αναπαράστασης	Περιγραφή
STEM	\$	Επιστρέφει ένα σύνολο από λέξεις που έχουν την ίδια γραμματική ρίζα με την λέξη που του δίνεται ως παράμετρος.
FUZZY	?	Επιστρέφει ένα σύνολο από λέξεις που έχουν την ίδια προφορά με την λέξη που του δίνεται ως παράμετρος.
SOUNDEX	!	Επιστρέφει ένα σύνολο από λέξεις που έχουν την ίδια ηχώ με την λέξη που του δίνεται ως παράμετρος

Παράδειγμα

Χρήση τελεστή	Παραγόμενα
\$scream	scream screaming screamed
\$cat	cat cats
\$sing	sang sung sing
\$commit	commit committed
?cat	cat cats calc case
?apply	apply apple applied April
?read	lead real
!Smith	Smythe, Smit

Θα πρέπει να πούμε ότι οι τελεστές επέκτασης μπορούν να εφαρμοστούν σε ένα πλήθος από λέξεις μονομιάς αρκεί αυτές να περικλείονται σε παρενθέσεις () ή αγκύλες {} . Για παράδειγμα η ερώτηση

"?(dog, cat, mouse)" είναι ισοδύναμη με την "?dog, ?cat, ?mouse" ή η ερώτηση
"?!(dog, !(cat, mouse))" είναι ισοδύναμη με την "?dog, (!?cat & !?mouse) "

Τέλος, αναφέρουμε ότι οι όροι μιας ερώτησης μπορούν να επεκταθούν και με την χρήση χαρακτήρων μπαλαντέρ. Στο ConText είναι διαθέσιμοι οι ειδικοί χαρακτήρες % και _ . Ο χαρακτήρας % μπορεί να αντιπροσωπεύσει οποιαδήποτε ακολουθία χαρακτήρων ενώ ο χαρακτήρας _ αντιπροσωπεύει οποιοδήποτε χαρακτήρα.

Για παράδειγμα, με την χρήση του χαρακτήρα % η ερώτηση "εκπαίδ%" θα επιστρέψει όλα τα κείμενα που περιέχουν τους όρους "εκπαίδευση", "εκπαίδευση", "εκπαιδευτικός" κ.α.

Τελεστές ελέγχου των αποτελεσμάτων

Οι τελεστές ελέγχου είναι δυαδικοί τελεστές και καθορίζουν τα αποτελέσματα της ερώτησης ανάλογα με τη παράμετρο που εφαρμόζεται στον αριστερό τελεστέο. Ο δεξιός τελεστέος είναι η ίδια η παράμετρος. Οι τελεστές ελέγχου των αποτελεσμάτων είναι :

Τελεστής	Τρόπος αναπαράστασης	Περιγραφή
WEIGHT	*	Εκχωρεί βάρος στον αριστερό τελεστέο. Ο δεξιός τελεστέος είναι το ίδιο το βάρος και μπορεί να πάρει τιμές από 0.1 μέχρι 10
THRESHOLD	>	Δεν επιστρέφονται κείμενα με βαθμολογία μικρότερη από αυτή που υποδεικνύει η δεξιά παράμετρος.
MAX	:	Καθορίζει τον αριθμό των κειμένων που επιστρέφονται σε κάθε ερώτηση. Για παράδειγμα :20 σημαίνει ότι μόνο τα πρώτα 20 θα επιστραφούν. Οι τιμές που μπορεί να πάρει είναι από 1 μέχρι $2^{32}-1$

Παράδειγμα

Το κριτήριο αναζήτησης "cat*3 | dog*1" ψάχνει για κείμενα που περιέχουν τις λέξεις "cat" ή "dog" αλλά μεγαλύτερη βαρύτητα δίνεται στην εμφάνιση της λέξης "cat".

Το κριτήριο αναζήτησης "relational databases > 75" υποδεικνύει ότι πρέπει να επιστραφούν μόνο τα κείμενα που έχουν βαθμολογία μεγαλύτερη του 75.

Το κριτήριο αναζήτησης "dance :20" υποδεικνύει ότι πρέπει να επιστραφούν μόνο τα 20 πρώτα κείμενα που περιέχουν την λέξη "dance".

3.3.1.4 Γλωσσολογικές υπηρεσίες

* Διατίθενται μόνο για την αγγλική γλώσσα

Το ConText Option μαζί με τις υπηρεσίες ανάκτησης κειμένου προσφέρει και εκτεταμένες δυνατότητες επεξεργασίας φυσικής γλώσσας που δυναμώνουν την ανάκτηση πληροφοριών ακόμα περισσότερο^[18]. Το ConText εφαρμόζει κανόνες και πρακτικές από ένα μεγάλο σύνολο γλωσσολογικών θεωριών προκειμένου να καταλάβει τον τρόπο με τον οποίο διερμηνεύει ο ανθρώπινος νους το γραπτό κείμενο.

Τα συστήματα τα οποία βασίζονται απλά σε αναγνώριση λέξεων και στην επανάληψη αυτών, συχνά χάνουν το νόημα του κειμένου και παράγουν ένα μεγάλο αριθμό από μη σχετικά αποτελέσματα. Το ConText "σπάζει" το κείμενο στα γραμματολογικά συστατικά του και προσπαθεί να βρει κατά πόσο το καθένα συμμετέχει στο τελικό νόημα του κειμένου. Στη συνέχεια χρησιμοποιεί αυτή την γνώση για την παραγωγή κεντρικών θεμάτων-ιδεών (themes) και νοηματικών περιλήψεων (gists) για το κάθε κείμενο. Η πραγματοποίηση των παραπάνω λειτουργιών γίνεται με την βοήθεια και την συνεργασία πέντε βασικών γλωσσολογικών μονάδων:

- Λεξικό (Lexicon)
- Μηχανή Λεκτικής Ανάλυσης (Parsing Engine)
- Θεματικός Αναλυτής (Theme Analyzer)
- Μηχανή Παραγωγής Εννοιών (Conceptualization Engine)
- Μηχανή Κατηγοριοποίησης (Classification Engine)

Λεξικό

Το λεξικό που διαθέτει το ConText αποτελεί μία στατική βάση γνώσης (Knowledge base) και παρέχει πληροφορίες για τις λέξεις και φράσεις που συναντά ο λεκτικός αναλυτής. Πιο συγκεκριμένα, το λεξικό περιέχει περισσότερες από 1,000,000 αγγλικές λέξεις και φράσεις μαζί με πλήθος γλωσσολογικών πληροφοριών για κάθε μία από αυτές (σύνταξη, γραμματική, εννοιολογική βαρύτητα της λέξης κ.λ.π.) καθώς και ένα πλήθος κανόνων που χρησιμοποιούνται για την παραγωγή θεματικών ενοτήτων. Επίσης, το λεξικό είναι σχεδιασμένο έτσι ώστε να αναγνωρίζει την ορολογία που χρησιμοποιείται σε περισσότερες από 1,000 επιχειρήσεις και επιστημονικά πεδία δημιουργώντας σχήματα κατηγοριοποίησης και ομαδοποίησης εκατοντάδων εννοιών ορίζοντας έτσι την σημασιολογική όψη που έχει το ConText για τον εξωτερικό κόσμο.

Μηχανή Λεκτικής Ανάλυσης

Η μηχανή λεκτικής ανάλυσης προσομοιώνει την πολύπλοκη ανθρώπινη διεργασία που λαμβάνει μέρος κατά το διάβασμα ενός κειμένου. Ο λεκτικός αναλυτής μπορεί και αναγνωρίζει παραγράφους, προτάσεις, φράσεις και λεκτικές μονάδες (λέξεις) του κειμένου και χρησιμοποιεί τις πληροφορίες του λεξικού για την απόδοση γραμματικών και σημασιολογικών ετικετών σε αυτά ώστε να αρχίσει η γλωσσολογική ανάλυση.

Οταν αναγνωριστεί η γραμματολογική λειτουργία κάθε λέξης μέσα σε μία πρόταση χρησιμοποιώντας τις πληροφορίες του λεξικού και αυτές που απορρέουν από τη θέση της μέσα στην πρόταση και τη σχέση της με τις άλλες γειτονικές λέξεις, η μηχανή λεκτικής ανάλυσης προσπαθεί να μαντέψει την εννοιολογική λειτουργία της λέξης μέσα στην πρόταση και καθώς η ανάλυση επεκτείνεται σε μεγαλύτερα κομμάτια του κειμένου (προτάσεις, παράγραφοι και τέλος ολόκληρο το κείμενο) συνεχώς αυξάνει τις πληροφορίες που είναι διαθέσιμες για αυτή τη λέξη μέσα στη βάση γνώσης.

Θεματικός Αναλυτής

Ο θεματικός αναλυτής εξετάζει το κείμενο σε επίπεδο προτάσεων και προσπαθεί να μαντέψει την συνεισφορά και λειτουργία της κάθε λέξης στο πληροφοριακό περιεχόμενο μιας πρότασης. Με αυτό τον τρόπο προσπαθεί να βρει τα συστατικές μονάδες που συνεισφέρουν και προσδιορίζουν το θέμα των προτάσεων.

Μηχανή Παραγωγής Εννοιών

Η μηχανή παραγωγής εννοιών χρησιμοποιεί τις πληροφορίες που παρέχει ο θεματικός αναλυτής για την εξαγωγή των θεμάτων που υπάρχουν στο κείμενο και του υπολογισμού του βάρους καθενός από αυτά σε σχέση με τα υπόλοιπα του κειμένου.

Μηχανή Κατηγοριοποίησης

Η μηχανή κατηγοριοποίησης χρησιμοποιεί τις σημασιολογικές πληροφορίες που υπάρχουν στο λεξικό για την κατηγοριοποίηση των θεμάτων του κειμένου σε κάποια εμπορική ή επιστημονική ομάδα εννοιών.

Με την βοήθεια των παραπάνω πέντε γλωσσολογικών μονάδων, το ConText μπορεί και παράγει δύο τύπους αποτελεσμάτων :

- Κεντρικά θέματα-ιδέες κειμένου
- Νοηματική περίληψη κειμένου

Κεντρικά θέματα-ιδέες κειμένου

Αυτός ο τύπος αποτελεσμάτων παρουσιάζει ένα προφίλ των κύριων θεμάτων που υπάρχουν και αναπτύσσονται στο κείμενο. Προσφέρουν μια γρήγορη ματιά σχετικά με το τι "λέει" το κείμενο. Μέχρι και 16 κεντρικά θέματα-ιδέες μπορούν να παραχθούν για ένα κείμενο στο καθένα από τα οποία αποδίδεται ένα σχετικό βάρος σε σχέση με τα υπόλοιπα ώστε να μπορούμε να διακρίνουμε την συμβολή του καθενός στο νοηματικό περιεχόμενο του κειμένου. Επίσης τα θέματα που παράγονται από ένα κείμενο μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση του κάτω από μία έννοια. Για παράδειγμα, εάν ένα κείμενο αναφέρεται διεξοδικά στο MS-DOS και στο UNIX, είναι πιθανό το ConText να επιστρέψει το MS-DOS και το UNIX ως κεντρικά θέματα-ιδέες του κειμένου. Ωστόσο, είναι πιθανό το ConText χρησιμοποιώντας την βάση γνώσης να επιστρέψει και το Operating systems σαν κεντρικό θέμα του κειμένου και στη συνέχεια να το κατηγοριοποιήσει κάτω από αυτή την έννοια.

Νοηματική περίληψη κειμένου

Αυτού του είδους το αποτέλεσμα παρέχει μια νοηματική περίληψη του κειμένου και αποτελείται από επιλεγμένες παραγράφους που θεωρούνται ότι αναφέρονται στα κεντρικά θέματα του κειμένου. Για το σχηματισμό την περίληψης χρησιμοποιούνται ολόκληρες παράγραφοι και όχι ξεχωριστές προτάσεις διότι θεωρούνται ότι παρέχουν

νοηματική συνοχή και καλύτερο πλαίσιο κατανόησης του κειμένου σε σύγκριση με τις επιλεγμένες προτάσεις.

Προκειμένου να δημιουργηθεί η περίληψη για ένα κείμενο, το ConText συγκρίνει τα θέματα των παραγράφων με τα κεντρικά θέματα-ιδέες που υπάρχουν στο κείμενο και επιλέγει αυτές τις παραγράφους που θεωρούνται ότι ταιριάζουν καλύτερα με τα κεντρικά θέματα-ιδέες του κειμένου.

Σύστημα Ανάκτησης Ελληνικών Κειμένων

Επίλογος

Το Σύστημα Ανάκτησης Ελληνικών Κειμένων είναι ένα πρόγραμμα που προσπαθεί να λαμβάνει στοιχεία από παραγράφους και να βρει τα κεντρικά θέματα των ίδιων από την παραγράφη που περιλαμβάνει τα θέματα. Η προσέγγιση που έχει ο προγράμμας είναι να βρει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση. Το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση δεν είναι το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση. Το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση δεν είναι το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση.

Το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση προσπαθεί να βρει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση. Το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση προσπαθεί να βρει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση. Το πρόγραμμα που περιλαμβάνει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση προσπαθεί να βρει τα κεντρικά θέματα των παραγράφων που περιλαμβάνουν τα θέματα που έχουν οριστεί στην προστίτηση.

4.2 Δημιουργία της Βάσης

ΚΕΦΑΛΑΙΟ 4

Σύστημα Ανάκτησης Ελληνικών Κειμένων

4.1 Γενικά

Στο κεφάλαιο αυτό θα παρουσιάσουμε την αρχιτεκτονική και λειτουργία ενός συστήματος ανάκτησης ελληνικών κειμένων που αναπτύχθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας. Αν και όπως φαίνεται από τον τίτλο του κεφαλαίου γίνεται λόγος για ένα σύστημα ανάκτηση ελληνικών κειμένων, το σύστημα που αναπτύχθηκε είναι διγλωσσικό και μπορεί και υποστηρίζει αναζητήσεις και στην αγγλική γλώσσα. Μεγαλύτερη βαρύτητα ωστόσο έχει δοθεί στον χειρισμό και επεξεργασία των ελληνικών αναζητήσεων. Το σύστημα που αναπτύχθηκε προσπαθεί να ενσωματώσει όλα εκείνα τα χαρακτηριστικά των IRS όπως για παράδειγμα αφαίρεση τετριμμένων λέξεων και αποκοπή καταλήξεων της ελληνικής, κατάταξη των αποτελεσμάτων, χρήση τελεστών στην ερώτηση του χρήστη κ.α. πάντα υπό τους περιορισμούς που έθετε η πλατφόρμα λειτουργίας και το εργαλείο ανάπτυξής του.

Θα πρέπει να πούμε ότι το σύστημα που αναπτύχθηκε αποτελεί εφαρμογή WWW (Web application) πράγμα που σημαίνει ότι μπορεί να υποστηρίξει ταυτόχρονα πλήθος χρηστών από οποιοδήποτε μέρος του κόσμου αρκεί να είναι γνωστή η διεύθυνση που δέχεται τις αιτήσεις το σύστημα ανάκτησης. Στην ουσία αποτελεί μία μηχανή αναζήτησης παρόμοιες με αυτές που υπάρχουν στο διαδίκτυο αλλά είναι προσανατολισμένη σε αποθέματα (resources) πληροφοριών που υπάρχουν μέσα στον ελληνικό χώρο (domain). Η ανάπτυξη του συστήματος ανάκτησης έγινε με το RDBMS Oracle 7.3.1.3 και για την μηχανή αναζήτησης χρησιμοποιήθηκε το Oracle ConText Option. Τέλος, η συλλογή των Web σελίδων και η δημιουργία της βάσης έγινε με την χρήση ενός αυτόνομου προγράμματος γραμμένο σε Visual Basic 5.0.

4.2 Δημιουργία της Βάσης

Οπως είπαμε και σε προηγούμενες παραγράφους, το υλικό πάνω στο οποίο πραγματοποιούν αναζήτησεις οι δύαφορες μηχανές αναζήτησης του διαδικτύου είναι οι σελίδες που υπάρχουν σε αυτό. Οι Web σελίδες δηλαδή είναι για τις μηχανές αναζήτησης ότι τα κείμενα της συλλογής για τα παραδοσιακά IRS.

Για την ανακάλυψη των αποθεμάτων και δημιουργία της βάσης (δηλαδή της συλλογής) του συστήματος ανάκτησης αναπτύχθηκε και χρησιμοποιήθηκε ένας απλός Web Crawler ο οποίος ήταν αυτόνομος από το όλο σύστημα ανάκτησης αλλά πάντα προσαρμοσμένος στις απαιτήσεις του. Στο διαδίκτυο υπάρχουν έτοιμα προγράμματα που ανήκουν στην κατηγορία των offline browsers τα οποία δοθέντος της διεύθυνσης μιας περιοχής μπορούν και "κατεβάζουν" τις σελίδες αυτής της περιοχής τοπικά στον υπολογιστή και επιτρέπουν με αυτόν τον τρόπο την περιήγηση του χρήστη στις σελίδες αυτές off-line πράγμα που σημαίνει εξοικονόμηση χρόνου και ταχύτητας. Αυτά τα προγράμματα θα μπορούσαν να χρησιμοποιηθούν -αντί της δημιουργίας κάποιου από την αρχή - για το κατέβασμα των σελίδων και την δεικτοδότησή τους από τη μηχανή αναζήτησης αλλά τελικά δεν προτιμήθηκε αυτή η λύση για τους παρακάτω λόγους:

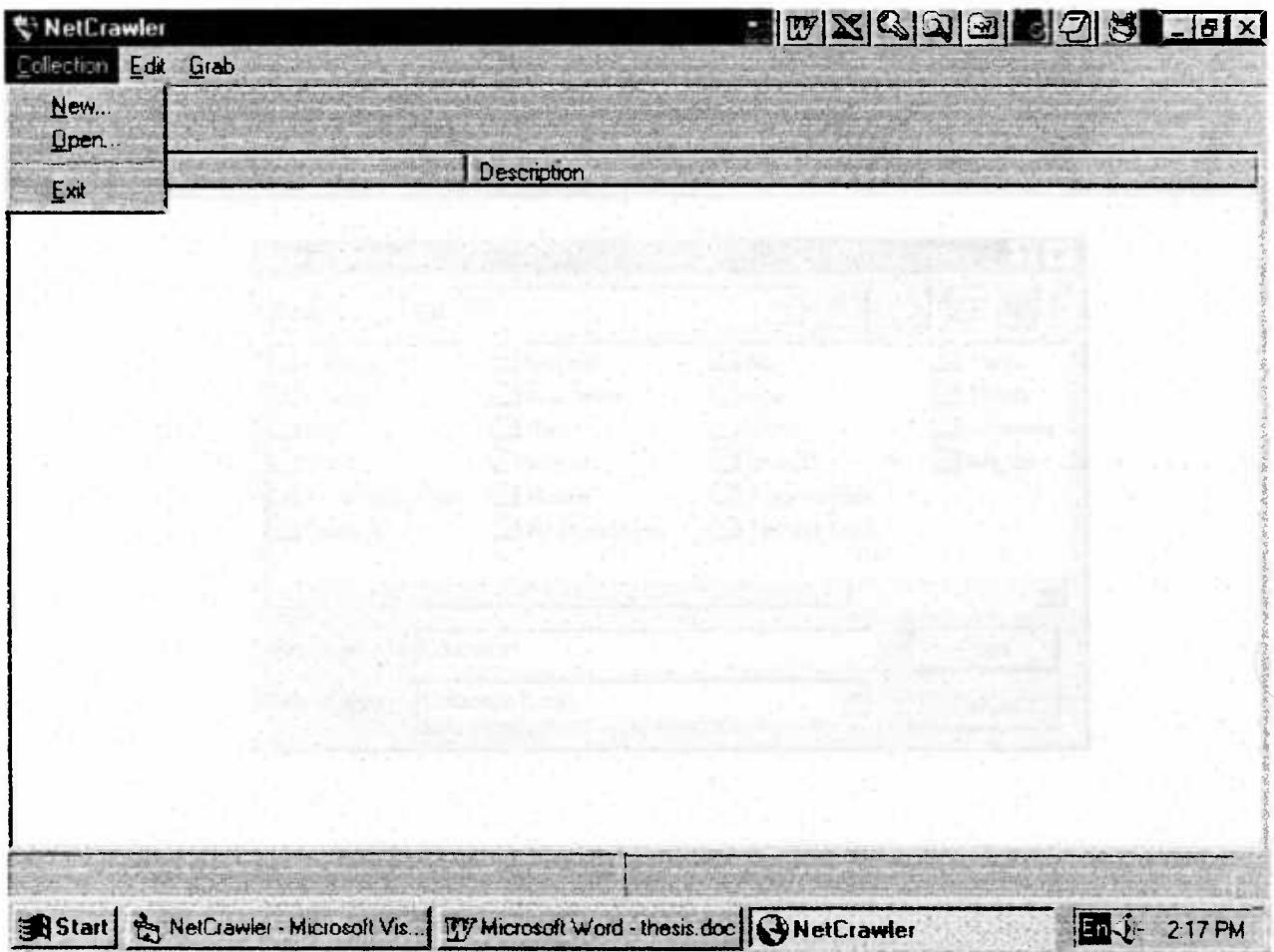
- Τα πιο πολλά από αυτά τα προγράμματα ήταν δοκιμαστικά ελεύθερα διαθέσιμα (shareware) και είχαν περιορισμό και ως προς της διάρκεια χρήστης τους αλλά και ως προς τον αριθμό των σελίδων που μπορούσαν να κατεβάσουν. Τα πιο πολλά περιόριζαν τον αριθμό των σελίδων που μπορούσαν να κατεβάσουν από μία περιοχή σε μερικές δεκάδες ενώ υπήρχαν περιοχές με εκατοντάδες σελίδες.
- Από τι στιγμή που οι σελίδες αποθηκεύονταν τοπικά στον υπολογιστή δεν υπήρχε τρόπος να μάθουμε την πραγματική διεύθυνση - URL's των σελίδων, πράγμα που ήταν απαραίτητο για την υλοποίηση του συστήματος ανάκτησης.
- Τέλος, η χρησιμοποίηση ενός τέτοιου προγράμματος θα είχε ως αποτέλεσμα την αυστηρή εξάρτηση του όλου συστήματος ανάκτησης από αυτό. Η αλλαγή του τρόπου λειτουργίας αυτού του προγράμματος λόγω καινούργιας έκδοσης του ή το τέλος της διάρκειας χρήστης του που έθετε το ίδιο το πρόγραμμα θα είχε ως αποτέλεσμα να μην μπορούσε να χρησιμοποιηθεί για το χτίσιμο ή ανανέωση της βάσης του συστήματος ανάκτησης. Επιπλέον, ακόμα και αν γινόταν η χρήση ενός τέτοιου προγράμματος έπρεπε παρόλα αυτά να υλοποιηθεί κάποιου είδους πρόγραμμα που να επεξεργαζόταν τις HTML σελίδες και να τις φόρτωνε στη βάση.

4.2.1 Αρχιτεκτονική και Λειτουργία του Web Crawler

Ο Web crawler που αναπτύχθηκε έχει τα βασικά χαρακτηριστικά και ακολουθεί παρόμοιο τρόπο λειτουργίας με αυτούς που χρησιμοποιούν οι μεγάλες μηχανές αναζήτησης του Internet. Για την λειτουργία του απαιτείται ο ορισμός μιας αρχικής διεύθυνσης από την οποία θα αρχίσει το κατέβασμα των σελίδων και ακολουθώντας

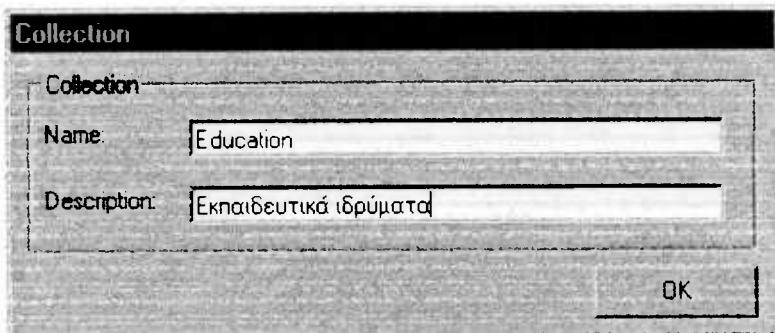
τους συνδέσμους που υπάρχουν σε αυτές τις σελίδες συνεχίζει ως ότου κατεβάσει όλες τις σελίδες από μία περιοχή.

Πιο συγκεκριμένα, όπως φαίνεται στην εικόνα 4.1, ο χρήστης έχει την δυνατότητα να δημιουργήσει από την αρχή ή να ανοίξει μια ήδη υπάρχουσα συλλογή.

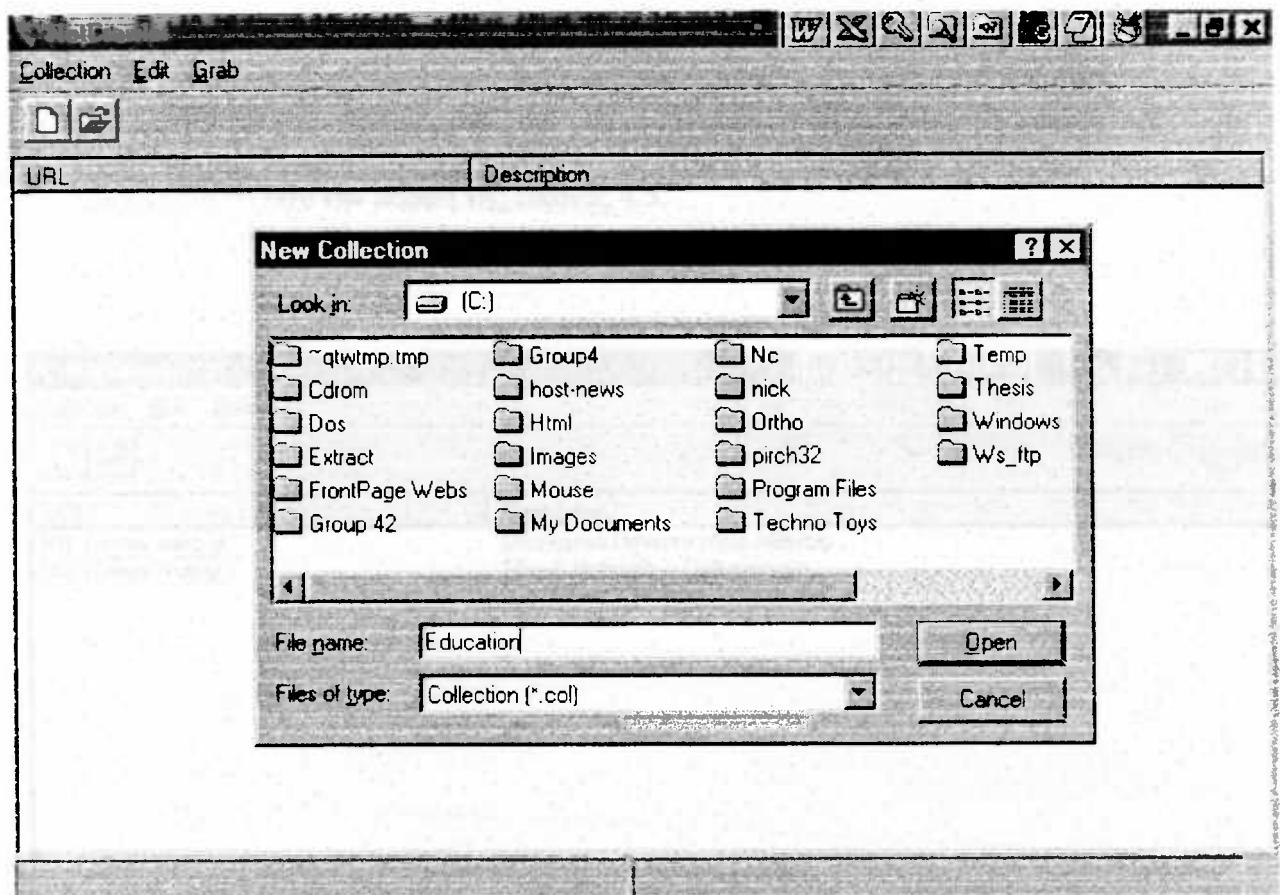


Εικόνα 4.1: Δημιουργία συλλογής

Για την δημιουργία μιας συλλογής είναι απαραίτητο ο χρήστης να ορίσει ένα όνομα αρχείου στο δίσκο στο οποίο θα αποθηκευτεί το όνομα και η περιγραφή της συλλογής (βλ. εικόνα 4.2, 4.3). Αυτό το αρχείο παίρνει την εξ ορισμού επέκταση *.col* από το *collection*. Με την δημιουργία αυτού του αρχείου δημιουργείται αυτόματα στον ίδιο κατάλογο με αυτό, ο κατάλογος <όνομα αρχείου συλλογής>.DB στον οποίο θα αποθηκευτούν οι HTML σελίδες.

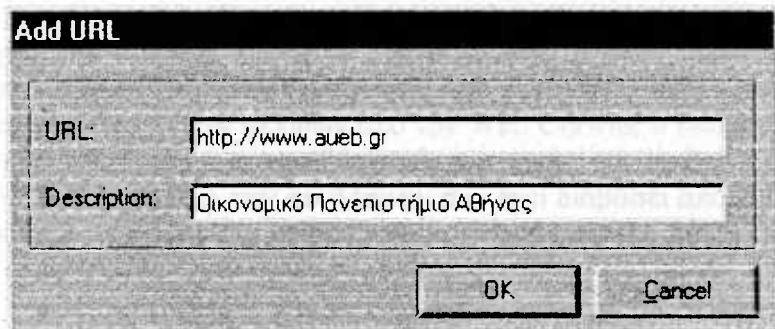


Εικόνα 4.2: Όνομα και περιγραφή συλλογής



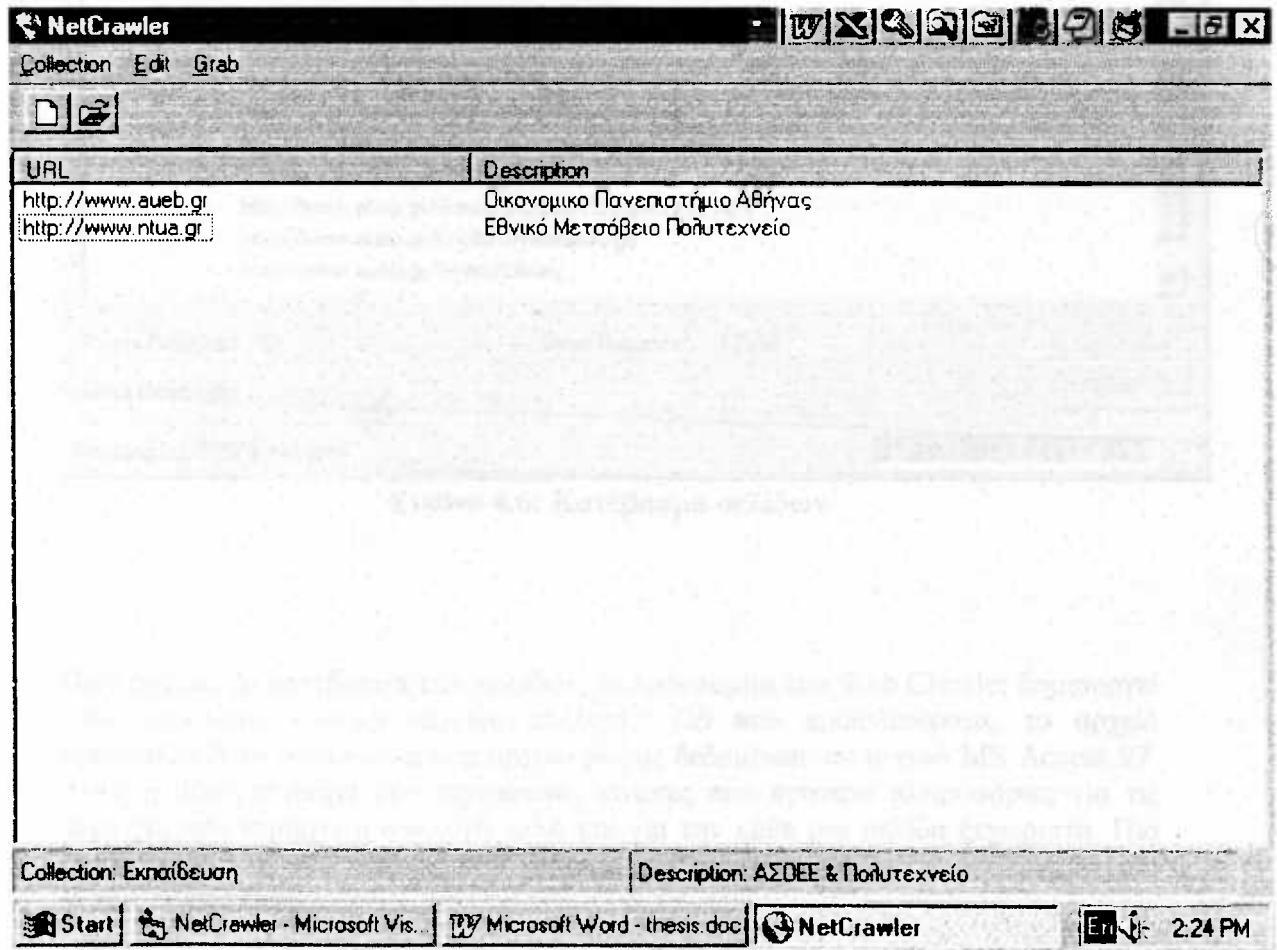
Εικόνα 4.3: Επιλογή αρχείου συλλογής

Στην συνέχεια ο χρήστης μπορεί να προσθέσει (ή/και να αφαιρέσει) μέσω της επιλογής Edit/Add URL (ή/και Edit/Delete URL) τα URLs των περιοχών από τα οποία ο Web Crawler θα κατεβάσει τις σελίδες, μαζί με μια σύντομη περιγραφή του κάθε URL (βλ. εικόνα 4.4).



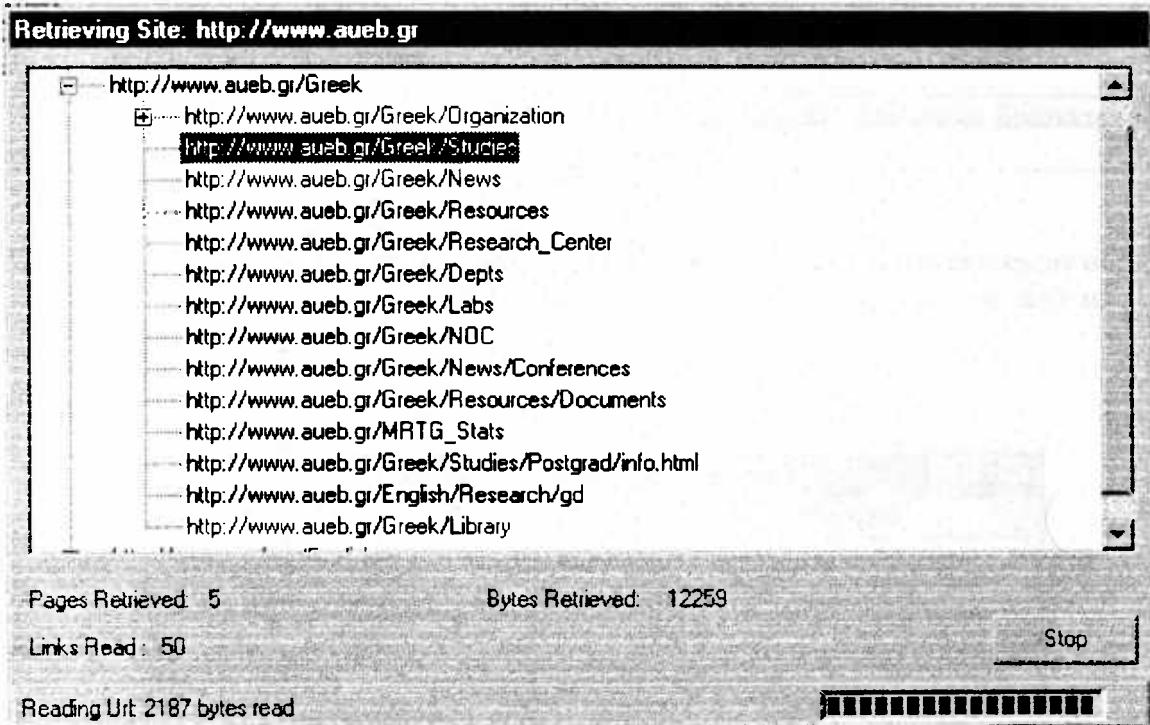
Εικόνα 4.4: Προσθήτη νέας περιοχής

Για παράδειγμα, αν δημιουργούσαμε μια συλλογή που θα περιείχε τις σελίδες του Οικονομικού Πανεπιστήμιου Αθηνών και του Εθνικού Μετσόβειου Πολυτεχνείου το πρόγραμμα θα είχε την μορφή της εικόνας 4.5.



Εικόνα 4.5: Παράδειγμα συλλογής

Από την στιγμή που ορίσαμε τα URLs της συλλογής, ο Web crawler μπορεί να αρχίσει το κατέβασμα των σελίδων επιλέγοντας από το μενού Grab/Start. Στην εικόνα 4.6 φαίνεται το κατέβασμα των σελίδων από τον Web Crawler ο οποίος πληροφορεί τον αριθμό των σελίδων που έχει κατεβάσει μέχρι στιγμής, τον αριθμό των bytes που έχει κατεβάσει και των αριθμό των συνδέσμων που έχει διαβάσει από τις μέχρι τώρα σελίδες. Θα πρέπει να πούμε ότι οι σύνδεσμοι σε κάθε σελίδα και η σειρά με την οποία γίνεται το κατέβασμα των σελίδων απεικονίζεται υπό την μορφή δένδρου όπου ο φωτισμένος κόμβος του δέντρου υποδεικνύει κάθε φορά την τρέχουσα σελίδα.



Εικόνα 4.6: Κατέβασμα σελίδων

Πριν αρχίσει το κατέβασμα των σελίδων, το πρόγραμμα του Web Crawler δημιουργεί στον κατάλογο <όνομα αρχείου συλλογής>.DB που προσαναφέραμε, το αρχείο contents.mdb το οποίο είναι ένα αρχείο βάσης δεδομένων σε μορφή MS Access 97. Αυτή η βάση περιέχει δύο σχεσιακούς πίνακες που κρατάνε πληροφορίες για τις περιοχές που περιέχει η συλλογή αλλά και για την κάθε μία σελίδα ξεχωριστά. Πιο συγκεκριμένα οι δύο πίνακες που χρησιμοποιεί και ενημερώνει το πρόγραμμα του Web Crawler είναι :

Sites		
<i>site_URL</i>	HyperLink	Το URL της περιοχής από την οποία θα κατεβαστούν οι HTML σελίδες
<i>site Description</i>	Text(150)	Σύντομη περιγραφή της περιοχής

HTMLpages		
<i>page ID</i>	AutoNumber	Ενα μοναδικό αναγνωριστικό για κάθε σελίδα
<i>page Site</i>	Hyperlink	Η περιοχή από την οποία προήλθε η σελίδα
<i>page URL</i>	Hyperlink	Το πλήρες URL της σελίδας
<i>page Title</i>	Text(100)	Ο τίτλος της σελίδας
<i>page Description</i>	Text(200)	Περιγραφή της σελίδας
<i>page Size</i>	Number	Το μέγεθος της σελίδας σε bytes
<i>page_LastModified</i>	Date/Time	Η ημερομηνία κατεβάσματος της σελίδας από τον Web Crawler
<i>page_Text</i>	Text(100)	Το πλήρες όνομα αρχείου στο οποίο βρίσκεται τοπικά η σελίδα

Στις εικόνες 4.7 και 4.8 μπορούμε να δούμε μερικά από τα δεδομένα που περιέχουν οι δύο παραπάνω πίνακες μετά τη λειτουργία και το κατέβασμα των σελίδων από το Web Crawler.

Sites : Table			
		Site_URL	Site_Description
▶	*	http://www.aueb.gr	Πρικονομικό Πανεπιστήμιο Αθηνών
	*	http://www.ntua.gr	Εθνικό Μετσόβειο Πολυτεχνείο
	*		
Record:	1	< > *	1 > * of 2

Εικόνα 4.7: Δεδομένα του πίνακα Sites

HTMLpages : Table								
	Page_ID	Page_Site	Page_URL	Page_Title	Page_Description	Page_Size	Page_LastModified	Page_Text
▶	2	http://www.aueb.gr	http://www.aueb.gr/Greek	Athens University: Αγγλική Έκδοση		3888	12/19/97	C:\NetCr2.html
	3	http://www.aueb.gr	http://www.aueb.gr/English	Athens University: Greek Version	Όι	3726	12/19/97	C:\NetCr3.html
	4	http://www.aueb.gr	http://www.aueb.gr/Greek/Organization	Organization Oikos Πανεπιστημιακή Οργά		2313	12/19/97	C:\NetCr4.html
	5	http://www.aueb.gr	http://www.aueb.gr/Greek/Studies	Spoedes sto Oik. Σπουδές Μαθήματ		2187	12/19/97	C:\NetCr5.html
	6	http://www.aueb.gr	http://www.aueb.gr/Greek/News	Nea Dikonomik Nέα Ανακοινώσεις		2200	12/19/97	C:\NetCr6.html
	7	http://www.aueb.gr	http://www.aueb.gr/Greek/Resources	Phges sto Oikon Πηγές Οδηγοί και Ι		2403	12/19/97	C:\NetCr7.html
	8	http://www.aueb.gr	http://www.aueb.gr/Greek/Research	Kentro Ereyneas Γενικά Το KENTRO EF		6032	12/19/97	C:\NetCr8.html
	9	http://www.aueb.gr	http://www.aueb.gr/Greek/Depts	Tmhmata Oikouk Τμήματα Τμῆμα Πλi		2239	12/19/97	C:\NetCr9.html
	10	http://www.aueb.gr	http://www.aueb.gr/Greek/Labs	Ergastήρια στ. Εργαστήρια Στο Οικοi		2786	12/19/97	C:\NetCr10.html
	11	http://www.aueb.gr	http://www.aueb.gr/Greek/NDIC	Kέντρο Διαχειρ Κέντρο Διαχείρισης Δ		7739	12/19/97	C:\NetCr11.html
	12	http://www.aueb.gr	http://www.aueb.gr/Greek/News/Conte	Conferences Συνέδρια Hercma 9E		468	12/19/97	C:\NetCr12.html
	13	http://www.aueb.gr	http://www.aueb.gr/Greek/Resources/	Documents Οδηγοί Τα παρακάτω		1312	12/19/97	C:\NetCr13.html
	14	http://www.aueb.gr	http://www.aueb.gr/MRTG/Stats	Router Overview Router Overview: sisyfc		3119	12/19/97	C:\NetCr14.html
	15	http://www.aueb.gr	http://www.aueb.gr/Greek/Studies/Post	Information on M Ενδιαφέροστε για με-		778	12/19/97	C:\NetCr15.html

Εικόνα 4.8: Δεδομένα του πίνακα HTMLpages

Τα βήματα που ακολουθεί ο Web Crawler για το κατέβασμα των σελίδων και την δημιουργία της βάσης είναι τα εξής:

Βήμα 1

Ο Web Crawler αρχίζει από την κεντρική σελίδα του κάθε site η οποία εξ ορισμού είναι η index.html (για παράδειγμα όταν δώσουμε σε έναν browser το URL <http://www.aueb.gr> η σελίδα που εμφανίζεται είναι η <http://www.aueb.gr/index.html>) και κατεβάζει την συγκεκριμένη σελίδα χωρίς όμως να την αποθηκεύσει τοπικά στον δίσκο.

Βήμα 2

Στην συνέχεια σαρώνει την τρέχουσα σελίδα και εντοπίζει όλους τους συνδέσμους που υπάρχουν σε αυτή. Εντοπίζει μόνο τους στατικούς συνδέσμους και αγνοεί αυτούς που υπάρχουν μέσα σε frames ή image maps. Το ίδιο κάνει και για συνδέσμους που δείχνουν σε CGI scripts ή άλλα προγράμματα.

Βήμα 3

Κάθε σχετικός και όχι πλήρες σύνδεσμος που ανακαλύφθηκε στο προηγούμενο βήμα επεκτείνεται στη πλήρη του μορφή. Για παράδειγμα, σχετικοί σύνδεσμοι της μορφής/main.html ή/info.html μετατρέπονται σε πλήρες μονοπάτι της μορφής <http://<όνομα server>/<όνομα σελίδας>>. Από τους πλήρεις συνδέσμους που προέκυψαν αγνοούνται αυτοί που δείχνουν σε σελίδες εκτός της περιοχής που είναι ενεργή, ώστε να κατέβουν οι σελίδες μόνο αυτής της περιοχής. Επίσης, αγνοούνται και οι σύνδεσμοι σε σελίδες οι οποίες έχουν ήδη κατέβει και επεξεργαστεί ώστε να αποφευχθούν οι κύκλοι. Τέλος, οι σύνδεσμοι που απομένουν από την όλη επεξεργασία αποθηκεύονται ως τελευταία φύλλα του δέντρου ώστε να τους επεξεργαστεί στην συνέχεια ο Web Crawler

Βήμα 4

Εντοπίζεται ο τίτλος της ενεργής σελίδας με την βοήθεια των HTML ετικετών `<title>` και `</title>`. Αν η σελίδα δεν έχει τίτλο της δίνεται αυτόματα από το πρόγραμμα ο τίτλος "UnTitled". Σαν περιγραφή της σελίδας συλλέγονται οι πρώτοι 200 χαρακτήρες που υπάρχουν μετά τον τίτλο της σελίδας αγνοώντας τους χαρακτήρες των HTML ετικετών.

Βήμα 5

Στη συνέχεια αφαιρούνται από την σελίδα τα HTML tags και οι "λευκοί" χαρακτήρες (linefeed, tabs κ.α.) και το κείμενο που έχει προκύψει αποθηκεύεται τοπικά στο δίσκο.

στον κατάλογο που έχουμε αναφέρει. Το όνομα αρχείου που αποθηκεύεται η κάθε σελίδα είναι το NetCrl<page_Id>. Για παράδειγμα, η σελίδα με page_ID 23 αποθηκεύεται στο αρχείο NetCrl23. Με τα στοιχεία που συλλέχθηκαν από τα προηγούμενα βήματα ενημερώνεται ο πίνακα HTMLpages.

Βήμα 6

Ο Web Crawler συνεχίζει την λειτουργία του με αυτόν τον τρόπο για όλες τις σελίδες που έχουν απομείνει και τις οποίες δεν έχει ακόμα κατεβάσει ωσότου είτε κατεβάσει όλες τις σελίδες όλων των περιοχών που υπάρχουν στην συλλογή είτε διακόψει την λειτουργία του ο χρήστης.

Στην συνέχεια αφού έχουν κατέβει όλες οι σελίδες της συλλογής τοπικά στον δίσκο, μπορεί να αρχίσει το φόρτωμα της βάσης και η δεικτοδότηση των σελίδων με την επιλογή Grab/Load Database από το μενού. Με αυτήν την λειτουργία το πρόγραμμα του Web Crawler χρησιμοποιεί την τεχνολογία ODBC (Open DataBase Connectivity) για να φορτώσει τα δεδομένα που υπάρχουν στον Access πίνακα HTMLpages σε αντίστοιχο πίνακα της Oracle που θα αναφέρουμε παρακάτω.

Η δημιουργία της συλλογής του συστήματος ανάκτησης βασίστηκε στον χάρτη αποθεμάτων που υπάρχουν στο ελληνικό χώρο ο οποίος είναι διαθέσιμος σε μερικές σελίδες του Forthnet, συγκεκριμένα στο σύνδεσμο HellasMap της κεντρικής σελίδας. Ακολουθώντας την παραπάνω διαδικασία σχηματίστηκε μια συλλογή από 2256 περίπου σελίδες που ανήκουν στη κατηγορία "Health and Medicine". Αυτή η συλλογή προήλθε από το κατέβασμα των σελίδων από 30 περίπου περιοχές που υπάρχουν στον ελληνικό χώρο κάτω από αυτή την κατηγορία. Βέβαια θα πρέπει να πούμε ότι ορισμένες από αυτές τις περιοχές περιείχαν σελίδες μόνο στην αγγλική έκδοση τους ενώ άλλων οι σελίδες παράγονταν δυναμικά οπότε και ήταν αδύνατη η συλλογή τους από τον Web Crawler. Στον πίνακα 4.9 αναφέρουμε ενδεικτικά τις περιοχές από τις οποίες σχηματίστηκε η συλλογή.

Athens & Saronic Gulf

http://www.forthnet.gr/omoio	"A.Stavrakaki - I.Siglinaki" Pharmacies
http://www.mdnet.gr/adelco	ADELCO
http://www.forthnet.gr/asclepeion	Asclepeion Voulas Hospital
http://www.avlmea.gr/	AVL MEA AG Medical Equipment
http://www.ebedent.com	Ebedent S.A. Medical and Dental Supplies
http://www.elinyae.gr	Hellenic Institute for Occupational Health and Safety (EKIMUAE)
http://www.eyenet.gr	HELLENIC OPHTHALMOLOGICAL SOCIETY
http://www.forthnet.gr/hua/	HELLENIC UROLOGICAL ASSOCIATION
http://www.iatron.com	IATRONE S.A. Medical & Diagnostic Equipment
http://www.istos.net.gr/mdm	M.D.M.

http://www.eexi.gr/seiap/	Medical Resident Association of Athens and Pireus (S.E.I.A.P.)
http://www.mednet.gr/	Mednet Hellas
http://www.domi.gr/metaxas	METAXAS DIAGNOSTICS
http://www.hol.gr/greece/medical/opt/	Ophthalmology in Cyberspace
http://www.businesshellas.com/ortho	Orthopaedics - soft orthopaedic products
http://www.etba.gr/~dalu	Pediatric Pages
http://www.etba.gr/~dalu	Pediatric pages. Information for Parents
http://www.aias.net/RelaxPalace	RELAX PALACE elderly care unit
http://www.forthnet.gr/sigmamed	Sigma Medical
http://www.aias.net/thalpori	THALPORI elderly care unit
Eastern Crete	
http://www.forthnet.gr/gpapad/	Dr. George Papadakis
http://www.natural-cosmetics.gr	Natural Cosmetics
http://www.edu.uch.gr/~venicu/	Venizelion Hospital - Intensive Care Unit
Macedonia	
http://www.waternet.diavlos.gr/social/elepap/	E.L.E.P.A.P.
http://www.waternet.diavlos.gr/newsstand/hospital/	Geniko Nosokomeio Edessas
Thessaloniki	
http://www.asklipios.gr	Asklipios medical instruments
http://www.biotrast.techpath.gr/	BIOTRAST S.A.
http://www.disabled.gr/	Disabled Hellas
http://www.med.auth.gr/medsurf/	MEDICUS - Greek Medical Indexer
http://www.med.auth.gr/Neuropage	Neuropage
http://www.diavlos.gr/orto96/ortowww/orto96.htm	Orthopedic Pages
http://www.diavlos.gr/eaeibe/fotis.htm	Society of Anesthesiology and Intensive Medicine of Northern Greece
http://www.med.auth.gr/~karanik/	The Medical Acupuncture Web Page

Πίνακας 4.9: Οι περιοχές που περιέχονται στη συλλογή του Συστήματος Ανάκτησης

4.3 Αρχιτεκτονική του Συστήματος Ανάκτησης

Το Σύστημα Ανάκτησης Ελληνικών Κειμένων δημιουργήθηκε με το Oracle ConText Option και όλες οι λειτουργίες του υλοποιούνται διαμέσου σχεσιακών πινάκων. Ο κύριος πίνακας του συστήματος ανάκτησης είναι ο *HTMLpages* ο οποίος κρατά πληροφορίες σχετικά με την κάθε σελίδα που υπάρχει στη συλλογή. Η μορφή και η επεξήγηση των πεδίων του πίνακα παρουσιάζονται στο παρακάτω σχήμα.

<i>HTMLpages</i>		
<i>page_ID</i>	Number	Ενα μοναδικό αναγνωριστικό για κάθε σελίδα
<i>page_URL</i>	Varchar2(100)	Το πλήρες URL της σελίδας
<i>page_Title</i>	Varchar2(100)	Ο τίτλος της σελίδας
<i>page_Description</i>	Varchar2(200)	Περιγραφή της σελίδας
<i>page_Size</i>	Number	Το μέγεθος της σελίδας σε bytes
<i>page_Text</i>	Text(100)	Το πλήρες όνομα αρχείου στο οποίο βρίσκεται τοπικά η σελίδα

Κάθε εγγραφή στον πίνακα *HTMLpages* αντιστοιχεί σε μία σελίδα και το περιεχόμενο της κάθε σελίδας βρίσκεται στο φυσικό αρχείο που προσδιορίζεται από το πεδίο *page_Text*.

Ο πίνακας *HTMLpages* γεμίζει με δεδομένα κάθε φορά που τρέχουμε το πρόγραμμα του Web Crawler και επιλέγουμε να φορτώσουμε την βάση όπως περιγράψαμε παραπάνω. Βλέπουμε δηλαδή ότι υπάρχει αντιστοιχία μεταξύ του πίνακα *HTMLpages* στην βάση *Contents.mdb* του Web Crawler και του πίνακα *HMTLpages* που χρησιμοποιεί το ConText. Ο λόγος που επιλέχθηκε αυτή η αρχιτεκτονική των δύο κατά τα άλλα πανομοιότυπων πινάκων ήταν ότι θέλαμε να κρατήσουμε το πρόγραμμα του Web Crawler αυτόνομο από το όλο σύστημα ανάκτησης. Το πρόγραμμα του Web Crawler θα μπορούσε να τρέχει και να ενημερώνει το πίνακα *HMTLpages* της βάσης *Contents.mdb* όσο συχνά θέλουμε χωρίς να αλλοιώνονται τα δεδομένα του πίνακα *HTMLpages* του ConText και επηρεάζεται η λειτουργία του συστήματος ανάκτησης. Οταν είμαστε πλέον σίγουροι για τις σελίδες που έχει κατεβάσει ο Web Crawler, τότε μπορούμε να αντιγράψουμε τα περιεχόμενα του πίνακα *HTMLpages* της βάσης *Contents.mdb* στον αντίστοιχο Oracle πίνακα και να αρχίσει από 'κει και πέρα η δεικτοδότηση των σελίδων. Παρατηρούμε δηλαδή ότι ο πλεονασμός που υπάρχει προσφέρει πρόσθιτη ευελιξία στην λειτουργία του όλου συστήματος.

Επίσης, το σύστημα ανάκτησης κάνει χρήση και δύο άλλων σχεσιακών πινάκων που χρησιμοποιούνται κατά την λειτουργία της ανάκτησης και επιστροφής των αποτελεσμάτων. Ο πρώτος είναι ο πίνακας *Temp_Res* (Temporary Results) ο οποίος χρησιμοποιείται για να κράτα τα αποτελέσματα της κάθε αναζήτησης και ο δεύτερος είναι ο πίνακας *Stopwords* ο οποίος χρησιμοποιείται κατά την επεξεργασία της ερώτησης του χρήστη και κρατά τις τετριμμένες λέξεις της ελληνικής γλώσσας^[2], 376 σε αριθμό. Η μορφή και η επεξήγηση των πεδίων των δύο πινάκων παρουσιάζεται στο παρακάτω σχήμα.

<u>Temp_Res</u>		
TEXTKEY	Varchar2(64)	Αποθηκεύεται το αναγνωριστικό του κειμένου που ικανοποιεί τα κριτήρια της αναζήτησης. Ετσι αν μία σελίδα από τον πίνακα HTMLpages με αναγνωριστικό page_ID ικανοποιεί τα κριτήρια αναζήτησης, το αναγνωριστικό της αποθηκεύεται σε αυτό το πεδίο.
SCORE	Number	Η βαθμολογία (βαθμός σχετικότητας) που πήρε το συγκεκριμένο κείμενο.
CONID	Number	Ενα αναγνωριστικό της ερώτησης του χρήστη. Επειδή στο πίνακα Temp_Res αποθηκεύονται τα αποτελέσματα όλων των ερώτησεων όλων των χρηστών πρέπει να υπάρχει τρόπος να διακρίνουμε ποιά αποτελέσματα αντιστοιχούν σε ποια ερώτηση.

<u>Stopwords</u>		
Sw_ID	number	Ενα αναγνωριστικό για κάθε τετριμμένη λέξη
Stopw	Varchar2(20)	Η τετριμμένη λέξη

Για να μπορεί το ConText να προσφέρει υπηρεσίες ανάκτησης από τον πίνακα *HTMLpages* και ειδικότερα από την στήλη *page_Text*, έχουμε αποδώσει σε αυτόν μια συγκεκριμένη πολιτική η οποία μετατρέπει αυτήν την στήλη σε στήλη κειμένου. Η πολιτική αυτή έχει το όνομα *pages_policy* και οι ρυθμίσεις που έχουν γίνει για την κάθε μία κατηγορία που απαιτεί το ConText (βλ. παρ. 3.3.1.1) είναι οι εξής:

Κατηγορία Data store : OSFILE

Το περιεχόμενο των σελίδων αποθηκεύεται σε φυσικά αρχεία στο δίσκο και μόνο το όνομα τους αποθηκεύεται στον πίνακα *HTMLpages* στη στήλη *page_Text*. Με αυτόν τον τρόπο δεν επιβαρύνουμε σε μέγεθος και απόδοση των πίνακα *HTMLpages* αλλά επιπλέον μπορούμε να σβήσουμε τα φυσικά αρχεία των σελίδων στον δίσκο αφού έχει γίνει η δεικτοδότησή τους και η κατασκευή των ευρετηρίων. Το τελευταίο είναι ιδιαίτερα σημαντικό όταν αποφασίσουμε να δεικτοδοτήσουμε μεγάλο μέρος του ελληνικού χώρου στον οποίο ο όγκος των πληροφοριών είναι πολύ μεγάλος για να αποθηκεύεται τοπικά.

Κατηγορία Filter : FILTER NOP

Το περιεχόμενο των σελίδων είναι σε απλή ASCII μορφή ύστερα από την επεξεργασία που έχει γίνει στο πρόγραμμα του Web Crawler. Ο λόγος που δεν χρησιμοποιήσαμε την επιλογή HTML FILTER είναι διότι αυτό το φύλτρο λειτουργεί

μόνο αν τα περιεχόμενα των σελίδων αποθηκεύονταν απευθείας στον πίνακα *HTMLpages*.

Κατηγορία Lexer : BASIC LEXER

Χρησιμοποιούνται οι προκαθορισμένες ρυθμίσεις.

Κατηγορία Engine : GENERIC ENGINE

Χρησιμοποιούνται οι προκαθορισμένες ρυθμίσεις.

Κατηγορία Wordlist : GENERIC WORDLIST

Χρησιμοποιούνται οι προκαθορισμένες ρυθμίσεις. Δεν ενεργοποιούμε τη λειτουργία Soundex στις ερωτήσεις του χρήστη. Υπάρχει η δυνατότητα αποκοπής καταλήξεων και Fuzzy ταιριάσματος μόνο για τους αγγλικούς όρους χρησιμοποιώντας τους τελεστές που έχουμε αναφέρει. Η υλοποίηση της αποκοπής καταλήξεων για τους ελληνικούς όρους περιγράφεται παρακάτω

Κατηγορία Stoplist : GENERIC STOPLIST

Σε αυτήν την κατηγορία αντικαταστήσαμε τις αγγλικές τετριμμένες λέξεις με τις αντίστοιχες ελληνικές. Παρόλο που είχαμε στην διάθεση μας 376 ελληνικές τετριμμένες λέξεις επελέγησαν από αυτές μόνο 255 διότι μέχρι τόσες υποστηρίζει το ConText στην παρούσα έκδοσή του 7.3.

Η δημιουργία της πολιτικής *pages_policy* γίνεται με την χρήση των παρακάτω PL/SQL εντολών

```
ctx_ddl.create_policy (policy_name => 'pages_policy',
                      colspec      => 'HTMLpages.page_Text',
                      dstore_pref  => 'DEFAULT_OSFILe',
                      filter_pref   => 'DEFAULT_NULL_FILTER',
                      lexer_pref    => 'DEFAULT_LEXER',
                      engine_pref   => 'DEFAULT_INDEX',
                      wordlist_pref => 'NO_SOUNDEx',
                      stoplist_pref => 'GREEK_STOPLIST')
```

όπου οι τετριμμένες λέξεις της ελληνικής γλώσσας ορίστηκαν με τις εντολές

```
ctx_ddl.set_attribute('STOP_WORD', 'ΚΑΙ', 1);
ctx_ddl.set_attribute('STOP_WORD', 'ΔΕΝ', 2);
ctx_ddl.set_attribute('STOP_WORD', 'ΜΗ', 3);
ctx_ddl.set_attribute('STOP_WORD', 'ΤΟΥ', 4);
ctx_ddl.set_attribute('STOP_WORD', 'ΤΗΣ', 5);
ctx_ddl.set_attribute('STOP_WORD', 'ΤΑ', 6);
...
... κ.ο.κ. για όλες τις τετριμμένες λέξεις
...
```

```
ctx_ddl.create_preference('GREEK_STOPLIST',
    'The Greek stopwords',
    'GENERIC STOP LIST');
```

Από τη στιγμή που αποδόθηκε η πολιτική *pages_policy* στον πίνακα *HTMLpages* μπορεί να γίνει η δεικτοδότηση των περιεχομένων των σελίδων ώστε να είναι δυνατή η ανάκτηση. Αυτό πετυχαίνεται με την PL/SQL εντολή

```
ctx_ddl.create_index('pages_policy');
```

η οποία παίρνει ως όρισμα το όνομα μιας πολιτικής και με βάση τις ρυθμίσεις που έχουν γίνει σε αυτή, υλοποιεί την δεικτοδότηση των περιεχομένων της στήλης κειμένου που στην περίπτωσή μας είναι η *page_Text*. Η εκτέλεση αυτής της εντολής έχει ως αποτέλεσμα την αυτόματη δημιουργία από το ConText τριών πινάκων δεικτοδότησης:

DR_nnnnn_IIT
DR_nnnnn_KTB
DR_nnnnn_LST

όπου *nnnnn* είναι ένα πενταψήφιο αναγνωριστικό που δίνεται αυτόματα από το σύστημα σε κάθε πολιτική που σχετίζεται με κάποιον πίνακα.

Ο πίνακας *DR_nnnnn_IIT* είναι ο κύριος πίνακας δεικτοδότησης και υλοποιεί μια ανεστραμμένη λίστα όρων. Αποθηκεύει κάθε λέξη που υπάρχει στο περιεχόμενο των στηλών κειμένου μαζί με τα κείμενα στα οποία αυτή υπάρχει καθώς και την θέση εμφάνισής της μέσα σε αυτές

<i>DR_nnnnn_IIT</i>		
<i>WORD_TEXT</i>	Varchar2(64)	Η λέξη που δεικτοδοτείται
<i>FIRST_DOC</i>	Number(38)	Το DocId του πρώτου κειμένου που υπάρχει στη λίστα WORD INFO
<i>DOCLSIZE</i>	Number(38)	Το μήκος, σε bytes, της λίστας WORD INFO
<i>WORD_INFO</i>	Long Raw	Λίστα των DocIds των κειμένων που υπάρχει η λέξη και η θέση εμφάνισής της μέσα σε αυτά.

Ο πίνακας *DR_nnnnn_KTB* αντιστοιχεί σε κάθε κείμενο που υπάρχει στη βάση ένα μοναδικό αναγνωριστικό. Δημιουργείτε δηλαδή μίας προς μία (1:1) αντιστοιχία του κλειδιού που δίνουμε εμείς σε κάθε κείμενο - στην περίπτωσή μας το κλειδί είναι το *page_ID* - και του κλειδιού που δίνει αυτόματα το ConText. Σε όλες τις εσωτερικές λειτουργίες του ConText χρησιμοποιούνται τα αναγνωριστικά των κειμένων όπως ορίζονται στον πίνακα *DR_nnnnn_KTB* και όχι αυτά που έχει δώσει χειρονακτικά ο χρήστης.

<i>DR_nnnnn_KTB</i>		
<i>TEXTKEY</i>	Varchar2(32)	Το αναγνωριστικό του κειμένου που έχει δώσει ο χρήστης. Στη περίπτωσή μας η τιμή του πεδίου page ID του πίνακα HTMLpages.
<i>DOCID</i>	Number(38)	Το αναγνωριστικό που δίνει το ConText.

Για λόγους πληρότητας της παρουσίασης των πινάκων δεικτοδότησης που χρησιμοποιεί το ConText αναφέρουμε, ότι ο πίνακας *DR_nnnnn_LST* κρατάει το επόμενο διαθέσιμο εσωτερικό DocId το οποίο χρησιμοποιείται στον πίνακα *DR_nnnnn_KTB* που μόλις αναφέραμε καθώς και όλα τα εσωτερικά DocIds των κειμένων που έχουν αλλάξει ή διαγραφεί από την βάση - δηλαδή από τον πίνακα *HTMLpages* - ώστε να ξέρει το ConText ποια κείμενα χρειάζονται επαναδεικτοδότηση και ποια δεν θα πρέπει να ληφθούν υπόψη κατά την ανάκτηση δηλαδή κατά το ψάξιμο του πίνακα *DR_nnnnn_IIT*.

Τέλος, υπάρχει και ο πίνακας *DR_nnnnn_IIW* ο οποίος δημιουργείται μόνο αν έχουμε ενεργοποιήσει την επιλογή SOUNDEX στις ερωτήσεις του χρήστη. Αυτός ο πίνακας κρατά κάθε λέξη που αναγνωρίζεται από την SOUNDEX συνάρτηση μαζί με την ομάδα των λέξεων στο οποίο ανήκει αυτή η λέξη.

<i>DR_nnnnn_IIW</i>		
<i>WORD</i>	Varchar2(15)	Η λέξη που αναγνωρίστηκε από την SOUNDEX συνάρτηση
<i>GROUP1</i>	Varchar2(15)	Το αναγνωριστικό του πρώτου SOUNDEX group στο οποίο ανήκει η λέξη
<i>GROUP2</i>	Varchar2(15)	Δεσμευμένο για μελλοντική χρήση
<i>GROUP3</i>	Varchar2(15)	Δεσμευμένο για μελλοντική χρήση

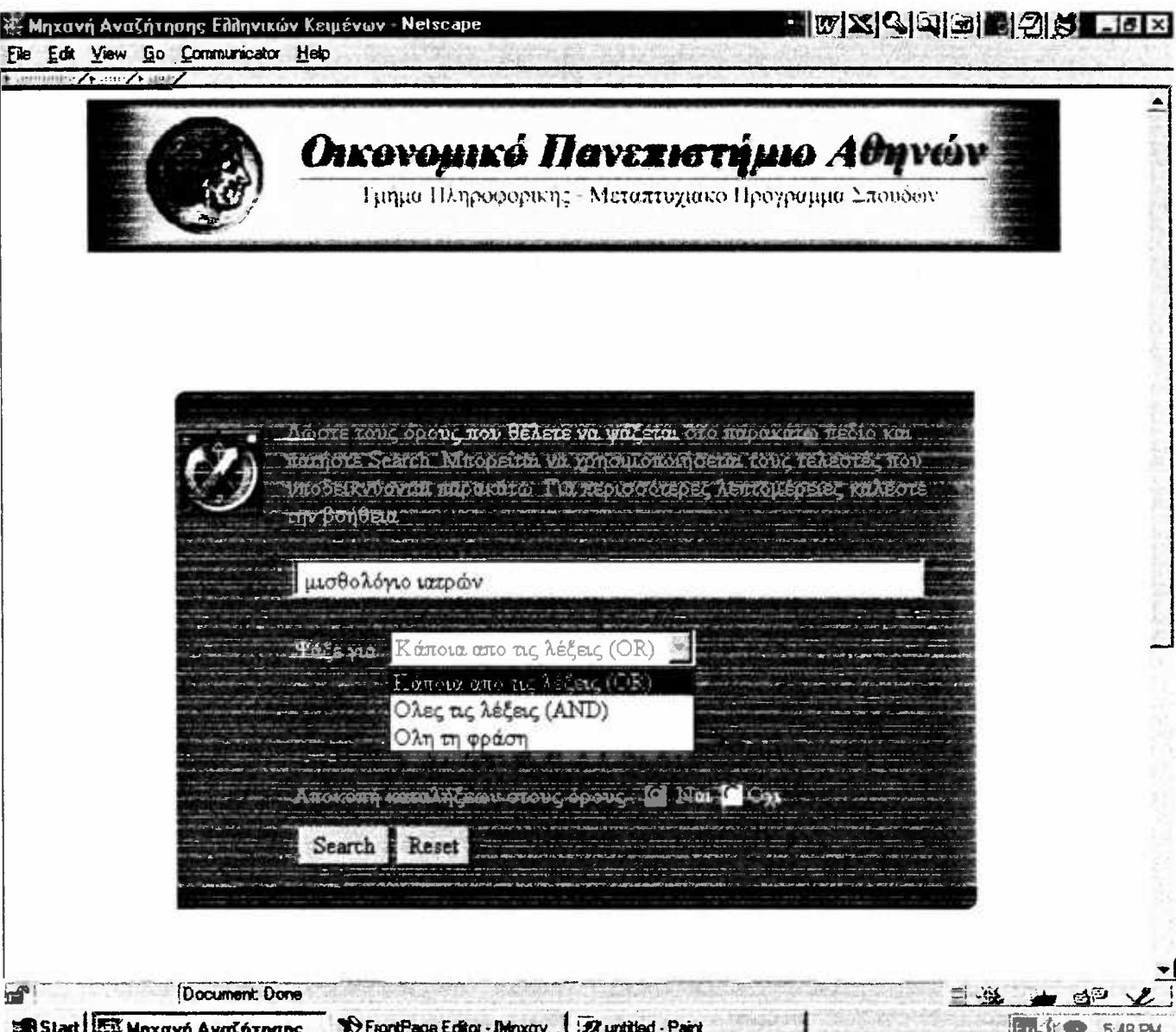
Θα πρέπει να πούμε ότι το ConText παρά το γεγονός ότι υποστηρίζει την δυνατότητα αποκοπής καταλήξεων στην αγγλική γλώσσα, στον πίνακα *DR_nnnnn_IIT* δεν αποθηκεύει τις ρίζες των όρων αλλά όλους τους όρους που συναντά στα περιεχόμενα των κειμένων, άσχετα αν έχουν κοινή ρίζα ή όχι. Με αυτό τον τρόπο το μέγεθος του πίνακα *DR_nnnnn_IIW* και το πλήθος των όρων γίνεται ιδιαίτερα μεγάλο για μεγάλες συλλογές αλλά έχει το πλεονέκτημα του ότι αφήνει στον χρήστη την απόφαση για το αν θέλει να χρησιμοποιήσει αποκοπή καταλήξεων στην ερώτησή του ή αν θέλει να κάνει αναζήτηση με ακριβές ταίριασμα. Στην περίπτωση της αποκοπής καταλήξεων, αυτή υλοποιείται διαμέσου του τελεστή επέκτασης όρου \$ όπως έχουμε ήδη αναφέρει και με βάση τους επιπλέον όρους που παράγει αυτός ο τελεστής γίνεται η αναζήτηση στον πίνακα δεικτοδότησης *DR_nnnnn_IIT*. Παρατηρούμε δηλαδή ότι στο ConText η λειτουργία της αποκοπής καταλήξεων λειτουργεί με αντίστροφο τρόπο από αυτόν που συναντάμε σε παραδοσιακά συστήματα IRS.

Επίσης θα πρέπει να αναφέρουμε ότι δεν χρειάζεται αναδιοργάνωση των πινάκων δεικτοδότησης κάθε φορά που γίνεται εισαγωγή ή/και διαγραφή καινούργιου κειμένου στη συλλογή. Στην περίπτωση καινούργιου κειμένου, το ίδιο το ConText βάζει μια αίτηση δεικτοδότησης του σε μία ουρά (queue) και όποτε υπάρχει διαθέσιμος ConText εξυπηρετητής γίνεται η δεικτοδότηση των περιεχομένων του και η ενημέρωση των αντίστοιχων πινάκων χωρίς να απαιτείται η παρέμβαση του χρήστη για τον χειρισμό της δεικτοδότησης. Στην περίπτωση της διαγραφής, απλά ενημερώνεται ο πίνακας *DR_nnppn_LST* με το αναγνωριστικό του κειμένου που διαγράφηκε και έτσι οι όροι αυτού του κειμένου δεν συμμετέχουν στις αναζητήσεις. Βέβαια στη περίπτωση που έχουν διαγραφεί πολλά κείμενα από την συλλογή, είναι προτιμότερο να διαγραφούν οι πίνακες δεικτοδότησης των κειμένων (εντολή *ctx_ddl.drop_index('pages_policy')*) και να χτιστούν ξανά από την αρχή. Ενδεικτικά αναφέρουμε ότι η δεικτοδότηση των 2256 σελίδων της συλλογής και η δημιουργία των αντίστοιχων πινάκων κράτησε περίπου 5 λεπτά.

4.4 Λειτουργία του Συστήματος Ανάκτησης

Οπως έχουμε ήδη αναφέρει, το σύστημα ανάκτησης που υλοποιήθηκε αποτελεί WWW εφαρμογή και για το λόγο αυτό χρησιμοποιήθηκε η τεχνολογία που παρέχει η Oracle για την υποστήριξη και λειτουργία τέτοιων εφαρμογών όπως Oracle Web Server, Oracle Web Listener, Oracle Universal Server κ.λ.π. προσφέροντας ολοκλήρωση και ενοποίηση διαφορετικών τεχνολογιών σε ένα ενιαίο περιβάλλον προγραμματισμού.

Το σύστημα ανάκτησης μπορεί και δέχεται αιτήσεις στην διεύθυνση <http://heltrun.aueb.gr:6052/retrieve.htm>, αρκεί να είναι σε λειτουργία ο Oracle Web server - για να είναι εφικτή η HTTP (HyperText Transport Protocol) επικοινωνία μεταξύ Web browser και Web server - και ένας τουλάχιστον ConText εξυπηρετητής ο οποίος είναι υπεύθυνος για την υποστήριξη των λειτουργιών της ανάκτησης. Οταν ισχύουν και οι δύο αυτές προϋποθέσεις, τότε παρουσιάζεται στην διεύθυνση που αναφέραμε η κεντρική σελίδα του συστήματος ανάκτησης (βλ. εικόνα 4.10)



Εικόνα 4.10: Κεντρική σελίδα του συστήματος ανάκτησης

Οπως φαίνεται και στην εικόνα, το σύστημα ανάκτησης μπορεί και προσφέρει στον χρήστη ορισμένες επιλογές αναζήτησης διαμέσου γραφικής διεπαφής χωρίς να χρειάζεται η πληκτρολόγηση επιπλέον τελεστών στην περιοχή εισαγωγής της ερώτησης του χρήστη. Πιο συγκεκριμένα, ο χρήστης μπορεί να επιλέξει να χρησιμοποιήσει αποκοπή καταλήξεων ή όχι στην ερώτησή του, ενώ μπορεί να επιλέξει να συνδέονται οι όροι της ερώτησης του με τους λογικούς τελεστές "AND" ή "OR". Ο λογικός τελεστής "OR" είναι ο προκαθορισμένος. Τέλος, ο χρήστης έχει και την επιλογή να εκτελέσει μια αναζήτηση με ακριβές ταίριασμα χωρίς να γίνει καμία περαιτέρω επεξεργασία στην ερώτησή του. Αυτή η τελευταία επιλογή χρησιμοποιείται όπως θα δούμε παρακάτω για την εκτέλεση αναζητήσεων στα

αγγλικά χρησιμοποιώντας το πλήθος των τελεστών που προσφέρει το ConText. Το σύστημα ανάκτησης θεωρεί εξ ορισμού ότι η ερώτηση του χρήστη είναι στην ελληνική γλώσσα και αυτό φαίνεται από τα βήματα τα οποία εκτελεί και την επεξεργασία που κάνει στην ερώτηση του χρήστη προτού εκτελέσει μια αναζήτηση.

Η ερώτηση του χρήστη μπορεί να έχει την μορφή όρων ή μιας φράσης ή ακόμα μπορεί να είναι και σε φυσική γλώσσα. Οταν ο χρήστης πατήσει το κουμπί της αναζήτησης τότε το σύστημα ανάκτησης εκτελεί τα ακόλουθα βήματα:

Βήμα 1

Σαρώνεται η ερώτηση του χρήστη και αφαιρούνται όλες οι τετριμμένες λέξεις της ελληνικής γλώσσας. Αυτό γίνεται με την βοήθεια του πίνακα *stopwords* που αναφέραμε και ο οποίος περιέχει το πλήρες σύνολο των τετριμμένων λέξεων της ελληνικής. Αν πρόκειται για ερώτηση του χρήστη στα αγγλικά αυτή θα παραμείνει αναλλοίωτη καθώς δεν υπάρχουν ελληνικές τετριμμένες λέξεις σε αυτήν.

Βήμα 2

Οι όροι της ερώτησης που απέμειναν από το προηγούμενο βήμα τοποθετούνται σε ένα προσωρινό πίνακα και αποκόπονται οι καταλήξεις τους αναλόγως με το αν το έχει επιλέξει ο χρήστης ή όχι. Η αποκοπή καταλήξεων της ελληνικής υλοποιείται με έμμεσο τρόπο με την χρήση χαρακτήρων μπαλαντέρ. Πιο συγκεκριμένα, σε κάθε όρο της ερώτησης εφαρμόζεται μία συνάρτηση αποκοπής καταλήξεων της ελληνικής^[4] η οποία δοθέντος ενός όρου επιστρέφει την ρίζα αυτού του όρου. Στην συνέχεια σε κάθε ρίζα των όρων της ερώτησης επισυνάπτεται ο χαρακτήρας "%", ο οποίος μπορεί να αντιπροσωπεύσει οποιαδήποτε ακολουθία χαρακτήρων οποιουδήποτε μήκους. Για παράδειγμα, ο όρος "Εκπαίδευση" ύστερα από την επεξεργασία που αναφέραμε, μετατρέπεται σε "Εκπαίδ%" και επειδή στο ευρετήριο *DR_nnppp_11T* αποθηκεύονται οι όροι στην πλήρη τους μορφή και όχι οι ρίζες τους, μια αναζήτηση με την λέξη "Εκπαίδ%" θα επιστρέψει και τα κείμενα που περιέχουν τους όρους "Εκπαίδευτικός", "Εκπαίδευση", "Εκπαίδευτική" κ.λ.π. τα αποτελέσματα δηλαδή που θα επέστρεφε ένα παραδοσιακό IRS κάνοντας αποκοπή καταλήξεων. Θα πρέπει να πούμε ότι όλοι οι όροι των κειμένων αποθηκεύονται στο ευρετήριο *DR_nnppp_11T* με κεφαλαία και σε UNICODE format (δηλαδή ο χαρακτήρας "έ" αποθηκεύεται ως "E") και πριν εκτελεστεί η αναζήτηση το ConText μετατρέπει αυτόματα σε κεφαλαία την ερώτηση του χρήστη κατά τρόπο μη ορατό και έτσι δεν χρειάζεται να ανησυχούμε για διαφορές μεταξύ μικρών-κεφαλαίων και κωδικών των χαρακτήρων της ερώτησης. Παρατηρούμε, ότι όλη η διαδικασία που περιγράφεται στο βήμα 2 έχει ισχύ και πετυχαίνει μόνο για ελληνικές ερωτήσεις διότι στους όρους της ερώτησης εφαρμόζεται συνάρτηση αποκοπής καταλήξεων της ελληνικής και όχι της αγγλικής γλώσσας.

Βήμα 3

Οι επεξεργασμένοι όροι που προέκυψαν στο προηγούμενο βήμα, συνδέονται μεταξύ τους με τους λογικούς τελεστές "AND" ή "OR" ανάλογα με την επιλογή του χρήστη. Για παράδειγμα η ερώτηση "μισθολόγιο ιατρών" με την χρήση αποκοπής καταλήξεων και του τελεστή "OR" μετατρέπεται σε "μισθολογ% | ιατρ%". Αν ο χρήστης έχει επιλέξει αναζήτηση με ακριβές ταίριασμα τότε η ερώτηση του δεν υφίσταται καμία επεξεργασία - δεν πραγματοποιούνται τα βήματα 1 και 2 - και όπως είναι δίνεται στο ConText να εκτελέσει την αναζήτηση. Με αυτόν τον τρόπο ο χρήστης έχει την ευχέρεια να εφαρμόσει όλους τους τελεστές που του παρέχει το ConText στην ερώτησή του για να κάνει πολύπλοκες αναζητήσεις. Για παράδειγμα, το αγγλικό stemming θα μπορούσε να το πετύχει με τον τελεστή \$ ή θα μπορούσε να αποδώσει βάρη στους όρους της ερώτησης, είτε αγγλικούς είτε ελληνικούς είτε ταυτόχρονα και στις δύο γλώσσες.

Βήμα 4

Αφού έχει γίνει η επεξεργασία της ερώτησης του χρήστη, αυτή δίνεται στο ConText για να εκτελέσει την αναζήτηση και τα αποτελέσματα αποθηκεύονται στο πίνακα Temp_Res που έχουμε ήδη περιγράψει. Σαν αναγνωριστικό της κάθε ερώτησης δίνουμε το SessionID που είναι ένα αναγνωριστικό συνόδου που δίνεται αυτόματα από το σύστημα κάθε φορά που πραγματοποιείται μία σύνοδος μεταξύ ενός πελάτη (Client) και του Web εξυπηρετητή. Αυτό το αναγνωριστικό εμφανίζεται στο πίνακα Temp_Res στο πεδίο CONID. Με αυτό τον τρόπο μπορούν να εξυπηρετηθούν ταυτόχρονα πολλοί χρήστες χωρίς να μπερδεύονται τα αποτελέσματα των ερωτήσεων τους.

Βήμα 5

Στην συνέχεια χρησιμοποιείται η πράξη της σύνδεσης (inner join) των πινάκων HTMLpages και Temp_Res (διαμέσου των κλειδιών page_ID και TEXTKEY αντίστοιχα) ώστε να πάρουμε τα στοιχεία που θέλουμε (όπως τίτλο, περιγραφή κ.λ.π..) για κάθε σελίδα που ικανοποιεί τα κριτήρια αναζήτησης. Τα αποτελέσματα επιστρέφονται σε μορφή που φαίνεται στη εικόνα 4.11 όπου για κάθε σελίδα εμφανίζεται ο τίτλος της, μία περιγραφή της, η βαθμολογία της, η διεύθυνσή της-URL το οποίο αποτελεί ενεργό σύνδεσμο και ο χρήστης μπορεί να μεταφερθεί σε αυτή αν το επιθυμεί, και τέλος το μέγεθός της. Τα αποτελέσματα επιστρέφονται κατά φθίνουσα σειρά σχετικότητας και στην κορυφή της σελίδας των αποτελεσμάτων είναι τυπωμένη η ερώτηση του χρήστη πριν και μετά την επεξεργασία της που έγινε στα βήματα 1,2,3 καθώς και ο αριθμός των σελίδων που ικανοποιούν τα κριτήρια αναζήτησης. Θα πρέπει να πούμε ότι η βαθμολογία που εμφανίζεται ότι παίρνει η κάθε σελίδα δεν είναι αυτό το οποίο δίνει αυτόματα το ConText αλλά υπολογίζεται από τον τύπο

$$Score_{New} = 100 * Score_{ConText} / MAX Score_{ConText}$$

Όπου $Score_{New}$, η βαθμολογία της κάθε σελίδας, $Score_{ConText}$, η βαθμολογία που δίνει το ConText σε κάθε σελίδα, και $MAX Score_{ConText}$ η μέγιστη βαθμολογία που έχει δόσει το ConText στα αποτελέσματα. Με αυτόν τον τρόπο, η βαθμολογία της κάθε σελίδας βγαίνει ποσοστό επί τοις 100 (%) και εξασφαλίζεται ότι η σελίδα με τη μέγιστη βαθμολογία θα πάρει ποσοστό 100%.

Αξιολόγηση Αποτελέσματων

Αποτελέσματα ανάκτησης - Netscape
File Edit View Go Communicator Help

Εκπαιδευτικό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής - Μεταπτυχιακό Πρόγραμμα Σπουδών

ΚΡΙΤΗΡΙΟ ΑΝΑΖΗΤΗΣΗΣ: ΜΙΣΘΟΛΟΓΙΟ ΙΑΤΡΩΝ
ΕΠΙΦΕΡΓΑΣΜΕΝΟ ΚΡΙΤΗΡΙΟ ΑΝΑΖΗΤΗΣΗΣ: ΜΙΣΘΟΛΟΓ% & ΙΑΤΡ%

ΒΡΕΘΗΚΑΝ 3 ΣΕΛΙΔΕΣ ΣΕ ΣΥΝΟΛΟ 2256 ΣΕΛΙΔΩΝ ΤΗΣ ΣΥΛΛΟΓΗΣ

Σ Σύλλογος Ειδικευομένων Ιατρών Αθηνών-Πειραιώς (Σ.Ε.Ι.Α.Π.) - Νέα
ΝΕΟ ΙΑΤΡΙΚΟ ΜΙΣΘΟΛΟΓΙΟ Μετά τα πρόσφατα δημοσιεύματα του Τύπου και τη δημοσίευση της πρότασης της κοινής επιτροπής μελέτης των νέου ιατρικού μισθολογίου, εκφράζουμε την ανάθεση των Νέων Γι [100%] <http://www.eexi.gr/seiap/seiap1.htm> Size - 17 KB

Σ Σύλλογος Ειδικευομένων Ιατρών Αθηνών-Πειραιώς (Σ.Ε.Ι.Α.Π.) - Απόψεις Μελών
Αγαπητοί, Συνάδελφοι Το καλοκάρι που πέρασε υπήρξε γεμάτο εξέλιξεις στα θέματα που απασχολούν τους νοσοκομειακούς γιατρούς Η.Ε.Ι.Ν.Α.Π. και το νέο Δ.Σ. μέσα σε κλίμα ομοψυχίας και με γνώμον [83%] <http://www.eexi.gr/seiap/article.htm> Size - 7 KB

Σ Σύνδεσμος Εκπαιδευτικού Ιατρών Αθηνών-Πειραιώς
ΝΕΟ ΜΙΣΘΟΛΟΓΙΟ ΝΕΟΣ ΣΥΛΛΟΓΟΣ ΑΠΟΨΕΙΣ ΜΕΛΩΝ E-Mail: seiap@eexi.gr Υπεύθυνος Σελίδων Γιάννης

Document Done Start Microsoft Word - Αποτελέσματα ανάκτησης Netscape Αποτελέσματα ανάκτησης Αποτελέσματα ανάκτησης 8:07 PM

Εικόνα 4.11: Αποτελέσματα ανάκτησης

Βήμα 6

Τέλος, μετά από κάθε αναζήτηση και εμφάνιση των αποτελεσμάτων, διαγράφονται τα αποτελέσματα από τον πίνακα *Temp_Res* ώστε να μην γεμίζει συνέχεια ο πίνακας με άχρηστα αποτελέσματα περασμένων αναζητήσεων και μειώνεται η ταχύτητα προσπέλασής του.



ΚΕΦΑΛΑΙΟ 5

Αξιολόγηση Αποτελεσμάτων

Μεγάλη προσπάθεια και έρευνα με συνεχώς καινούργιες προσεγγίσεις και τεχνικές έχει γίνει από την επιστημονική κοινότητα για την λύση του προβλήματος της αξιολόγησης (evaluation) των συστημάτων ανάκτησης πληροφοριών. Παρόλα αυτά, το πρόβλημα της αξιολόγησης περιλαμβάνει πολλές παραμέτρους και παράγοντες που πρέπει να συνεκτιμήθουν και η λύση του δείχνει ακόμα μακριά και αυτό φαίνεται από το πλήθος των δημοσιεύσεων και των άρθρων που έχουν γίνει κατά καιρούς και που σχετίζονται με αυτό το θέμα.

Η αξιολόγηση ενός συστήματος ανάκτησης αναφέρεται στην εκτίμηση της απόδοσης (efficiency) και της αποτελεσματικότητάς του (effectiveness). Η απόδοση αναφέρεται στην πολυπλοκότητα των πράξεων και στο κόστος λειτουργιών του συστήματος και ένα μέτρο της είναι η ταχύτητα απόκρισης του συστήματος ανάκτησης. Η αποτελεσματικότητα αναφέρεται στην ικανότητα του συστήματος να μπορεί να ανακτεί ένα μεγάλο μέρος από τα σχετικά κείμενα και να απορρίπτει ένα μεγάλο μέρος από τα άσχετα. Συνήθως, ένα σύστημα ανάκτησης αξιολογείται με βάση το πόσο καλά ικανοποιεί επτά βασικές παραμέτρους:

1. Την πληρότητα (coverage) της συλλογής δηλαδή το ποσοστό στο οποίο αυτή καλύπτει τα σχετικά κείμενα.
2. Το χρόνο απόκρισης (time lag) του συστήματος που είναι το μέσο χρονικό διάστημα που μεσολαβεί μεταξύ της υποβολής της ερώτησης του χρήστη και της απάντησης από το σύστημα ανάκτησης.
3. Το τρόπο παρουσίασης (presentation) των αποτελεσμάτων.
4. Τη προσπάθεια (effort) που απαιτείται για την διατύπωση της ερώτησης από τη πλευρά του χρήστη.
5. Το κόστος εισαγωγής νέων κειμένων στη συλλογή.
6. Την απόκριση (recall) του συστήματος που είναι το ποσοστό της σχετικής πληροφορίας που ανακτήθηκε σε μία ερώτηση του χρήστη.

7. Την ακρίβεια (precision) του συστήματος που είναι το ποσοστό της ανακτηθέντας πληροφορίας που είναι σχετική με την ερώτηση του χρήστη.

PRECISION =

Οι παράμετροι 1-5 μπορούν εύκολα μέσα από πειράματα να εκτιμηθούν. Η απόκριση και η ακρίβεια είναι οι παράμετροι εκείνοι οι οποίοι χαρακτηρίζουν ένα σύστημα ανάκτησης και αποτελούν μέτρο της αποτελεσματικότητας του. Όσο πιο αποτελεσματικό είναι ένα σύστημα ανάκτησης τόσο πιο πολύ θα ικανοποιεί το χρήστη αφού θα είναι ικανό να ανακτεί τα σχετικά κείμενα για κάθε ερώτηση και ταυτόχρονα να απορρίπτει τα μη σχετικά.

Στο παρελθόν έχει γίνει διαμάχη σχετικά με το αν η ακρίβεια και η απόκριση είναι οι κατάλληλες παράμετροι για την μέτρηση της αποτελεσματικότητας των συστημάτων ανάκτησης. Μια εναλλακτική πρόταση ήταν η χρήση της απόκρισης και του ποσοστού των μη σχετικών κειμένων που ανακτήθηκαν (Fallout). Παρόλα αυτά το ζεινάρι απόκριση-ακρίβεια συνεχίζει να υπερισχύει ως μέτρο της αποτελεσματικότητας των συστημάτων ανάκτησης λόγω του ότι πρόκειται για δύο καλά ορισμένες και κατανοητές ποσότητες. Στον πίνακα 5.1 φαίνεται η σχέση που υπάρχει ανάμεσα στα σχετικά/μη-σχετικά και ανακτηθέντα/μη-ανακτηθέντα κείμενα της συλλογής για μία ερώτηση του χρήστη.

	ΣΧΕΤΙΚΑ	ΜΗ-ΣΧΕΤΙΚΑ	
ΑΝΑΚΤΗΘΕΝΤΑ	$A \cap B$	$\bar{A} \cap B$	B
ΜΗ-ΑΝΑΚΤΗΘΕΝΤΑ	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	\bar{B}
	A	\bar{A}	N

(N = Αριθμός των κειμένων της συλλογής)

Πίνακας 5.1: Διαχωρισμός των κειμένων της συλλογής

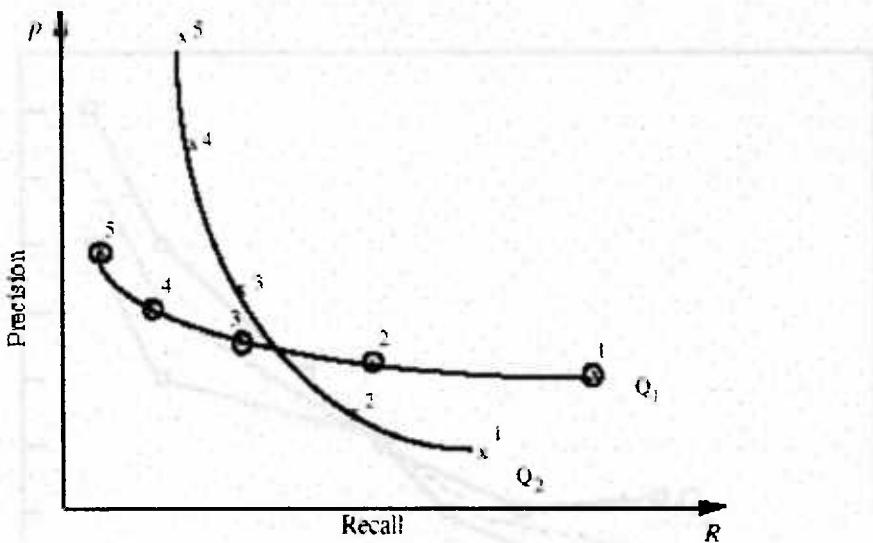
Από τον παραπάνω πίνακα μπορούμε να υπολογίσουμε την απόκριση, την ακρίβεια και το Fallout ως εξής :

$$\text{PRECISION} = \frac{|A \cap B|}{|B|}$$

$$\text{RECALL} = \frac{|A \cap B|}{|A|}$$

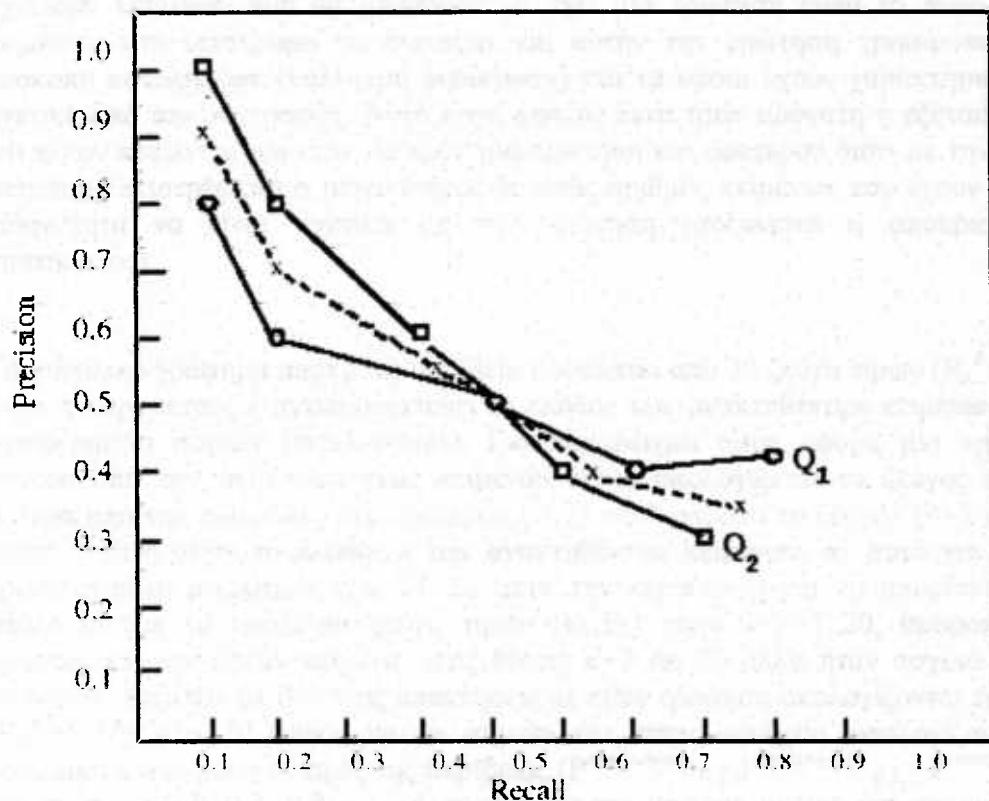
$$\text{FALLOUT} = \frac{|\bar{A} \cap B|}{|\bar{A}|}$$

Για κάθε ερώτηση που τίθεται στο σύστημα μπορούμε να κατασκευάσουμε τον παραπάνω πίνακα και με βάση αυτόν να υπολογίσουμε την ακρίβεια και την απόκριση του συστήματος ανάκτησης για τη συγκεκριμένη ερώτηση. Για καλύτερο έλεγχο της αποτελεσματικότητας του συστήματος συνηθίζεται η απόκριση και η ακρίβεια για μία ερώτηση να μην υπολογίζονται μία μόνο φορά μετά την επιστροφή όλων των κειμένων για αυτήν την ερώτηση, αλλά να υπολογίζονται λιγότερες όπου η παράμετρος λ εκφράζει το πλήθος των κειμένων που έχουν μέχρι στιγμής ανακτηθεί. Επισήμως όσον αφορά μία ερώτηση, ύστερα από την ανάκτηση ενός κειμένου ($\lambda=1$), δύο κειμένων ($\lambda=2$) κ.ο.κ. μέχρι το πλήθος των κειμένων που επιστρέφει το σύστημα ανάκτησης για αυτήν ερώτηση, μπορεί να κατασκευαστεί ο πίνακας 5.1 και να υπολογιστούν κατ' επέκταση τα ζεύγη απόκριση-ακρίβεια για κάθε μία τιμή της παραμέτρου λ . Με αυτόν τον τρόπο μπορούμε να παρακολουθούμε την πρόοδο των τιμών της απόκρισης και της ακρίβειας μέχρι να ανακτηθούν όλα τα κείμενα. Εάν λιγότεροι από το P_λ παριστάνεται την ακρίβεια, το R_λ παριστάνει την απόκριση και το ζεύγος απόκριση-ακρίβεια παριστάνεται από το διατεταγμένο ζεύγος (R_λ, P_λ) . Το σύνολο αυτό των διατεταγμένων ζευγών σχηματίζει το γράφημα απόκριση-ακρίβεια. Γεωμετρικά όταν αυτά τα σημεία ενωθούν σχηματίζουν μια καμπύλη που είναι η καμπύλη της απόκρισης-ακρίβειας (Recall-Precision graph) και δείχνει την αποτελεσματικότητα του συστήματος ανάκτησης για μια συγκεκριμένη ερώτηση του χρήστη (βλ. σχήμα 5.2).



Σχήμα 5.2: Καμπύλη απόκριση-ακρίβεια για δύο ερωτήσεις Q_1 και Q_2 και πέντε τιμές της παραμέτρου λ

Επειδή η εξαγωγή συμπερασμάτων σχετικά με την απόκριση και την ακρίβεια του συστήματος δεν είναι αξιόπιστη όταν οι τιμές τους υπολογίζονται από μία μόνο ερώτηση, στη πράξη υπολογίζουμε τις μέσες τιμές της απόκρισης και της ακρίβειας όπως αυτές προκύπτουν για ένα αριθμό ερωτήσεων. Η συνολική αποτελεσματικότητα του συστήματος ανάκτησης προκύπτει όταν το σύνολο των καμπυλών απόκριση-ακρίβεια, μία για κάθε ερώτηση, συνδυαστεί ώστε να παραχθεί η συνολική καμπύλη απόκριση-ακρίβεια. Χρησιμοποιώντας τα αποτελέσματα που έδωσε το σύστημα ανάκτησης για κάθε μία ερώτηση του χρήστη, υπολογίζουμε την μέση απόκριση (R^{avg}) και μέση ακρίβεια (P^{avg}) για κάθε διακεκριμένη τιμή της τιμή παραμέτρου λ . Τα διατεταγμένα ζεύγη ($R_{\lambda}^{\text{avg}}, P_{\lambda}^{\text{avg}}$) που προκύπτουν με αυτό τον τρόπο σχηματίζουν την συνολική καμπύλη απόκριση-ακρίβεια. Στο σχήμα 5.3 φαίνεται με διακεκομμένη γραμμή η συνολική καμπύλη απόκριση-ακρίβεια που παράγεται από τις καμπύλες απόκριση-ακρίβεια των δύο ερωτήσεων που παρουσιάσαμε στο σχήμα 5.2. Τέλος συνηθίζεται, το συνολικό γράφημα απόκριση-ακρίβεια να σχηματίζεται από τα διατεταγμένη ζεύγη ($R_{\lambda}^{\text{avg}}, P_{\lambda}^{\text{avg}}$) όπου η ακρίβεια υπολογίζεται για διακεκριμένες τιμές τις απόκρισης συνήθως 0.1, 0.2, 0.3, ..., 1. Ο υπολογισμός της ακρίβειας για τις διακεκριμένες τιμές της απόκρισης που αναφέραμε γίνεται με την εφαρμογή παρεμβολής (interpolation) στις αρχικές τιμές των ζευγών ($R_{\lambda}^{\text{avg}}, P_{\lambda}^{\text{avg}}$).



Σχήμα 5.3: Συνολική καμπύλη απόκριση-ακρίβεια για δύο ερωτήσεις

5.1.1 Αποτελέσματα Πειράματος

Ακολουθώντας την μεθοδολογία που περιγράψαμε παραπάνω, εκτελέσαμε πειράματα στο σύστημα ανάκτησης προκειμένου να μετρήσουμε την αποτελεσματικότητα του και ειδικότερα την μεταβολή της ανάλογα με το αν χρησιμοποιούμε αποκοπή καταλήξεων ή όχι στους όρους της ερώτησης. Για το σκοπό αυτό, υποβάλαμε στο σύστημα ανάκτησης δύο φορές 20 ερωτήματα σε ελληνική γλώσσα χρησιμοποιώντας την πρώτη φορά αποκοπή καταλήξεων ενώ τη δεύτερη όχι. Οι ερωτήσεις περιείχαν μεταβλητό αριθμό όρων (από δύο έως τέσσερις όρους σε κάθε ερώτηση) ενώ και στις δύο περιπτώσεις οι όροι των ερωτήσεων συνδέονταν με τον λογικό τελεστή AND για να "στενέψουμε" και να κάνουμε πιο συγκεκριμένη την ερώτηση.

Εφόσον δεν υπήρχε στη διάθεση μας δοκιμαστική συλλογή (test collection) στην οποία είναι γνωστά τα σχετικά κείμενα για κάθε ερώτηση, η σχετικότητα ενός κειμένου σε σχέση με μία ερώτηση αφέθηκε στην υποκειμενική κρίση του συγγραφέα. Η ανακάλυψη των σχετικών κειμένων για μία ερώτηση έγινε με τον ακόλουθο τρόπο: Η ερώτηση υποβαλλόταν στο σύστημα ανάκτησης χρησιμοποιώντας αποκοπή καταλήξεων και το σύνολο των σχετικών κειμένων για αυτήν τη ερώτηση προερχόταν από τα κείμενα που επιστρέφονταν και που

υποκειμενική κρίση θεωρούνταν σχετικά. Με αυτό τον τρόπο ο μέγιστος αριθμός σχετικών κειμένων που θα μπορούσε να έχει μία ερώτηση είναι το σύνολο των κειμένων που επιστρέφει το σύστημα για αυτήν την ερώτηση χρησιμοποιώντας αποκοπή καταλήξεων (καλύτερη περίπτωση) και τα οποία έχουν χαρακτηριστεί ως σχετικά από τον συγγραφέα. Αυτό έγινε αφενός διότι ήταν αδύνατη η εξέταση 2000 και πλέον κειμένων χωριστά για κάθε μία ερώτηση και αφετέρου διότι με την χρήση stemming επιστρέφεται ο μεγαλύτερος δυνατός αριθμός κειμένων που έχουν κάποια πιθανότητα να είναι σχετικά με την ερώτηση (αυξάνεται η απόκριση του συστήματος).

Το συνολικό γράφημα απόκριση-ακρίβεια προκύπτει από 20 ζεύγη τιμών ($R_{\lambda}^{avg}, P_{\lambda}^{avg}$) όπου η παράμετρος λ αντιπροσωπεύει το πλήθος των ανακτηθέντων κειμένων σε 20 διακεκριμένα σημεία (break-points). Για παράδειγμα όσον αφορά μία ερώτηση, ύστερα από την ανάκτηση ενός κειμένου ($\lambda=1$) υπολογίζεται το ζεύγος (R_1, P_1), ύστερα από την ανάκτηση δύο κειμένων ($\lambda=2$) υπολογίζεται το ζεύγος (R_2, P_2) κ.ο.κ. μέχρι $\lambda=20$ ή μέχρι το πλήθος λ των ανακτηθέντων κειμένων, αν αυτό για κάποια ερώτηση είναι μικρότερο του 20. Σε αυτή την περίπτωση για να μπορέσουμε να υπολογίσουμε τα υπόλοιπα ζεύγη τιμών (R_{λ}, P_{λ}) όπου $\lambda=\kappa+1..20$, θεωρούμε ότι εικονικά επιστράφηκαν κείμενα στις θέσεις $\kappa+1$ ώς 20 αλλά ήταν άσχετα με την ερώτηση. Κατόπιν με βάση τις απαντήσεις σε κάθε ερώτηση υπολογίζονται τα ζεύγη ($R_{\lambda}^{avg}, P_{\lambda}^{avg}$) στα 20 διακεκριμένα σημεία και στην συνέχεια υπολογίζονται με γραμμική παρεμβολή οι τιμές της ακρίβειας ($P_{interpolation}^{0.1}, P_{interpolation}^{0.2}, \dots, P_{interpolation}^{1}$) για τις τιμές 0.1, 0.2, 0.3, ..., 1 της απόκρισης από τις οποίες και προκύπτει το συνολικό γράφημα απόκριση-ακρίβεια.

Παρακάτω δίνουμε το πλήρες σύνολο των 20 ερωτήσεων του πειράματος πριν και μετά την επεξεργασία τους από το σύστημα ανάκτησης.

Μετά την επεξεργασία			
Ερώτηση	Πριν την επεξεργασία	Με αποκοπή καταλήξεων	Χωρίς αποκοπή καταλήξεων
1	ΜΙΣΘΟΛΟΓΙΟ ΙΑΤΡΩΝ	ΜΙΣΘΟΛΟΓ% & ΙΑΤΡ%	ΜΙΣΘΟΛΟΓΙΟ & ΙΑΤΡΩΝ
2	ΠΑΙΔΙΚΕΣ ΑΡΡΩΣΤΙΕΣ	ΠΑΙΔ% & ΑΡΡΩΣΤ%	ΠΑΙΔΙΚΕΣ & ΑΡΡΩΣΤΙΕΣ
3	ΠΛΑΣΤΙΚΗ ΧΕΙΡΟΥΡΓΙΚΗ	ΠΛΑΣΤ% & ΧΕΙΡΟΥΡΓ%	ΠΛΑΣΤΙΚΗ & ΧΕΙΡΟΥΡΓΙΚΗ
4	ΙΑΤΡΙΚΟ ΣΥΝΕΔΡΙΟ ΚΑΙ ΑΝΑΙΣΘΗΣΙΟΛΟΓΙΑ	ΙΑΤΡ% & ΣΥΝΕΔΡ% & ΑΝΑΙΣΘΗΣΙΟΛΟΓ%	ΙΑΤΡΙΚΟ & ΣΥΝΕΔΡΙΟ & ΑΝΑΙΣΘΗΣΙΟΛΟΓΙΑ
5	ΡΙΝΟΠΛΑΣΤΙΚΗ	ΡΙΝΟΠΛΑΣΤ%	ΡΙΝΟΠΛΑΣΤΙΚΗ
6	ΟΡΘΟΠΕΔΙΚΕΣ ΚΛΙΝΙΚΕΣ	ΟΡΘΟΠΕΔ% & ΚΛΙΝ%	ΟΡΘΟΠΕΔΙΚΕΣ & ΚΛΙΝΙΚΕΣ
7	ΕΥΡΩΠΑΪΚΗ ΕΝΩΣΗ ΚΑΙ ΥΓΕΙΑ	ΕΥΡΩΠΑ% & ΕΝΩΣ% & ΥΤΕ%	ΕΥΡΩΠΑΪΚΗ & ΕΝΩΣΗ & ΥΤΕΙΑ
8	ΕΤΑΙΡΕΙΑ ΠΡΟΣΤΑΣΙΑΣ ΚΑΙ ΑΠΟΚΑΤΑΣΤΑΣΗΣ ΛΑΝΔΗΡΩΝ	ΕΤΑΙΡ% & ΠΡΟΣΤΑΣ% & ΑΠΟΚΑΤΑΣΤ% & ΑΝΑΠΗΡ%	ΕΤΑΙΡΕΙΑ & ΠΡΟΣΤΑΣΙΑΣ & ΑΠΟΚΑΤΑΣΤΑΣΗΣ & ΛΑΝΔΗΡΩΝ
9	ΟΡΚΟΣ ΤΟΥ ΙΠΠΟΚΡΑΤΗ	ΟΡΚ% & ΙΠΠΟΚΡ%	ΟΡΚΟΣ & ΙΠΠΟΚΡΑΤΗ
10	ΕΓΚΕΦΑΛΙΚΟ ΕΠΕΙΣΟΔΙΟ ΚΑΙ ΑΡΤΗΡΙΑΚΗ ΠΙΕΣΗ	ΕΓΚΕΦΑΛ% & ΕΠΕΙΣΟΔ% & ΑΡΤΗΡ% & ΠΙΕΣ%	ΕΓΚΕΦΑΛΙΚΟ & ΕΠΕΙΣΟΔΙΟ & ΑΡΤΗΡΙΑΚΗ & ΠΙΕΣΗ
11	ΓΑΣΤΡΕΝΤΕΡΙΚΕΣ ΛΟΙΜΩΞΕΙΣ	ΓΑΣΤΡΕΝΤΕΡ% & ΛΟΙΜΩΞ%	ΓΑΣΤΡΕΝΤΕΡΙΚΕΣ & ΛΟΙΜΩΞΕΙΣ
12	ΑΠΟΣΤΟΛΕΣ ΤΩΝ ΓΙΑΤΡΩΝ ΤΟΥ ΚΟΣΜΟΥ	ΑΠΟΣΤΟΛ% & ΓΙΑΤΡ% & ΚΟΣΜ%	ΑΠΟΣΤΟΛΕΣ & ΓΙΑΤΡΩΝ & ΚΟΣΜΟΥ

13	ΠΑΡΕΝΕΡΓΙΕΣ ΕΜΒΟΛΙΩΝ	ΠΑΡΕΝΕΡΓ% & ΕΜΒΟΛ%	ΠΑΡΕΝΕΡΓΙΕΙΣ & ΕΜΒΟΛΙΩΝ
14	ΜΗΝΙΓΓΙΤΙΔΑ ΚΑΙ ΘΕΡΑΠΕΙΑ	ΜΗΝΙΓΓ% & ΘΕΡΑΠ%	ΜΗΝΙΓΓΙΤΙΔΑ & ΘΕΡΑΠΕΙΑ
15	ΑΙΜΑΤΟΛΟΓΙΚΟΣ ΑΝΑΛΥΤΗΣ	ΑΙΜΑΤΟΛΟΦ% & ΑΝΑΛΥΤ%	ΑΙΜΑΤΟΛΟΓΙΚΟΣ & ΑΝΑΛΥΤΗΣ
16	ΙΑΤΡΙΚΑ ΤΜΗΜΑΤΑ ΑΣΚΛΗΠΙΕΙΟΥ ΒΟΥΛΑΣ	ΙΑΤΡ% & ΤΜΗΜ% & ΑΣΚΛΗΠ% & ΒΟΥΛ%	ΙΑΤΡΙΚΑ & ΤΜΗΜΑΤΑ & ΑΣΚΛΗΠΙΕΙΟΥ & ΒΟΥΛΑΣ
17	ΕΠΙΣΤΗΜΟΝΙΚΟ ΠΡΟΣΩΠΙΚΟ ΑΣΚΛΗΠΙΕΙΟΥ ΒΟΥΛΑΣ	ΕΠΙΣΤΗΜΟΝ% & ΠΡΟΣΩΠ% & ΑΣΚΛΗΠ% & ΒΟΥΛ%	ΕΠΙΣΤΗΜΟΝΙΚΟ & ΠΡΟΣΩΠΙΚΟ & ΑΣΚΛΗΠΙΕΙΟΥ & ΒΟΥΛΑΣ
18	ΟΜΟΙΟΠΑΘΗΤΙΚΗ ΙΑΤΡΙΚΗ	ΟΜΟΙΟΠ% & ΙΑΤΡ%	ΟΜΟΙΟΠΑΘΗΤΙΚΗ & ΙΑΤΡΙΚΗ
19	ΙΣΧΑΙΜΙΚΑ ΕΓΚΕΦΑΛΙΚΑ ΕΠΕΙΣΟΔΙΑ	ΙΣΧΑΙΜ% & ΕΓΚΕΦΑΛ% & ΕΠΕΙΣΟΔ%	ΙΣΧΑΙΜΙΚΑ & ΕΓΚΕΦΑΛΙΚΑ & ΕΠΕΙΣΟΔΙΑ
20	ΣΥΜΠΤΩΜΑΤΑ ΛΑΡΥΓΓΙΤΙΔΑΣ	ΣΥΜΠΤΩΜ% & ΛΑΡΥΓΓ%	ΣΥΜΠΤΩΜΑΤΑ & ΛΑΡΥΓΓΙΤΙΔΑΣ

Ενδεικτικά στις επόμενες παραγράφους παρουσιάζουμε τα αποτελέσματα που έδωσε το σύστημα ανάκτησης για 5 από τις παραπάνω ερωτήσεις με και χωρίς αποκοπή καταλήξεων στους όρους της ερώτησης. Θα πρέπει πάντως να πούμε ότι οι τελικές τιμές της απόκρισης και της ακρίβειας και οι αντίστοιχοι πίνακες αποτελεσμάτων προέκιναν από όλο το σύνολο των 20 ερωτήσεων.

5.1.1.1 Αποτελέσματα με αποκοπή καταλήξεων στους όρους

ΕΡΩΤΗΣΗ 1: ΜΙΣΘΟΛΟΓΙΤΟ ΙΑΤΡΩΝ ΜΙΣΘΟΛΟΓ% & ΙΑΤΡ%						
ΒΡΕΘΗΚΑΝ:	3					
ΣΧΕΤΙΚΕΣ:	3					
DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια	
500	Σύλλογος Ειδικευουμένων Ιατρών Αθηνών-Πειραιώς (Σ.Ε.Ι.Α.Π.) - Νέα	100	ΝΑΙ	0.3333	1.0000	
501	Σύλλογος Ειδικευουμένων Ιατρών Αθηνών-Πειραιώς (Σ.Ε.Ι.Α.Π.) - Απόψεις Μελών	83	ΝΑΙ	0.6667	1.0000	
499	Σύνδεσμος Εκπαιδευούμενών Ιατρών Αθηνών-Πειραιώς	17	ΝΑΙ	1.0000	1.0000	

ΕΡΩΤΗΣΗ 2: ΠΑΙΔΙΚΕΣ ΑΡΡΩΣΤΙΕΣ ΠΑΙΔ% & ΑΡΡΩΣΤ%						
ΒΡΕΘΗΚΑΝ:	21					
ΣΧΕΤΙΚΕΣ:	15					
DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια	
540	αι έξαφνα, στην ώρα που ξεντύνετε το παιδί, παρατηρείτε στην πλάτη του ένα μεγάλο στυρί γεμάτο υγρά,	100	ΝΑΙ	0.0667	1.0000	
521	ΠΡΟΕΤΟΙΜΑΣΤΕΙΤΕ ΓΙΑ ΤΟ ΦΘΙΝΟΠΩΡΟ ΚΑΙ ΤΟΝ ΧΕΙΜΩΝΑ	60	ΝΑΙ	0.1333	1.0000	
545	news1	60	ΝΑΙ	0.2000	1.0000	

527	Vaccinations	40	ΟΧΙ	0.2000	0.7500
528	Μαντού	40	ΝΑΙ	0.2667	0.8000
529	fever1	30	ΝΑΙ	0.3333	0.8333
536	UnTitled	30	ΝΑΙ	0.4000	0.8571
538	ΒΡΟΓΧΙΟΛΙΤΙΣ	30	ΝΑΙ	0.4667	0.8750
537	Φαρυγγίτιδα, Αμυγδαλίτιδα, ρινίτιδα	30	ΝΑΙ	0.5333	0.8889
557	babywalk	30	ΝΑΙ	0.6000	0.9000
535	meningitis	30	ΝΑΙ	0.6667	0.9091
541	ΓΑΣΤΡΕΝΤΕΡΙΚΕΣ ΛΟΙΜΩΞΕΙΣ	20	ΝΑΙ	0.7333	0.9167
551	fddirctions	20	ΝΑΙ	0.8000	0.9231
526	newborn	10	ΝΑΙ	0.8667	0.9286
941	προαναισθητική αξιολόγηση χειρουργικού ασθενούς	10	ΟΧΙ	0.8667	0.8667
1166	προαναισθητική αξιολόγηση χειρουργικού ασθενούς	10	ΟΧΙ	0.8667	0.8125
849	peripitosi18	10	ΟΧΙ	0.8667	0.7647
549	Μπορεί ο τραυλισμός να "θεραπευτεί"	10	ΝΑΙ	0.9333	0.7778
531	accidents	10	ΟΧΙ	0.9333	0.7368
563	breastfeed	10	ΟΧΙ	0.9333	0.7000
543	eczema	10	ΝΑΙ	1.0000	0.7143

**ΕΡΩΤΗΣΗ 3: ΠΛΑΣΤΙΚΗ ΧΕΙΡΟΥΡΓΙΚΗ
ΠΛΑΣΤ% & ΧΕΙΡΟΥΡΓ%***

ΒΡΕΘΗΚΑΝ: 11

ΣΧΕΤΙΚΕΣ: 4

DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια
383	UnTitled	100	ΝΑΙ	0.2500	1.0000
1076	orthopediki kliniki ika	57	ΟΧΙ	0.2500	0.5000
931	prospelaseis	29	ΟΧΙ	0.2500	0.3333
353	NEWS	14	ΟΧΙ	0.2500	0.2500
484	Ποιές είναι οι αποστολές των γιατρών του κόσμου;	14	ΟΧΙ	0.2500	0.2000
939	epilegmena keimena - HIV	14	ΝΑΙ	0.5000	0.3333
1168	shunt	14	ΟΧΙ	0.5000	0.2857
1054	mpisxiniotis ioannis ergasies	14	ΟΧΙ	0.5000	0.2500
826	file:///UnTitled	14	ΟΧΙ	0.5000	0.2222
816	orthopediki kliniki Ag.paulos	14	ΝΑΙ	0.7500	0.3000
370	Asclepeion Hospital - ORL Department	14	ΝΑΙ	1.0000	0.3636

**ΕΡΩΤΗΣΗ 12: ΑΠΟΣΤΟΛΕΣ ΤΩΝ ΓΙΑΤΡΩΝ ΤΟΥ ΚΟΣΜΟΥ
ΑΠΟΣΤΟΛ% & ΓΙΑΤΡ% & ΚΟΣΜ%**

ΒΡΕΘΗΚΑΝ: 9

ΣΧΕΤΙΚΕΣ: 9

DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια
484	Ποιές είναι οι αποστολές των γιατρών του κόσμου;	100	ΝΑΙ	0.1111	1.0000
486	Πώς μπορώ να συμμετέχω και εγώ σε μια αποστολή των γιατρών του κόσμου;	60	ΝΑΙ	0.2222	1.0000
481	Ποιές είναι οι αργές των γιατρών του κόσμου;	20	ΝΑΙ	0.3333	1.0000
482	Από πού προέρχονται οι πόροι των γιατρών του κόσμου;	20	ΝΑΙ	0.4444	1.0000
485	Υπάρχει ασφάλεια σ' αυτές τις αποστολές;	20	ΝΑΙ	0.5556	1.0000
483	Ένα διοικητικό σχήμα διάφανο και	20	ΝΑΙ	0.6667	1.0000

	αποτελεσματικό.				
479	MDM ! Welcomc !	10	ΝΑΙ	0.7778	1.0000
495	Ποιές είναι οι αποστολές των γιατρών του κόσμου;	10	ΝΑΙ	0.8889	1.0000
497	Πως μπορώ να συμμετέχω και εγώ σε μία αποστολή των γιατρών του κόσμου;	10	ΝΑΙ	1.0000	1.0000

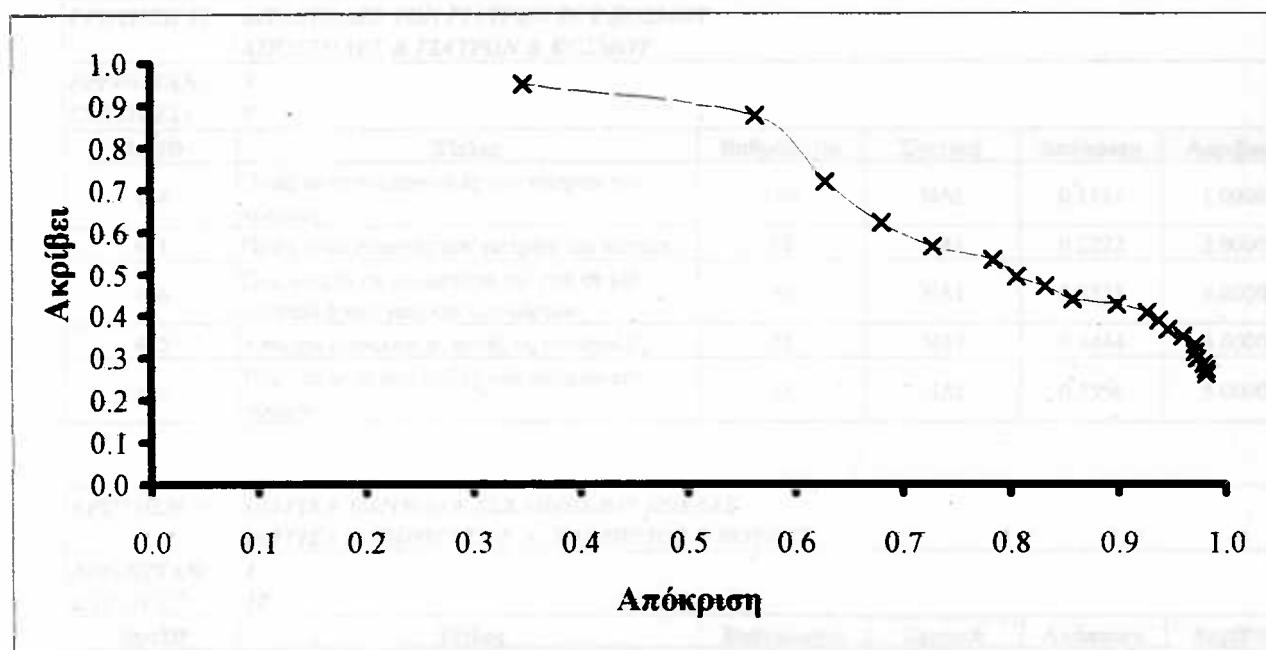
**ΕΡΩΤΗΣΗ 16: ΙΑΤΡΙΚΑ ΤΜΗΜΑΤΑ ΑΣΚΛΗΠΙΕΙΟΥ ΒΟΥΛΑΣ
ΙΑΤΡ% & ΤΜΗΜ% & ΑΣΚΛΗΠ% & ΒΟΥΛ%**

ΒΡΕΘΗΚΑΝ: 15**ΣΧΕΤΙΚΕΣ:** 10

DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια
363	Asclepeion Hospital	100	ΝΑΙ	0.1000	1.0000
361	Asclepeion Hospital, Athens - Greece	40	ΝΑΙ	0.2000	1.0000
381	Asclepeion Hospital-ORL DEP-	40	ΟΧΙ	0.2000	0.6667
364	Asclepeion Hospital - Department of Orthopaedics	20	ΝΑΙ	0.3000	0.7500
383	UnTitled	20	ΝΑΙ	0.4000	0.8000
784	Asclepeion Hospital - Department of Orthopaedics	20	ΝΑΙ	0.5000	0.8333
991	prospelaseis	20	ΟΧΙ	0.5000	0.7143
1086	Asclepeion Hospital - Department of Orthopaedics	20	ΝΑΙ	0.6000	0.7500
1107	file:///UnTitled	20	ΟΧΙ	0.6000	0.6667
1009	imerida3	20	ΟΧΙ	0.6000	0.6000
975	Asclepeion Hospital - Department of Orthopaedics	20	ΝΑΙ	0.7000	0.6364
727	arxeio synedrion	20	ΟΧΙ	0.7000	0.5833
368	Asclepeion Hospital - Department of Orthopaedics	20	ΝΑΙ	0.8000	0.6154
370	Asclepeion Hospital - ORL Department	20	ΝΑΙ	0.9000	0.6429
379	Asclepeion Hospital - Department of Orthopaedics	20	ΝΑΙ	1.0000	0.6667

Με βάση τα παραπάνω αποτελέσματα υπολογίζονται τα παρακάτω 20 ζεύγη τιμών ($R_{\lambda}^{avg}, P_{\lambda}^{avg}$) από τα οποία προκύπτει το γράφημα του σχήματος 5.4.

Break-point (λ)	Απόκριση ^{Avg}	Ακρίβεια ^{Avg}
1	0.3476	0.9500
2	0.5633	0.8750
3	0.6281	0.7167
4	0.6807	0.6208
5	0.7283	0.5650
6	0.7839	0.5333
7	0.8054	0.4929
8	0.8319	0.4688
9	0.8579	0.4389
10	0.8988	0.4250
11	0.9267	0.4091
12	0.9381	0.3875
13	0.9464	0.3654
14	0.9603	0.3500
15	0.9708	0.3333
16	0.9708	0.3125
17	0.9733	0.2971
18	0.9792	0.2861
19	0.9817	0.2737
20	0.9817	0.2600



Σχήμα 5.4: Γράφημα Απόκριση-Ακρίβεια με αποκοπή καταλήξεων στους όρους

5.1.1.2 Αποτελέσματα χωρίς αποκοπή καταλήξεων στους όρους

ΕΡΩΤΗΣΗ 1: ΜΙΣΘΟΛΟΓΙΟ ΙΑΤΡΩΝ ΜΙΣΘΟΛΟΓΙΟ & ΙΑΤΡΩΝ						
ΒΡΕΘΗΚΑΝ: 3						
ΣΧΕΤΙΚΕΣ: 3						
DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια	
500	Σύλλογος Ειδικευομένων Ιατρών Αθηνών-Πειραιώς (Σ.Ε.Ι.Α.Π.) - Νέα	100	ΝΑΙ	0.3333	1.0000	
499	Σύνδεσμος Εκπαιδευομένων Ιατρών Αθηνών-Πειραιώς	50	ΝΑΙ	0.6667	1.0000	
501	Σύλλογος Ειδικευομένων Ιατρών Αθηνών-Πειραιώς (Σ.Ε.Ι.Α.Π.) - Απόψεις Μελών	50	ΝΑΙ	1.0000	1.0000	

ΕΡΩΤΗΣΗ 2: ΠΑΙΔΙΚΕΣ ΑΡΡΩΣΤΙΕΣ ΠΑΙΔΙΚΕΣ & ΑΡΡΩΣΤΙΕΣ						
ΒΡΕΘΗΚΑΝ: 0						
ΣΧΕΤΙΚΕΣ: 15						

ΕΡΩΤΗΣΗ 3: ΠΛΑΣΤΙΚΗ ΧΕΙΡΟΥΡΓΙΚΗ ΠΛΑΣΤΙΚΗ & ΧΕΙΡΟΥΡΓΙΚΗ						
ΒΡΕΘΗΚΑΝ: 1						
ΣΧΕΤΙΚΕΣ: 4						
DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια	
383	UnTitled	100	ΝΑΙ	0.2500	1.0000	

**ΕΡΩΤΗΣΗ 12: ΑΠΟΣΤΟΛΕΣ ΤΩΝ ΓΙΑΤΡΩΝ ΤΟΥ ΚΟΣΜΟΥ
ΑΠΟΣΤΟΛΕΣ & ΓΙΑΤΡΩΝ & ΚΟΣΜΟΥ**
ΒΡΕΘΗΚΑΝ: 5**ΣΧΕΤΙΚΕΣ:** 9

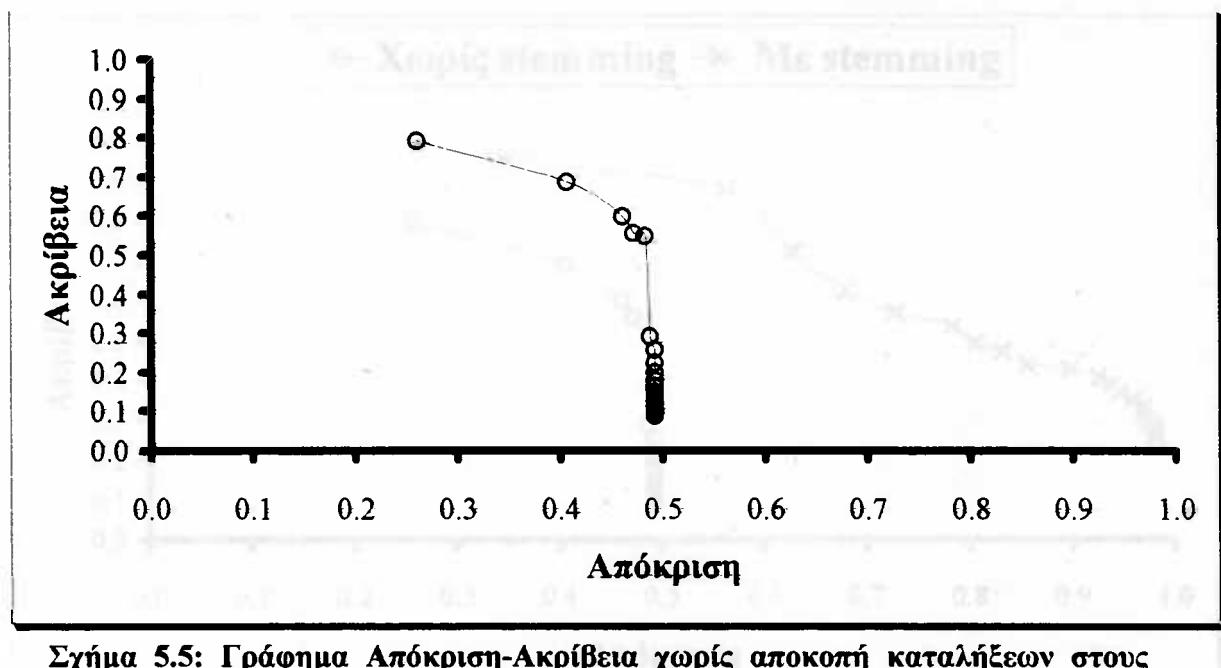
DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια
484	Ποιές είναι οι αποστολές των γιατρών του κόσμου;	100	ΝΑΙ	0.1111	1.0000
481	Ποιές είναι οι αρχές των γιατρών του κόσμου;	50	ΝΑΙ	0.2222	2.0000
486	Πώς μπορώ να συμμετέχω και εγώ σε μία αποστολή των γιατρών του κόσμου;	50	ΝΑΙ	0.3333	3.0000
485	Υπάρχει ασφάλεια σ' αυτές τις αποστολές;	25	ΝΑΙ	0.4444	4.0000
495	Ποιές είναι οι αποστολές των γιατρών του κόσμου;	25	ΝΑΙ	0.5556	5.0000

**ΕΡΩΤΗΣΗ 16: ΙΑΤΡΙΚΑ ΤΜΗΜΑΤΑ ΑΣΚΛΗΠΙΕΙΟΥ ΒΟΥΛΑΣ
ΙΑΤΡΙΚΑ & ΤΜΗΜΑΤΑ & ΑΣΚΛΗΠΙΕΙΟΥ & ΒΟΥΛΑΣ**
ΒΡΕΘΗΚΑΝ: 1**ΣΧΕΤΙΚΕΣ:** 10

DocID	Τίτλος	Βαθμολογία	Σχετική	Απόκριση	Ακρίβεια
363	Asclepeion Hospital	100	ΝΑΙ	0.1000	1.0000

Με βάση τα παραπάνω αποτελέσματα υπολογίζονται τα παρακάτω 20 ζεύγη τιμών (R_i^{avg}, P_i^{avg}) από τα οποία προκύπτει το γράφημα του σχήματος 5.5.

Break-point (λ)	Απόκριση ^{Avg}	Ακρίβεια ^{Avg}
1	0.2603	0.7895
2	0.4069	0.6842
3	0.4614	0.5965
4	0.4721	0.5526
5	0.4827	0.5474
6	0.4875	0.2895
7	0.4923	0.2556
8	0.4923	0.2237
9	0.4923	0.1988
10	0.4923	0.1789
11	0.4923	0.1627
12	0.4923	0.1491
13	0.4923	0.1377
14	0.4923	0.1278
15	0.4923	0.1193
16	0.4923	0.1118
17	0.4923	0.1053
18	0.4923	0.0994
19	0.4923	0.0942
20	0.4923	0.0895



Σχήμα 5.5: Γράφημα Απόκριση-Ακρίβεια χωρίς αποκοπή καταλήξεων στους όρους

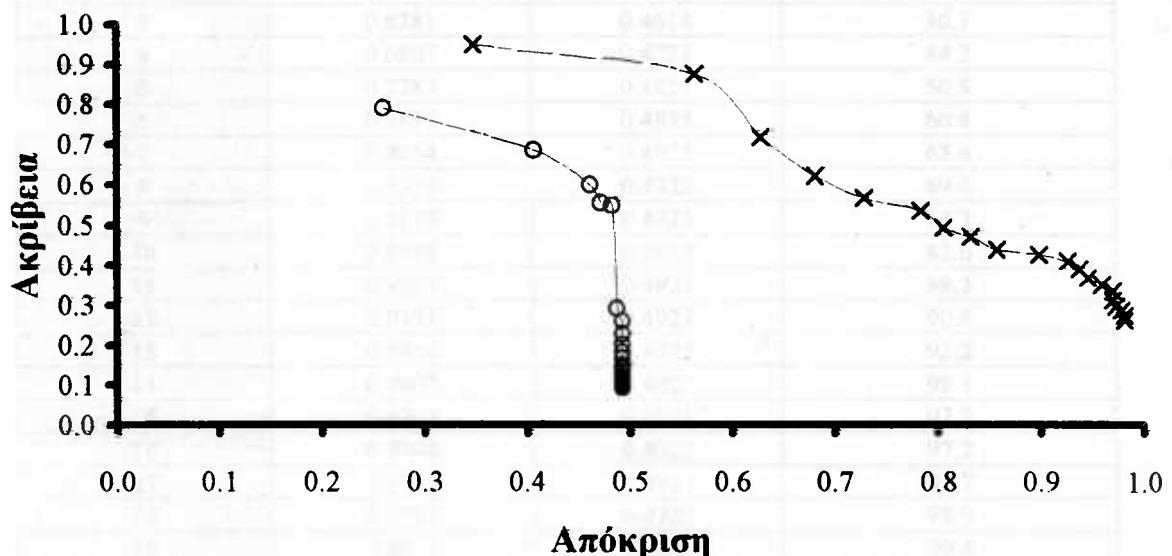
5.1.1.3 Συγκριτικά Αποτελέσματα - Συμπεράσματα

Χρησιμοποιώντας τα παραπάνω αποτελέσματα μπορούμε να κατασκευάσουμε το συγκριτικό γράφημα του σχήματος 5.6 που περιέχει τα γραφήματα Απόκριση-Ακρίβεια με και χωρίς αποκοπή καταλήξεων στους όρους.

Η επόμενη σελίδα παρουσιάζει την απόκριση και την ακρίβεια της διεκπεργμένης απόκρισης. Η πλευρά της χάρτης της απόκρισης σέρνεται κατά την αύξηση της ακρίβειας, ενώ πλευρά της ακρίβειας σέρνεται κατά την αύξηση της απόκρισης. Οι δύο σειρές δείχνουν την απόκριση ανάλογη στην ακρίβεια, αλλά με διαφορετικό τρόπο. Η σειρά με αποκοπή καταλήξεων στους όρους παρουσιάζει μια μεγαλύτερη απόκριση σε όρους με χαμηλή ακρίβεια, αλλά μια μικρότερη απόκριση σε όρους με ψηλή ακρίβεια.

Τα παραπάνω αποτελέσματα φαίνονται και σταύρωσης 5.7 και 5.8 με δύο σειρές απόκρισης σε μέτρια, της απόκρισης και της ακρίβειας, την απόκρισης απόκρισης σε μεγαλοκανόνες και μεγαλοκανόνες απόκρισης:

⊖ Χωρίς stemming ✖ Με stemming



Σχήμα 5.6: Συγκριτικό γράφημα Απόκριση-Ακρίβεια

Στο σχήμα 5.6 φαίνεται καθαρά μία βελτίωση των αποτελεσμάτων και σε ακρίβεια και σε απόκριση όταν χρησιμοποιούμε αποκοπή καταλήξεων στους όρους της ερώτησης. Η αποκοπή καταλήξεων της Ελληνικής όπως ήταν αναμενόμενο είχε ως αποτέλεσμα την αύξηση των αριθμού των ανακτηθέντων κειμένων για κάθε ερώτηση το οποίο παρουσιάζεται με μία αύξηση της μέσης απόκρισης R_{λ}^{avg} στα διακεκριμένα σημεία λ . Η αύξηση της τιμής της απόκρισης είχε και ως αποτέλεσμα και την αύξηση της μέσης ακρίβειας P_{λ}^{avg} και αυτό ερμηνεύεται από το γεγονός ότι χωρίς την αποκοπή καταλήξεων το σύστημα ανάκτησης σε πολλά ερωτήματα δεν επέστρεφε καθόλου κείμενα (δηλ. πλήθος ανακτηθέντων κειμένων=0) οπότε και η τιμή της ακρίβειας στα διακεκριμένα σημεία λ ήταν 0.0 πράγμα που μείωνε την τιμή της μέσης ακρίβειας P_{λ}^{avg} .

Τα παραπάνω συμπεράσματα φαίνονται και στους πίνακες 5.7 και 5.8 οι οποίοι δείχνουν την βελτίωση της απόκρισης και της ακρίβειας του συστήματος ανάκτησης όταν χρησιμοποιούμε αποκοπή καταλήξεων στους όρους.

Break-point (λ)	Απόκριση ^{Aνε}		Βελτίωση επί τοις εκατό (%)
	Με αποκοπή καταλήξεων	Χωρίς αποκοπή καταλήξεων	
1	0.3476	0.2603	33.5
2	0.5633	0.4069	38.4
3	0.6281	0.4614	36.1
4	0.6807	0.4721	44.2
5	0.7283	0.4827	50.9
6	0.7839	0.4875	60.8
7	0.8054	0.4923	63.6
8	0.8319	0.4923	69.0
9	0.8579	0.4923	74.3
10	0.8988	0.4923	82.6
11	0.9267	0.4923	88.2
12	0.9381	0.4923	90.6
13	0.9464	0.4923	92.2
14	0.9603	0.4923	95.1
15	0.9708	0.4923	97.2
16	0.9708	0.4923	97.2
17	0.9733	0.4923	97.7
18	0.9792	0.4923	98.9
19	0.9817	0.4923	99.4
20	0.9817	0.4923	99.4

Πίνακα 5.7: Βελτίωση της απόκρισης

Break-point (λ)	Ακρίβεια ^{Aνε}		Βελτίωση επί τοις εκατό (%)
	Με αποκοπή καταλήξεων	Χωρίς αποκοπή καταλήξεων	
1	0.9500	0.7895	20.3%
2	0.8750	0.6842	27.9
3	0.7167	0.5965	20.2
4	0.6208	0.5526	12.3
5	0.5650	0.5474	3.2
6	0.5333	0.2895	84.2
7	0.4929	0.2556	92.8
8	0.4688	0.2237	109.6
9	0.4389	0.1988	120.8
10	0.4250	0.1789	137.6
11	0.4091	0.1627	151.4
12	0.3875	0.1491	159.9
13	0.3654	0.1377	165.4
14	0.3500	0.1278	173.9
15	0.3333	0.1193	179.4
16	0.3125	0.1118	179.5
17	0.2971	0.1053	182.1
18	0.2861	0.0994	187.8
19	0.2737	0.0942	190.6
20	0.2600	0.0895	190.5

Πίνακα 5.8: Βελτίωση της ακρίβειας

Για λόγους πληρότητας της αξιολόγησης των αποτελεσμάτων του συστήματος ανάκτησης, θα μπορούσαμε να συγκρίνουμε τα παραπάνω αποτελέσματα με τα αντίστοιχα αποτελέσματα που δίνονται ότι μπορούν και υποστηρίζουν αναζήτησεις σε διάφορες γλώσσες. Βέβαια τα αποτελέσματα που επιστρέφουν δεν είναι άμεσα συγκρίσιμα με αυτά που παρουσιάσαμε παραπάνω για δύο κυρίως λόγους:

1. Η συλλογή των κειμένων των μηχανών αναζήτησης είναι τεράστια σε μέγεθος και πιθανών να καλύπτει όλα τα αποθέματα που υπάρχουν στον Ελληνικό χώρο πλήθους εκατομμυρίων σελίδων και όχι μόνο μερικών χιλιάδων όπως συμβαίνει με τη συλλογή του συστήματος ανάκτησης.
2. Η ενημέρωση της συλλογής των μεγάλων μηχανών αναζήτησης γίνεται κατά αραιότερα διαστήματα λόγου του μεγάλου μεγέθους της. Ετσι οι σελίδες που υπάρχουν στην συλλογή του συστήματος ανάκτησης μπορεί να μην έχουν ακόμα ανακαλυφθεί και δεικτοδοτηθεί από τις μηχανές αναζήτησης διότι είναι καινούργιες ή να έχουν δεικτοδοτηθεί παλαιότερες εκδόσεις αυτών.

Εχοντας υπόψη μας τις παραπάνω παρατηρήσεις, θέσαμε ενδεικτικά στην μηχανή αναζήτησης AltaVista ([URL:<http://www.altavista.digital.com>](http://www.altavista.digital.com)) μερικές από τις παραπάνω Ελληνικές ερωτήσεις και πήραμε τα ακόλουθα αποτελέσματα:

* Ο τελεστής "&" στην AltaVista όπως και στο ConText αντιπροσωπεύει το λογικό τελεστή AND.

Ερώτηση 1: μισθολόγιο & ιατρών

1 documents match your query

1. MedNet Hellas: Συλλογή Νομοθεσίας

VII. Δημόσιοι φορείς παροχής αιτρικών υπηρεσιών Εθνικό Σύστημα Υγείας. Εθνικό Σύστημα Υγείας - Υπηρεσιακή κατάσταση γιατρών κλάδου ΕΣΥ - Πειθαρχικά...

<http://www.mednet.gr/law/law7.htm> - size 145K - 11-Mar-97 - Greek

Ερώτηση 2: μισθολόγιο & ιατρού

No documents match the query.

Ερώτηση 3: αποστολές & γιατρόν & κόσμου

4 documents match your query.

1. Macedonian Press Agency: News in Greek, 97-06-01

Macedonian Press Agency: News in Greek, 97-06-01. Macedonian Press Agency: News in Elot928 Greek Directory - Previous Article - Next Article. From: The...



<http://www.hri.org/news/greek/mpegr/97-06-01.mpegr.html> - size 98K - 9-Sep-97
- Greek

2. Πως μπορώ να συμμετέχω και εγώ σε μία αποστολή των γιατρών του κόσμου;

Οι ΓΙΑΤΡΟΙ ΤΟΥ ΚΟΣΜΟΥ δέχονται στις αποστολές τους, γιατρούς, νοσηλευτές, άλλους υγειονομικούς, τεχνικούς και διοικητικούς υπαλλήλους, ανάλογα με το είδος.

<http://www.istos.net.gr/mdm/symetoxh.htm> - size 3K - 18-Dec-96 - Greek

3. Ποιές είναι οι αργές των γιατρών του κόσμου;

Σημαντικό συστατικό της δράσης των ΓΙΑΤΡΩΝ ΤΟΥ ΚΟΣΜΟΥ είναι η ΜΑΡΤΥΡΙΑ: Οι ΓΙΑΤΡΟΙ ΤΟΥ ΚΟΣΜΟΥ δεν είναι απλοί παραπηρητές και σιωπηλοί θεραπευτές του...

<http://www.istos.net.gr/mdm/arxes.htm> - size 4K - 7-Oct-96 - Greek

4. Ποιές είναι οι αποστολές των γιατρών του κόσμου;

Οι αποστολές των ΓΙΑΤΡΩΝ ΤΟΥ ΚΟΣΜΟΥ είναι τριών ειδών: Α) Αποστολές Διερεύνησης. Ι' ίνονται με στόχο να διαπιστωθούν επί τόπου οι πραγματικές συνθήκες, να...

<http://www.istos.net.gr/mdm/apostole.htm> - size 5K - 7-Oct-96 - Greek

Ερώτηση 4: αποστολές & γιατρών & κόσμων

No documents match the query.

Ερώτηση 5 : ιατρικά & τμήματα & ασκληπίειον & βούλας

1 documents match your query.

1. Asclepeion Hospital

ΓΕΝΙΚΟ ΠΕΡΙΦΕΡΙΑΚΟ ΝΟΣΟΚΟΜΕΙΟ "ΑΣΚΛΗΠΙΕΙΟ ΒΟΥΛΑΣ" ΑΘΗΝΑ. Το "Ασκληπιείο Βούλας" ιδρύθηκε από τον Ελληνικό Ερυθρό Σταυρό το 1933 με κύριο σκοπό την...

<http://www.forthnet.gr/asclepeion/asklepgr.htm> - size 3K - 27-Oct-96 - Greek

Στα παραπάνω αποτελέσματα, τα κείμενα που βρίσκονται μέσα σε διακεκομμένο πλαίσιο είναι τα ίδια που έχουν επιστραφεί και από το σύστημα ανάκτησης για την ίδια ερώτηση. Οι όροι των ερωτήσεων συνδυάστηκαν με τον λογικό τελεστή AND όπως κάναμε και στις ερωτήσεις του συστήματος ανάκτησης ενώ σε μερικές ερωτήσεις (βλ. ερώτηση 3) αφαιρέσαμε χειρονακτικά τις τετριμμένες λέξεις. Οσον αφορά τα ανακτηθέντα κείμενα, τα συμπεράσματα είναι τα εξής:

Ερώτηση 1: Ανακτήθηκε μόνο ένα κείμενο και αυτό δεν περιέχεται στα κείμενα που επέστρεψε το σύστημα ανάκτησης για την ίδια ερώτηση. Πιθανών η AltaVista να μην έχει δεικτοδοτήσει τις σελίδες που επιστρέφει το σύστημα ανάκτησης ενώ σίγουρα δεν περιέχεται στη συλλογή του συστήματος ανάκτησης η σελίδα που επέστρεψε η AltaVista.

Ερώτηση 2: Η ερώτηση 2 διαφέρει από την ερώτηση 1 μόνο στο δεύτερο όρο ("ιατρού" αντί "ιατρών"). Δεν ανακτήθηκε κανένα κείμενο πράγμα που σημαίνει ότι δεν γίνεται αποκοπή καταλήξεων στους όρους διότι αν συνέβαινε αυτό θα επιστρεφόταν τουλάχιστον το κείμενο που ανακτήθηκε στην ερώτηση 1.

Ερώτηση 3: Ανακτήθηκαν τέσσερα κείμενα από τα οποία τα τρία υπάρχουν στα κείμενα που επέστρεψε το σύστημα ανάκτησης για την ίδια ερώτηση. Ενδεικτικά αναφέρουμε ότι το σύστημα ανάκτησης επέστρεψε 9 κείμενα με αποκοπή καταλήξεων και 5 κείμενα χωρίς αποκοπή καταλήξεων. Το πρώτο κείμενο που επέστρεψε η AltaVista δεν περιέχεται στη συλλογή του συστήματος ανάκτησης.

Ερώτηση 4: Η ερώτηση 4 διαφέρει από την ερώτηση 3 μόνο στο τελευταίο όρο ("κόσμων" αντί "κόσμου"). Δεν ανακτήθηκε κανένα κείμενο πράγμα που σημαίνει ότι δεν γίνεται αποκοπή καταλήξεων στους όρους διότι αν συνέβαινε αυτό θα επιστρέφονταν τουλάχιστον τα κείμενα που ανακτήθηκαν στην ερώτηση 3.

Ερώτηση 5: Τα αποτελέσματα που έδωσε η AltaVista είναι τα ίδια ακριβώς με αυτά που έδωσε το σύστημα ανάκτησης χωρίς αποκοπή καταλήξεων.

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα

6.1 Πλεονεκτήματα / Μειονεκτήματα του ConText

Η υλοποίηση του συστήματος ανάκτησης με την χρήση του Oracle ConText έδειξε ότι αυτό το προϊόν μπορεί να ανταποκριθεί με άριστες επιδόσεις στην ανάγκη για αποθήκευση και ανάκτηση μικρού έως μεσαίου όγκου πληροφοριών. Για μεγάλο όγκο πληροφορίας οι συνθήκες και οι παράμετροι ενδεχομένως να αλλάζουν - χωρίς να έχουμε όμως πειραματική απόδειξη για αυτό - στηριζόμενη στο γεγονός ότι για μεγάλου μεγέθους συλλογές, ο χώρος που απαιτείται για την αποθήκευση των πινάκων δεικτοδότησης με τους όρους στην πλήρη τους μορφή χωρίς αποκοπή των καταλήξεων τους θα είναι τεράστιος, με άμεση συνέπεια την αύξηση του χρόνου αναζήτησης. Παρόλα αυτά το Oracle ConText χρησιμοποιείται ήδη και λειτουργεί σε εκατοντάδες εφαρμογές ανάκτησης πληροφοριών, ενώ οι μελλοντικές του εκδόσεις αναμένεται να έχουν ακόμη μεγαλύτερες ευκολίες και δυνατότητες ανάκτησης. Χαρακτηριστικά αναφέρουμε ότι μεγάλες ευρωπαϊκές εταιρίες παραγωγής λογισμικού όπως Webdevelopment, Q-Ray, VDA, Digital Collections, PROLIN, Scopus, TIBCO, Executech, Global Software Consultants, COM.sortium, SRA, Highland Technology, Mitratech και δεκάδες άλλες, χρησιμοποιούν το ConText σε κάθε εφαρμογή ανάκτησης πληροφοριών που αναπτύσσουν είτε πρόκειται για stand-alone είτε για intranet/internet εφαρμογή.

Το μεγάλο πλεονέκτημα από την χρησιμοποίηση του Oracle ConText είναι η ενοποίηση των λειτουργιών ενός RDBMS και ενός Full Text Retrieval συστήματος, και η ταυτόχρονη λειτουργία τους σε μοντέλο Πελάτη-Εξυπηρετητή (Client-Server) που σε σύγκριση με τα παραδοσιακά IRS αποτελεί μια καινοτομία. Το κάθε σύστημα διατηρεί συναλλοίωτα τα ιδιαίτερα χαρακτηριστικά και δυνατότητες του - παρέχονται οι κλασσικές λειτουργίες χειρισμού των δεδομένων που συναντάμε στα RDBMS όπως INSERT, UPDATE, DELETE κ.α. μαζί με εξειδικευμένες λειτουργίες που συναντάμε στα IRS, ξένες σε συστήματα RDBMS, όπως αποκοπή καταλήξεων, τετριψμένες λέξεις, ανεστραμμένη λίστα όρων κ.α. - ενώ παρέχεται μια κοινή διεπαφή προγραμματισμού και περιβάλλον λειτουργίας τους, διαμέσου της καθιερωμένης γλώσσας PL/SQL και επεκτάσεών της (έτοιμες βιβλιοθήκες προγραμμάτων

Packages, Cartridges). Πιο συγκεκριμένα, τα πλεονεκτήματα που προκύπτουν από την χρήση του Oracle ConText για την υλοποίηση του συστήματος ανάκτησης μπορούν να συνοψισθούν στα εξής:

1. Ολοκλήρωση και ενοποίηση των λειτουργιών RDBMS και IRS. Αυτός ο συνδυασμός είναι ιδανικός για τις απαιτήσεις ανάκτησης που συναντάμε σε μεγάλους σύγχρονους οργανισμούς όπου υπάρχει ανάγκη ανάκτησης τόσο δομημένης όσο και αδόμητης πληροφορίας. Για παράδειγμα μια κλασσική εφαρμογή αυτού του είδους, είναι η ανάγκη για ανάκτηση όχι μόνο κειμένων, άρθρων, υποσημειώσεων (memos) κ.α. που ικανοποιούν την ερώτηση του χρήστη από άποψη νοηματικού περιεχομένου αλλά ικανοποιούν ταυτόχρονα και άλλα δομημένα κριτήρια όπως η ημερομηνία του άρθρου ή το μέγεθος ενός κειμένου να βρίσκονται σε προκαθορισμένα όρια.
2. Το Oracle ConText αποτελεί μέρος του Oracle Universal Server ο οποίος και παρέχει ευκολίες που επιτρέπουν την ανάπτυξη Web εφαρμογών. Επιπλέον υπάρχει η δυνατότητα παράλληλης επεξεργασίας (parallel processing) όλων των λειτουργιών που σχετίζονται με την υποστήριξη και την λειτουργία της ανάκτησης, εφόσον υπάρχουν και λειτουργούν ταυτόχρονα περισσότεροι από έναν ConText εξυπηρετητές. Σε αυτή τη περίπτωση μέσω PL/SQL εντολών μπορούμε να κατανείμουμε τις λειτουργίες και το φόρτο εργασίας σε περισσότερους από έναν ConText εξυπηρετητές. Για παράδειγμα ένας θα μπορούσε να αναλάβει την δεικτοδότηση των κειμένων, άλλος την αναζήτηση ή την επιστροφή των αποτελεσμάτων, άλλος τη συλλογή και συντήρηση της ουράς των αιτήσεων των πελατών κ.λ.π. Οι παραπάνω δυνατότητες επιτρέπουν την ανάπτυξη συστημάτων ανάκτησης που μπορούν να υποστηρίζουν ταυτόχρονα πολλούς χρήστες από διάφορα μέρη του κόσμου έχοντας μικρούς χρόνους απόκρισης.
3. Σε αντίθεση με τα περισσότερα παραδοσιακά IRS στα οποία η συλλογή των κειμένων είναι συνήθως στατική και δεν είναι εύκολο να γίνουν εισαγωγές ή ενημερώσεις στη βάση, το Oracle ConText επιτρέπει την εισαγωγή, διαγραφή ή ενημέρωση των κειμένων της συλλογής κατά τρόπο διαφανές προς τον χρήστη. Κάθε φορά που συμβαίνει μία από τις παραπάνω πράξεις στον κύριο πίνακα της συλλογής, το ConText "παγιδεύει" (trap) τα παραπάνω γεγονότα και εκτελεί από μόνο του τις απαραίτητες ενημερώσεις στους πίνακες δεικτοδότησης.
4. Το ConText παρέχει πολυγλωσσική υποστήριξη (multilingual support) από την άπυψη ότι μπορεί και πραγματοποιεί χωρίς πρόβλημα αναζήτησεις σε οποιαδήποτε γλώσσα χρησιμοποιεί το λατινικό αλφάριθμο. Μπορεί και "καταλαβαίνει" δηλαδή και χαρακτήρες άλλων γλωσσών εκτός της αγγλικής και πραγματοποιεί σωστά την μετατροπή κεφαλαίων-μικρών για οποιαδήποτε γλώσσα. Βέβαια ορισμένες δυνατότητες και λειτουργίες όπως αποκοπή καταλήξεων, fuzzy ταίριασμα κ.α. εξακολουθούν να ισχύουν μόνο για την αγγλική γλώσσα αλλά οι μελλοντικές εκδόσεις του ConText σκοπεύουν να καλύψουν γραμματολογικά περισσότερες γλώσσες.
5. Δεν υπάρχει περιορισμός όσον αφορά το μέγεθος των κειμένων της συλλογής ενώ αυτά όπως έχουμε ήδη αναφέρει μπορούν να βρίσκονται σε διάφορα μορφότυπα χωρίς να επηρεάζεται η λειτουργία της ανάκτησης. Ειδικότερα θα πρέπει να

τονίσουμε ότι τα κείμενα μπορούν να είναι σε HTML μορφή, πράγμα που διευκολύνει ακόμα περισσότερο την υλοποίηση WWW συστημάτων ανάκτησης. Επιπλέον παρέχεται η δυνατότητα διαίρεσης ενός κειμένου και η αποθήκευσή του στη συλλογή κατά παραγράφους (βλ. παρ. 3.3.1.1 κατηγορία Master/Detail)

6. Παρέχονται πολλές από τις δυνατότητες που υπάρχουν στα παραδοσιακά IRS όπως αυτές της αποκοπής καταλήξεων, τετριμένων λέξεων, βαθμολόγηση των κειμένων, χρήση τελεστών στη ερώτηση του χρήστη κ.α.. Οπως συμβαίνει με όλα τα IRS έτσι και το ConText, υλοποιεί με τον δικό του τρόπο τα παραπάνω χαρακτηριστικά χρησιμοποιώντας τις δικές του εσωτερικές δομές δεδομένων με αποτέλεσμα άλλα χαρακτηριστικά να υπερτερούν και άλλα να μειονεκτούν σε σύγκριση με άλλα συστήματα ανάκτησης.
7. Τέλος, εφόσον το ConText αποτελεί επέκταση της Oracle RDBMS, παρέχονται οι ίδιες ευκολίες και δυνατότητες ανάκαμψης από λάθη (recovery from crashes), κλειδώματος και ασφάλειας των δεδομένων (data security), ελέγχου των χρηστών κ.α. που παρέχει ο Oracle Database Server.

Τα αδύνατα σημεία του Oracle ConText όπως αυτά φάνηκαν μέσα από την υλοποίηση του συστήματος ανάκτησης είναι τα εξής:

1. Το ConText υποστηρίζει την δυνατότητα αποκοπής καταλήξεων αλλά σε επίπεδο ερώτησης του χρήστη και όχι σε επίπεδο εσωτερικών δομών αποθήκευσης. Με άλλα λόγια, το stemming υλοποιείται διαμέσου τελεστών στην ερώτηση του χρήστη ενώ στο ανεστραμμένο αρχείο όρων αποθηκεύονται οι όροι στην πλήρη τους μορφή και όχι οι ρίζες τους. Βέβαια αυτό αποτελεί σχεδιαστική επιλογή της Oracle αναγκαία μεν για να μπορεί να υποστηρίζει ταυτόχρονα πολλές γλώσσες και να δίνει την δυνατότητα στον χρήστη να επιλέγει αυτός αν θέλει αποκοπή καταλήξεων ή όχι στην ερώτησή του, αλλά από την άλλη αυξάνεται το μέγεθος του ανεστραμμένου αρχείου όρων και κατ' επέκταση και ο χρόνος αναζήτησης.
2. Ο αριθμός των τετριμένων λέξεων περιορίζεται σε 255 με αποτέλεσμα να μην μπορούν να ενσωματωθούν πλήρως οι τετριμένες λέξεις περισσοτέρων των μιας γλωσσών για την περίπτωση δημιουργίας πολυγλωσσικού συστήματος ανάκτησης. Αρκεί να αναφέρουμε ότι το πλήρες σύνολο των τετριμένων λέξεων για την ελληνική γλώσσα ανέρχεται σε 376. Πάντως έχει ανακοινωθεί ότι η νεότερη έκδοση του ConText θα μπορεί να υποστηρίζει μέχρι 4096 τετριμένες λέξεις.
3. Η βαθμολόγηση των κειμένων της συλλογής - αν δεν έχουμε χρησιμοποιήσει τελεστές απόδοσης βαρών στους όρους της ερώτησης - βασίζεται εξ ορισμού στη συχνότητα εμφάνισης των όρων στα περιεχόμενα των κειμένων. Με αυτόν τον τρόπο όμως το ConText, μεροληπτεί και πριμοδοτεί τα μεγάλα σε μέγεθος κείμενα διότι φιλτριλογικά συναμένεται τι κείμενα αυτά να περιέχουν περισσότερες εμφανίσεις των όρων της ερώτησης. Εκτός αυτού η τελική βαθμολογία που δίνει το ConText σε κάθε κειμένου δεν είναι πολύ διευκρινιστική καθώς οποιοδήποτε κείμενο περιέχει έναν όρο πάνω από 10 φορές θα πάρει την ανώτερη βαθμολογία που είναι 100 (αριθμός εμφάνισης όρου X 10 μονάδες η κάθε εμφάνιση). Ετσι για παράδειγμα αν η ερώτηση ήταν "υπολογιστής", δύο

κείμενα που το ένα περιέχει τον όρο "υπολογιστής" 10 φορές και το άλλο περιέχει τον ίδιο όρο 18 φορές θα πάρουν την ίδια ανώτερη βαθμολογία που είναι 100.

4. Τα φύλτρα που χρησιμοποιεί το ConText προκειμένου να διαβάσει τα περιεχόμενα των κειμένων που δεν βρίσκονται σε απλή μορφή κειμένου (plain text), δεν λειτουργούν όταν τα κείμενα της συλλογής αποθηκεύονται σε φυσικά αρχεία στον δίσκο έξω από τον κύριο πίνακα της βάσης. Αυτό το πρόβλημα το αντιμετωπίσαμε έμμεσα στην υλοποίηση του συστήματος ανάκτησης με την αφαίρεση των tags από τις HTML σελίδες και την αποθήκευση τους ως απλά αρχεία κειμένου.
5. Το ConText στη παρούσα έκδοση του 7.3 δεν μπορεί να υποστηρίξει 100% δίγλωσσα κείμενα όσον αφορά την αποκοπή καταλήξεων και την αφαίρεση των τετριμένων λέξεων. Ο αριθμός των τετριμένων λέξεων που υποστηρίζεται είναι μέχρι 255 ανεξαρτήτου γλώσσας (π.χ. μπορούμε να έχουμε αγγλικές και ελληνικές τετριμένες λέξεις μαζί αλλά ο συνολικός αριθμός τους δεν θα πρέπει να ξεπερνά τις 255) οπότε είναι μάλλον αδύνατο να μπορούμε να συμπεριλάβουμε το πλήρες σύνολο των τετριμένων λέξεων δύο ή περισσοτέρων γλωσσών. Όσον αφορά την αποκοπή καταλήξεων, διατίθεται μόνο αγγλικό stemming και δεν υπάρχει τρόπος να ενσωματώσουμε την δική μας συνάρτηση αποκοπής καταλήξεων άλλης γλώσσας. Και τα δύο παραπάνω προβλήματα θα μπορούσαν να ξεπεραστούν με έμμεσο τρόπο όπως έγινε και στο σύστημα ανάκτησης που υλοποιήθηκε. Η αποκοπή καταλήξεων της αγγλικής ή της ελληνικής ή οποιασδήποτε άλλης γλώσσας θα μπορούσε να ινλοποιηθεί με την χρήση χαρακτήρων μπαλαντέρ εφόσον έχουμε πρώτα στη διάθεση μας την αντίστοιχη συνάρτηση αποκοπής καταλήξεων η οποία δοθέντος ενός όρου επιστρέφει την ρίζα του. Βέβαια παραμένει το πρόβλημα της ανακάλυψης της γλώσσας του όρου - αγγλικός, ελληνικός ή άλλης γλώσσας όρος - ώστε να χρησιμοποιηθεί η αντίστοιχη συνάρτηση αποκοπής καταλήξεων. Με τις τετριμένες λέξεις τα πράγματα είναι πιο εύκολα αφού μπορούμε να γεμίσουμε ένα πίνακα με όσες τετριμένες λέξεις θέλουμε οποιασδήποτε γλώσσας και με βάση αυτόν να τις αφαιρούμε από το περιεχόμενο των κειμένων πριν την δεικτοδότησή τους και από την ερώτηση του χρήστη πριν την εκτέλεση της αναζήτησης.
6. Το ConText αποτελεί ένα τελικό προϊόν της Oracle Corporation και δεν αφήνει κανένα περιθώριο επέμβασης του χρήστη ώστε να το προσαρμόσει στις δικές του ιδιαίτερες ανάγκες. Για παράδειγμα δεν μπορούμε να ενσωματώσουμε δική μας συνάρτηση αποκοπής καταλήξεων ή να γράψουμε την δική μας συνάρτηση ομοιότητας (similarity function) ώστε να γίνουν πειράματα και να καταλήξουμε στην βέλτιστη λύση από άποψη ποιότητας των αποτελεσμάτων. Με άλλα λόγια, λόγω του ότι το ConText είναι εμπορικό προϊόν και δεν έχει εκπαιδευτικούς σκοπούς δεν είναι καθόλου παραμετροποιήσιμο ώστε να μπορεί να προσαρμοστεί στις εκάστοτε ανάγκες του χρήστη αλλά και να μπορούμε να λάβουμε υπόψη τις ιδιαιτερότητες της κάθε συλλογής..

6.2 Βελτιώσεις - Περαιτέρω δουλειά

Λρχική φυλοδοξία της παρούσας διπλωματικής εργασίας ήταν η κατασκευή μιας μηχανής αναζήτησης για όλα τα αποθέματα πληροφοριών που υπήρχαν στην ελληνική γλώσσα. Κάτι τέτοιο ύμως στην συνέχεια αποδείχτηκε πολύ δύσκολο λόγω του όγκου των πληροφοριών - HTML σελίδων - που υπάρχουν στο διαδίκτυο αλλά και λόγου της δυσκολίας ανακάλυψης αυτών από το σύστημα ανάκτησης. Εξάλλου σκοπός της παρούσας διπλωματικής εργασίας ήταν η κατασκευή ενός πλοτικού συστήματος ανάκτησης και όχι μια ολοκληρωμένη λύση. Ετσι προτιμήθηκε ο περιορισμός του συστήματος ανάκτησης σε αναζήτησεις ενός συγκεκριμένου χώρου τόσο θεματολογικά όσο και διευθυνσιακά οριθετημένου.

Αν και συνήθως ο Web Crawler των μηχανών αναζήτησης θεωρείται αυτόνομο πρόγραμμα τρίτων κατασκευαστών, στο σύστημα ανάκτησης που υλοποιήθηκε βελτιώσεις θα μπορούσαν να γίνονται σε αυτόν ώστε να μπορεί να ανακαλύπτει και να ακολουθεί περισσότερους συνδέσμους και να πραγματοποιεί ένα τοπολογικό διαχωρισμό αυτών με απότερο σκοπό την κάλυψη και δεικτοδότηση των αποθεμάτων του ελληνικού χώρου. Επίσης θα μπορούσε να ενσωματωθεί σε αυτόν μια διαδικασία ανακάλυψης της γλώσσας στην οποία είναι γραμμένη μία HTML σελίδα και εφόσον είναι ελληνική να ανακτάται και να δεικτοδοτείται από το σύστημα ανάκτησης. Μία μεθοδολογία για αυτό το σκοπό είναι η χρήση n-grams^[19,20]. Βέβαια τέτοιες βελτιώσεις θα πρέπει να συνοδεύονται και από σχεδιαστικές αλλαγές στο σύστημα ανάκτησης όπως κατάργηση της χρήσης της Access βάσης δεδομένων και λειτουργία μόνο με το Oracle DBMS, άμεση δεικτοδότηση των σελίδων ώστε να μην γίνεται καθόλου τοπική απωθήκευσή τους γιατί ο όγκος των πληροφοριών θα είναι τεράστιος κ.λ.π. Προς την κατεύθυνση βελτίωσης της πληρότητας της συλλογής θα μπορούσε να συμβάλει και η ύπαρξη επιλογής στον χρήστη μέσο της οποίας θα μπορούσε να υποβάλει μια αίτηση δεικτοδότησης στο σύστημα ανάκτησης (επιλογή Add URL που υπάρχει στις σύγχρονες μηχανές αναζήτησης όπου ο χρήστης εισάγει ένα URL και η μηχανή αναζήτησης ανταποκρίνεται ότι εντός ορισμένου χρονικού διαστήματος θα δεικτοδοτήσει της σελίδες αυτής της περιοχής) ώστε να διευκολύνει και τον Web Crawler στην ανακάλυψη καινούργιων αποθεμάτων αλλά και να διατηρείται η βάση ενημερωμένη με τις πιο πρόσφατες αλλαγές.

Οσον αφορά την λειτουργία του ελληνικού stemming το οποίο αποτελεί σημαντική λειτουργία από άποψης ποιότητας των αποτελεσμάτων, θα μπορούσαμε να προτείνουμε ένα διαφορετικό τρόπο υλοποίησής του ο οποίος να προσεγγίζει αυτόν που συναντάμε στα παραδοσιακά IRS. Πιο συγκεκριμένα, κάθε φορά που θα αποθηκεύεται μια HTML σελίδα τοπικά στον δίσκο σε μορφή απλού κειμένου, θα μπορούσαμε να εφαρμόζαμε ένα φίλτρο το οποίο θα έκανε αποκοπή καταλήξεων στους όρους της οπότε όταν το ConText έκανε δεικτοδότηση των σελίδων θα αποθήκευε στο αναστραμμένο αρχείο όρων τις ρίζες των όρων πράγμα που θα μείωνε το χώρο αποθήκευσης και τον χρόνο αναζήτησης. Αυτή η λύση ύμως έχει επιτυχία και εφαρμογή εφόσον ισχύει μια βασική προϋπόθεση: Θα πρέπει να μπορούμε να καταλαβαίνουμε το περιεχόμενο της κάθε σελίδας προκαταβολικά ώστε να ξέρουμε ποιόν αλγόριθμο stemming να εφαρμόσουμε (αγγλικό, ελληνικό ή άλλο) ακόμα και

σε επίπεδο λέξης αφού πολλές σελίδες περιέχουν και αγγλικές και ελληνικές λέξεις ή και άλλων γλωσσών ταυτόχρονα.

Τέλος, βελτιώσεις θα μπορούσαν να γίνουν στην διεπαφή του συστήματος ανάκτησης με επιπλέον επιλογές που διευκολύνουν τον χρήστη. Τέτοιες θα μπορούσε να είναι η ύπαρξη επιλογής όσο αναφορά των αριθμών των κειμένων που επιστρέφονται σε κάθε σελίδα αποτελεσμάτων μαζί με την προσθήκη κουμπιών περιήγησης (navigation buttons) προηγούμενο-επόμενο, την επιλογή για το αν θα πρέπει να επιστρέφονται η περιγραφές των κειμένων στα αποτελέσματα ή όχι, την παροχή βοήθειας και την επιλογή απλής ή προχωρημένης (advanced) αναζήτησης ώστε η εφαρμογή των διαφόρων τελεστών να γίνεται με γραφικό τρόπο, δυνατότητες επανατροφοδότησης από την πλευρά του χρήστη και επαναδιαβάθμισης των αποτελεσμάτων κ.λ.π.

Το τελικό συμπέρασμα που προέκυψε από την διατριβή μου με το αντικείμενο της ανάκτησης πληροφοριών και ειδικά αυτών που υπάρχουν στον διαδίκτυο, είναι η ανάγκη για την ύπαρξη αποτελεσματικών εργαλείων τα οποία θα εφαρμόζουν στο μεγαλύτερο βαθμό στο οποίο αυτό είναι εφικτό, τεχνικές και εργαλεία από το χώρο της ανάκτησης πληροφοριών ώστε τα αποτελέσματα που δίνουν να είναι όσο το δυνατό καλύτερα και ακριβέστερα διότι ο όγκος πληροφοριών είναι πράγματι τεράστιος και το νοηματικό περιεχόμενο σε ορισμένες περιπτώσεις είναι ασαφές. Σαν πρώτη διαπίστωση, ο ελληνικός χώρος έδειξε ανώριμος για την ύπαρξη μιας μηχανής αναζήτησης που μπορεί και υποστηρίζει αναζητήσεις στην ελληνική γλώσσα και αυτό φαίνεται από το γεγονός ότι ολόκληρες περιοχές του ελληνικού χώρου περιέχουν μόνο αγγλικές εκδόσεις των σελίδων τους ή τουλάχιστον η αναζήτηση τους με αγγλικούς όρους είναι το κύριο μέλημα τους. Βέβαια αυτό θα πρέπει να εξεταστεί σε συνδυασμό με την διάδοση που έχει το Internet στην Ελλάδα και το είδος των ατόμων που το χρησιμοποιούν σαν εργαλείο για την πραγματοποίηση αναζητήσεων. Τα πανεπιστήμια, τα εκπαιδευτικά ιδρύματα και οι φορείς παροχής υπηρεσιών διαδικτύου (Internet Service Providers) αποτελούν σήμερα τον κύριο κορμό αποθεμάτων του ελληνικού χώρου ο οποίος σε ποσοστό, σε σύγκριση με ολόκληρο το διαδίκτυο, είναι απειροελάχιστος ενώ οι μη περιστασιακοί χρήστες του Internet προέρχονται από παρόμοιους χώρους και οι πληροφορίες που αναζητούν υπάρχουν συνήθως σε περιοχές εκτός Ελλάδας.

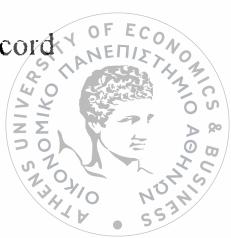
Το Internet όμως έχει ήδη αρχίσει και συνεχίζει με γοργούς ρυθμούς να εδραιώνεται και στην Ελλάδα σαν μέσο προβολής και διαφήμισης και στο μέλλον αναμένεται ολοένα και περισσότερες επιχειρήσεις και οργανισμοί που απευθύνονται και στο ελληνικό κοινό να αποκτήσουν παρουσία στο διαδίκτυο. Σε μία τέτοια προοπτική, η ανάγκη για ύπαρξη εξειδικευμένων συστημάτων ανάκτησης που καλύπτουν επαρκώς τα αποθέματα του ελληνικού χώρου θα είναι επιτακτική ενώ η ανάπτυξη φορέων που προσφέρουν αυτές τις υπηρεσίες αναζήτησης όπως συμβαίνει σήμερα με τις μεγάλες μηχανές αναζήτησης του εξωτερικού θα είναι μεγάλη και οικονομικώς προσιδοφόρα.

Η ύπαρξη αποτελεσματικής ανάκτησης ελληνικών κειμένων είναι σημαντική τόσο από επιστημονικής πλευράς όσο και πολιτισμικής. Πράγματι υπάρχει πλούτος υλικού τόσο επιστημονικού όσο και πολιτισμικού περιεχομένου (π.χ. μουσεία, ιδιωτικές

βιβλιοθήκες, ιδρύματα, μοναστήρια) που είναι επιτακτική η ανάγκη να περαστεί στο διαδίκτυο. Σε μία τέτοια προοπτική τα προβλήματα που πρέπει να ρέπεραστούν είναι πολλά και σχετίζονται με τις ιδιαιτερότητες της ελληνική γλώσσας όπως είδος ομιλίας και γραφής (καθαρεύουσα-καθομιλουμένη), σύστημα γραφής (μονοτονικό-πολυτονικό), διαφορετικά πρότυπα αναπαράστασης των ελληνικών χαρακτήρων (ΕΛΟΤ 928, 437 κ.α.) και άλλα πολλά που σχετίζονται με τον τρόπο που θα περαστεί και θα συντηρηθεί η πληροφορία στο διαδίκτυο. Κάτω από τις συνθήκες που ήδη επιβάλει ή που θα επιβάλει η σύγχρονη ελληνική πραγματικότητα το πρόβλημα της αποτελεσματικής ανάκτησης ελληνικών κειμένων αποτελεί ανοικτό ερευνητικό πεδίο με μακρύ δρόμο αλλά και μεγάλα οφέλη τόσο επιστημονικά όσο και κοινωνικά.

Βιβλιογραφία - Αναφορές

- [1] Smeaton F. Alan, "Information Retrieval: Advances in the last 5 years", 4th Congress on Law and Computers, May 1988, Technical Report, Dublin City University.
- [2] Καλαμπούκης Ζ. Θεόδωρος, "Ενα σύστημα ανάκτησης Ελληνικών κειμένων", 1995, Κέντρο Έρευνας Οικονομικού Πανεπιστημίου Αθηνών.
- [3] C. J. Van Rijsbergen, "Information Retrieval", 1979, 2nd Edition, Butterworths, London.
- [4] Kalamboukis.T.Z, "Suffix stripping with Modern Greek", Program, automated library and Information Systems, Vol. 29, 1995, 313-321.
- [5] Salton G., "Automatic Text Processing", The Transformation Analysis, and Retrieval of Information by computer, Reading, MA: Addison-Wesley, 1989.
- [6] Martijn Koster, NEXOR, "Robots in the Web: threat or treat?", April 1997
[<http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html>]
- [7] Kathleen Webster & Kathryn Paul, Information Technology, "Beyond Surfing: Tools and Techniques for Searching the Web", January 1996
[<http://magi.com/~mmelick/it96jan.htm>]
- [8] David Wells & Ansu Kurien, "Searching and Indexing", April 1996
[<http://www.onjs.com/survey/crawl.htm>]
- [9] Ann Eagan & Laura Bender, "Spiders and Worms and Crawlers, Oh my: Searching on the World Wide Web", 1996 [<http://www.library.ucsh.edu/untangle/eagan.html>]
- [10] Danny Sullivan, "Search Engine Watch", 1996-1997
[<http://www.searchenginewatch.com>]
- [11] Smeaton F. Alan, "An Information Retrieval System Implemented on top of a Relational Database", 1989
- [12] M. F. Porter, "An Algorithm for Suffix Striping", Program, Vol. 14, No. 2, 1980, pg. 130-137
- [13] R. C. Russel . U.S. Patent Office 1261167, April 1918
- [14] R. C. Russel . U.S. Patent Office 1435663, November 1922
- [15] L. Davidson, "Retrieval of misspelt names in an airline passenger record system", Communications of the ACM, vol. 5, 1962, pp.169-171



[16] "Oracle ConText Option Application Developer's Guide Release 1.1", 1995-1997 Oracle Corporation

[17] "Oracle ConText Option Administrator's Guide Release 1.1", 1995-1997 Oracle Corporation

[18] Timothy O' Brien, "Oracle ConText: Text Looms as the Next Frontier in Information Management", April 1996 for Oracle Corporation

[19] W. B. Cavnar, "Using an n-gram based document representation with a vector processing retrieval model", Proceedings of TREC 3.

[20] W. B. Cavnar, "N-gram based text filtering for TREC-2", Proceedings of TREC-2 Text Retrieval Conference 2, Donna Harman, ed. National Bureau of Standards, 1993.



General description

The subject of this Msc. Thesis is the development and evaluation of an Information Retrieval System (IRS) for Greek Documents. In the context of this Msc. Thesis, we have designed and developed a WWW search engine which is fully capable of handling and supporting user queries in English as well as in Greek language. Our further intention was the development of a WWW search Engine that could make use of techniques and characteristics which exists in the traditional information retrieval systems like removal of stopwords, suffix stripping, relevance ranking, use of operators in user queries e.t.c. in order to achieve better results and fully support the Greek language.

Starting ambition of this Msc. Thesis was the development of a search engine for all the sites that belongs in the Greek domain or better, for all the Internet resources which are in Greek language. This attempt soon proved to be very difficult to accomplish because of the huge number of HTML pages that there are in the Greek domain and because it was too difficult to discover all the Internet resources which are in Greek language. Besides, primary attention of the Msc. Thesis was the development of a prototype and not a complete functional system. That is why we choose to limit the IRS search to a space of specific address and subject.

For the achievement of the above goals, we developed a Web Robot program which was responsible for finding HTML pages on the Internet that need to be indexed and for downloading them for the creation of the system's collection. The result of the Web robot's function was the discovery and downloading of about 2500 HTML pages from about 30 Greek sites which thematic belongs to the "Health & Medicine" category.



The indexing of the contents of the HTML pages and the creation of the indexes as well as the engine of the IRS was based on the services provided by the Oracle ConText Option 7.3. These services were extended by adding the full set of the Greek stopwords, 376, and a suffix stripping algorithm for modern Greek that was responsible for the function of stemming on Greek terms.

In the end, we evaluated the results the IRS returned to us for two cases: when we apply suffix stripping to the terms of the queries and when we don't. For the evaluation of the system's results, we submitted to the system 20 queries in Greek language, in the first case with suffix stripping of the terms and in the second case without and then we calculated the recall-precision values and created the corresponding graphs. Furthermore, for the sake of completeness we compare the system's results with the results we took from Altavista for the same queries and arrive at valuable conclusions.

The structure of this Msc. Thesis contains six chapters :

In the first chapter, we explain basic concepts of information retrieval and we describe the theoretical base the IRS we developed relies on.

In the second chapter, we give a general description of the most famous search engines that exists today in the Internet as well as the problems that have to deal and get over with.



In the third chapter, we analyze the relationship between information retrieval systems and database management systems given that our system was developed with the Oracle RDBMS 7.3. We then make a quick presentation of such a system (bibliographic records system, *retriev*) and then we describe the text retrieval services that the Oracle ConText Option offers.

In the fourth chapter, we make a full description of the architecture and function of the Web Crawler program and the IRS for Greek documents. In both cases, we give pictures of the user interface and scenarios about how to use them.

In the fifth chapter, we mention the parameters that participate and the methodology that being used in evaluation of information retrieval systems. Then we present the results of the experiments we make on the IRS and give out the conclusions we arrive at.

In the sixth chapter, we mention the advantages and disadvantages about using the Oracle ConText Option in building information retrieval systems and we suggest some improvements and further research that can be done for building better information retrieval systems that can fully handle queries in Greek language.



