

Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΑΧΕΙΡΙΣΗ ΚΑΙ ΔΕΙΚΤΟΔΟΤΗΣΗ
ΕΛΕΥΘΕΡΟΥ ΚΕΙΜΕΝΟΥ



**Select docid, text, ...
From articles
Where contains(text, 'SQL') = 1
Order by date**



doc11 Oracle releases Application Server 4...
doc4 Microsoft releases SQL Server 7...
doc93 IBM announces new OLAP tools for DB2

ΣΠΟΥΔΑΣΤΗΣ : ΛΟΥΡΔΑΣ ΒΑΣΙΛΕΙΟΣ
ΕΙΣΗΓΗΤΗΣ : ΔΕΡΒΟΣ ΔΗΜΗΤΡΙΟΣ

ΜΑΡΤΙΟΣ 2000
ΘΕΣΣΑΛΟΝΙΚΗ

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ την οικογένειά μου για την ηθική υποστήριξή της.

Ευχαριστώ τον επιβλέποντα της πτυχιακής εργασίας κ. Δέρβο για την υποστήριξη, την καθοδήγηση, τις συμβουλές του για να φτάσει αυτή η εργασία στο τέλος της, καθώς επίσης και για την προσφορά της DB2.

Ευχαριστώ θερμά την εταιρία PC Systems για την προσφορά της Oracle 8.0.5 καθώς και για την οποιαδήποτε βοήθεια πρόσφερε όποτε χρειάστηκα. Η άσκηση της πρακτικής εργασίας μου στην εταιρία με προσέφερε παραπάνω από όσα αρχικά είχα φανταστεί και υπήρξε μια πολύτιμη εμπειρία για μένα.

Ευχαριστώ το συνάδελφο Φεύγα Αθανάσιο για την μικρή έστω καθοδήγηση όσον αφορά στις σελίδες asp και ο οποίος αυτή τη στιγμή υπηρετεί την πατρίδα.

Τέλος, θέλω να ευχαριστήσω φίλους και γνωστούς για το κουράγιο που μου έδωσαν ώστε να συνεχίζω κάθε φορά.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΙΕΣ	1
ΠΕΡΙΕΧΟΜΕΝΑ.....	4
ΚΕΦΑΛΑΙΟ 1	9
ΕΙΣΑΓΩΓΗ ΣΤΑ ΣΥΣΤΗΜΑΤΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ	9
1.1 Ορισμός του συστήματος ανάκτησης πληροφορίας	10
1.2 Αξιολόγηση ενός ΣΑΠ.....	10
1.3 Ακρίβεια και ανάκτηση.....	11
1.4 Ανασκόπηση λειτουργίας ενός ΣΑΠ.....	13
1.4.1 Κανονικοποίηση όρων	13
1.4.2 Επιλεκτική διανομή της πληροφορίας	16
1.4.3 Αναζήτηση εγγράφων στη βάση δεδομένων	16
1.4.4 Ευρετηριοποίηση	17
1.5 Συσχέτιση με τα συστήματα διαχείρισης βάσεων δεδομένων	17
1.6 Ψηφιακές βιβλιοθήκες και αποθήκες δεδομένων.....	18
ΚΕΦΑΛΑΙΟ 2	20
ΣΥΓΧΡΟΝΕΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ ΣΤΟ ΔΙΑΔΙΚΤΥΟ.....	20
2.1 Εισαγωγή.....	21
2.2 Ορισμός της μηχανής αναζήτησης.....	21
2.3 Τρόπος λειτουργίας	21
2.4 Χαρακτηριστικά των μηχανών αναζήτησης	22
2.4.1 Crawling.....	22
2.4.2 Ευρετηριοποίηση	23
2.4.3 Διαβάθμιση	23
2.4.4 Spamming	24
2.4.5 Πώς αντιμετωπίζονται τα κεφαλαία και πεζά.....	24
2.5 Μηχανές αναζήτησης και θεματικοί κατάλογοι.....	25
2.5.1 Μηχανές αναζήτησης.....	25
2.5.2 Θεματικοί κατάλογοι	25
2.6 Κατηγορίες μηχανών αναζήτησης	26
2.6.1 Γενικού περιεχομένου	26
2.6.2 Εξυπηρέτησης παιδικών αναγκών	26
2.6.3 Metacrawlers.....	26
2.6.4 Εύρεσης αρχείων πολυμέσων	27
2.6.5 Αναζήτησης ειδήσεων και νέων	27
2.6.6 Αναζήτησης συγκεκριμένου γεωγραφικού χώρου.....	27
2.6.7 Εξειδικευμένων θεμάτων	27

2.7	Βασική σύνταξη ερωτήσεων	28
2.7.1	Λογικοί τελεστές	29
2.7.2	Τελεστής NEAR	29
2.8	Μεγέθη των μηχανών αναζήτησης και των θεματικών καταλόγων	29
2.8.1	Μεγέθη των μηχανών αναζήτησης	29
2.8.2	Μεγέθη των θεματικών καταλόγων	31
2.9	Το μέλλον των μηχανών αναζήτησης	32
ΚΕΦΑΛΑΙΟ 3		33
IBM DB2 TEXT EXTENDER		33
3.1	Εισαγωγή στον IBM DB2 Text Extender	34
3.2	Ο Text Extender σε περιβάλλον πελάτη - διακομιστή	34
3.3	Είδη ευρετηρίων του Text Extender	36
3.3.1	Γλωσσολογικό ευρετήριο (<i>linguistic index</i>)	36
3.3.2	Ακριβές ευρετήριο (<i>precise index</i>)	37
3.3.3	Διπλό ευρετήριο (<i>dual index</i>)	37
3.3.4	Ngram ευρετήριο	37
3.4	Γλωσσολογική επεξεργασία για την ευρετηριοποίηση	38
3.4.1	Ανάλυση κειμένου	39
3.4.2	Αναγωγή των όρων στη βασική τους μορφή	40
3.4.3	Φιλτράρισμα των στοπ λέξεων	40
3.4.4	Απλοποίηση των σύνθετων όρων	40
3.4.5	Εξαγωγή χαρακτηριστικών	41
3.5	Γλωσσολογική επεξεργασία για την ανάκτηση	42
3.5.1	Επαύξηση συνωνύμων	43
3.5.2	Επαύξηση θησαυρού	43
3.5.3	Επαύξηση domain όρων και συντομογραφιών	43
3.5.4	Χαρακτήρες και λέξεις μάσκες	43
3.5.5	Επαύξηση ονομάτων	43
3.5.6	Ηχητική επαύξηση	44
3.6	Γλωσσολογική επεξεργασία για την πλοήγηση	44
3.6.1	Πρώτο στάδιο	44
3.6.2	Δεύτερο στάδιο	44
3.7	Γλωσσολογικές λειτουργίες που υποστηρίζει ο Text Extender ανά γλώσσα	45
3.8	Αρχές και έννοιες θησαυρού	46
3.8.1	Όροι	46
3.8.2	Σχέσεις	46
3.9	Διαχειριστικές εργασίες του Text Extender	48
3.9.1	Εργασίες στον διακομιστή Text Extender	49
3.9.2	Εργασίες στον πελάτη Text Extender	49
3.10	SQL ερωτήσεις	53

3.10.1	Σύνταξη ερώτησης	53
3.10.2	Περίληψη κανόνων και περιορισμών.....	59
3.10.3	Συναρτήσεις του Text Extender	60
ΚΕΦΑΛΑΙΟ 4		62
ORACLE CONTEXT CARTRIDGE.....		62
4.1	Εισαγωγή στο Oracle ConText cartridge	63
4.1.1	Χαρακτηριστικά του ConText cartridge	63
4.2	Ο διακομιστής ConText	63
4.2.1	Οι διακομιστές ConText και Oracle.....	63
4.2.2	Ρόλοι του διακομιστή ConText.....	64
4.3	Διαδικασία ευρετηριοποίησης.....	66
4.3.1	Πολιτικές (<i>policies</i>) ConText	66
4.3.2	Επιλογές ευρετηριοποίησης	68
4.4	Ενεργοποίηση του ConText cartridge, τερματισμός και εκκίνηση ενός διακομιστή ConText	69
4.4.1	Ενεργοποίηση του ConText	69
4.4.2	Εκκίνηση και τερματισμός διακομιστή.....	70
4.5	Ρόλοι και χρήστες του ConText στη βάση.....	71
4.6	Είδη ερωτήσεων	71
4.6.1	Ερωτήσεις δύο βημάτων	71
4.6.2	Ερωτήσεις ενός βήματος.....	72
4.6.3	Ερωτήσεις από μνήμη	72
4.6.4	Παράδειγμα ερώτησης δύο βημάτων	72
4.6.5	Παράδειγμα ερώτησης ενός βήματος.....	73
4.6.6	Παράδειγμα ερώτησης από μνήμη	73
4.7	Ερωτήσεις.....	74
4.7.1	Λογικοί τελεστές	75
4.7.2	Τελεστής WITHIN	75
4.7.3	Τελεστής NEAR.....	77
4.7.4	Τελεστές αλλαγής του σκορ.....	78
4.7.5	Τελεστές ελέγχου του συνόλου αποτελεσμάτων	79
4.7.6	Τελεστές επαύξησης	80
4.7.7	Τελεστές θησαυρού.....	81
4.7.8	Χαρακτήρες μάσκες (<i>wildcard characters</i>)	83
4.7.9	Χαρακτήρες καθορισμού γκρουπ	83
4.7.10	Αποθηκευμένες ερωτήσεις.....	83
4.7.11	Χαρακτήρες διαφυγής.....	84
4.8	Ερωτήσεις θεμάτων.....	84
4.8.1	Αρχές ευρετηριοποίησης.....	84
4.8.2	Γνωστά θέματα	85
4.8.3	Άγνωστα θέματα	85
4.8.4	Βάρος θέματος	85
4.8.5	Διαδικασία ερώτησης.....	86
4.8.6	Υπολογισμός σκορ	86

4.8.7	Διάκριση κεφαλαίων-πεζών	86
4.8.8	Δημιουργία ερωτήσεων θεμάτων	87
4.8.9	Τροποποίηση των ερωτήσεων (<i>refining queries</i>)	87
4.9	Διαδικασίες και συναρτήσεις για την εκτέλεση ερωτήσεων	88
4.9.1	Πακέτο CTX_QUERY	89
4.9.2	Πακέτο CTX_LING	91
4.9.3	Πακέτο CTX_DDL	93
4.9.4	Πακέτο CTX_DML	95
4.9.5	Συναρτήσεις CONTAINS και SCORE	96
4.9.6	Ερώτηση SELECT	96
ΚΕΦΑΛΑΙΟ 5		98
ΠΡΟΤΥΠΟ ΠΕΡΙΒΑΛΛΟΝ ΙΣΤΟΣΕΛΙΔΩΝ.....		98
5.1	Εισαγωγή.....	99
5.2	Σύντομη περιγραφή των asp σελίδων	99
5.3	Περιγραφή του μοντέλου ADO.....	100
5.3.1	Γενικά.....	100
5.3.2	ADO Αντικείμενα, συλλογές και γεγονότα	100
5.4	Περιγραφή του πρότυπου περιβάλλοντος	102
5.4.1	Σύνδεση με τη βάση.....	102
5.4.2	Ανάκτηση δεδομένων από τη βάση	102
5.4.3	Περιγραφή των ιστοσελίδων για τον Text Extender.....	103
5.4.4	Περιγραφή των ιστοσελίδων για το ConText cartridge	106
ΚΕΦΑΛΑΙΟ 6		110
ΕΠΙΛΟΓΟΣ.....		110
6.1	Γενική ανασκόπηση	111
6.2	Δυσκολίες και προβλήματα.....	111
6.3	Εντυπώσεις.....	112
ΠΑΡΑΡΤΗΜΑ		115
Διαχειριστικές εντολές του Text Extender για τον πελάτη		116
Διαχειριστικές εντολές του Text Extender για τον διακομιστή.....		118
ΓΛΩΣΣΑΡΙΟ.....		120
Αγγλικό.....		120
Ελληνικό.....		123
ΒΙΒΛΙΟΓΡΑΦΙΑ		126

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ ΣΤΑ ΣΥΣΤΗΜΑΤΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

1.1 Ορισμός του συστήματος ανάκτησης πληροφορίας

Σύστημα ανάκτησης πληροφορίας (information retrieval system) ονομάζεται το σύστημα που είναι ικανό να αποθηκεύει, ενημερώνει, ανακτά και διαχειρίζεται την πληροφορία στη γενικότερη δυνατή μορφοποίησή της (π.χ. ελεύθερο ή/και ημιδομημένο κείμενο, εικόνες, διαγράμματα, κτλ.). Παρόλη την ποικίλη δομή της πληροφορίας, ο πλέον γενικός τύπος μορφοποίησης είναι το κείμενο. Οι άλλοι τύποι χρησιμοποιούνται κυρίως αφού έχουν συσχετιστεί με κείμενα και προκύπτουν αφού γίνει αναζήτηση με βάση το κείμενο αυτά. Ωστόσο, στο προσεχές μέλλον, θα γίνει δυνατή η *ταύτιση σχεδίων (pattern matching)* και σε άλλους τύπους πληροφορίας.

Ο όρος *μονάδα πληροφορίας (item)* χρησιμοποιείται για να αντιπροσωπεύσει την μικρότερη πλήρη μονάδα κειμένου που επεξεργάζεται και χειρίζεται το σύστημα. Ο ορισμός της μονάδας πληροφορίας μπορεί να ποικίλει ανάλογα με το πώς το σύστημα διαχειρίζεται την πληροφορία. Για παράδειγμα, η μονάδα θα μπορούσε σε κάποιες περιπτώσεις να είναι ένα βιβλίο, ενώ σε άλλες, η μονάδα θα ήταν ένα κεφάλαιο ή μια παράγραφος. Ανάλογα σε κάθε περίπτωση υπάρχουν διαφορετικά επίπεδα αφαίρεσης.

Ένα σύστημα ανάκτησης πληροφορίας (ΣΑΠ χάριν συντομίας) αποτελείται από λογισμικό που βοηθά το χρήστη στην εύρεση της πληροφορίας που αναζητεί. Το ΣΑΠ μπορεί να χρησιμοποιεί κοινό ή εξειδικευμένο υλικό (*hardware*) για την διαδικασία αναζήτησης.

Τα ΣΑΠ υποδιαιρούνται σε δύο κατηγορίες : στα συστήματα *πλήρους ταύτισης (exact match)* και στα συστήματα *μερικής ταύτισης (partial match)*. Στα πρώτα, που ονομάζονται επίσης και *συστήματα μοντέλου Boole*, εξετάζεται αν τα περιεχόμενα της ερώτησης περιέχονται ή όχι σε κείμενα της συλλογής. Ο χρήστης διατυπώνει την ερώτηση χρησιμοποιώντας τους λογικούς τελεστές AND, OR, NOT και XOR. Τα συστήματα πλήρους ταύτισης ήταν και τα πρώτα που χρησιμοποιήθηκαν και συνεχίζουν να είναι ευρείας χρήσεως ακόμη και σήμερα. Τα *συστήματα μερικής ταύτισης* διαφέρουν από τα πλήρους ταύτισης στην ασάφεια της αναπαράστασης, η οποία πρέπει να σημειωθεί ότι δεν υποδηλώνει απώλεια πληροφορίας. Η σημασιολογία των κειμένων μοντελοποιείται σε ενδιαμέσες παραστάσεις, των οποίων το είδος διακρίνει τα επιμέρους μοντέλα ΣΑΠ. Η ασάφεια στην αναπαράσταση ωστόσο αντανακλάται και στο αποτέλεσμα μιας ερώτησης, δηλαδή δεν είναι σίγουρο ότι τα κείμενα του αποτελέσματος αποτελούν αποδεκτές απαντήσεις, καθώς επίσης ότι ούτε το σύνολο των κατάλληλων κειμένων που επιλέγεται είναι σωστό.

Τα ΣΑΠ χρησιμοποιούνταν αρχικά για την ανάκτηση βιβλιογραφικού τύπου αναφορών. Η επεξεργασία γινόταν με *μαζικό (batch)* τρόπο εκτός των ωρών παραγωγής. Με την εξέλιξη του υλικού, τα σύγχρονα ΣΑΠ έχουν πλέον δυνατότητα διαχείρισης ολόκληρων κειμένων και όχι μόνο επιλεκτικών κομματιών, ενώ η απόκριση γίνεται σε πραγματικό χρόνο.

1.2 Αξιολόγηση ενός ΣΑΠ

Ο χρήστης υποβάλλει μια ερώτηση στο ΣΑΠ. Αφού εκείνο την επεξεργαστεί, παρουσιάζει στο χρήστη το αποτέλεσμα. Το αποτέλεσμα περιέχει *μη σχετιζόμενη (non-relevant)* και *σχετιζόμενη (relevant)* πληροφορία που εξυπηρετεί την πληροφοριακή ανάγκη του χρήστη. Η μη σχετιζόμενη πληροφορία οφείλεται στην αδυναμία διάκρισης του ΣΑΠ των κειμένων που ικανοποιούν το χρήστη, έτσι το αποτέλεσμα εκτός της χρήσιμης πληροφορίας περιέχει και «θόρυβο».

Οι βασικοί στόχοι ενός ΣΑΠ είναι η *αποτελεσματικότητα* και η *αποδοτικότητα* του. *Αποτελεσματικό* είναι το ΣΑΠ που μπορεί να διακρίνει με επιτυχία κείμενα που ικανο-

ποιούν την πληροφοριακή ανάγκη του χρήστη. Αποδοτικό είναι το ΣΑΠ που για την ικανοποίηση της ερώτησης του χρήστη δεσμεύει πόρους (*resources*) του συστήματος με τέτοιο τρόπο ώστε να παραμένει ανταγωνιστικό σε σχέση με άλλες υλοποιήσεις. Η αξιολόγηση του ΣΑΠ γίνεται κυρίως πειραματικά. Για το σκοπό αυτό υπάρχουν μετρικές, οι οποίες αναπτύσσονται αναλυτικά παρακάτω (§1.3).

Από την πλευρά του χρήστη, οι έννοιες *σχετιζόμενο* και *αναγκαίο* είναι ταυτόσημες. Ωστόσο, από την πλευρά του ΣΑΠ οι έννοιες αυτές δεν ταυτίζονται πάντα. Η σχετιζόμενη πληροφορία, αυτή που ικανοποιεί την πληροφοριακή ανάγκη του χρήστη, δεν είναι πάντα και αναγκαία, ο χρήστης μπορεί να ήταν ήδη γνώστης της. Το ΣΑΠ δεν μπορεί να κάνει διάκριση μεταξύ των δύο.

1.3 Ακρίβεια και ανάκτηση

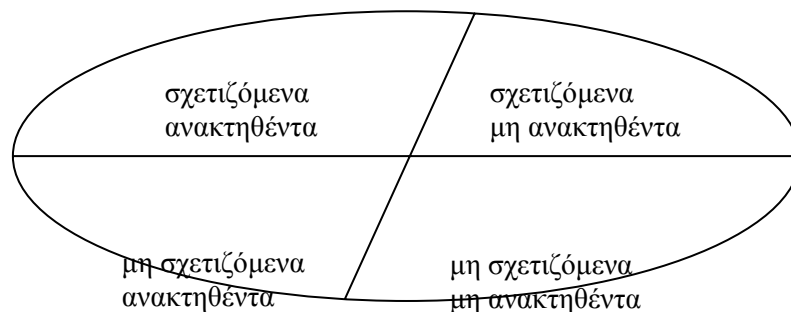
Υπάρχουν δυο μετρικές, άρρηκτα συσχετισμένες με τα ΣΑΠ, η *ακρίβεια* (*precision*) και η *ανάκτηση* (*recall*). Ορίζονται ως εξής :

$$\text{Ακρίβεια} = \frac{\text{Αριθμός σχετιζόμενων εγγράφων που ανακτώνται}}{\text{Συνολικός αριθμός εγγράφων που ανακτώνται}} \quad (1)$$

$$\text{Ανάκτηση} = \frac{\text{Αριθμός σχετιζόμενων εγγράφων που ανακτώνται}}{\text{Συνολικός αριθμός σχετιζόμενων εγγράφων στη συλλογή}} \quad (2)$$

Και οι δύο μετρικές παίρνουν τιμή στο διάστημα $[0,1]$.

Όταν ο χρήστης διατυπώνει μια ερώτηση στο ΣΑΠ, μετά την επεξεργασία της, η συλλογή που περιέχει τα έγγραφα έχει μια εικόνα ως εξής (Σχήμα 1.1):



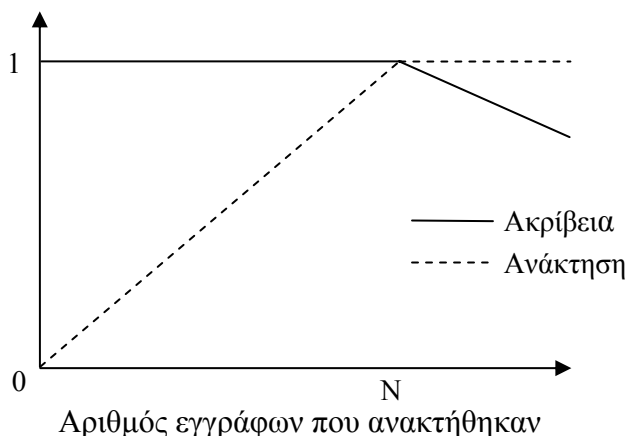
Σχήμα 1.1: Κατηγοριοποίηση εγγράφων της συλλογής σε σχέση με την ερώτηση

Το σύνολο των εγγράφων αποτελεί τα *σχετιζόμενα* και *μη σχετιζόμενα* έγγραφα. Από αυτά, ένα υποσύνολο από σχετιζόμενα και μη έγγραφα θα ανακτηθούν μετά την εκτέλεση της ερώτησης του χρήστη. Έτσι, τα έγγραφα που περιέχει η βάση χωρίζονται στα σχετιζόμενα ανακτηθέντα, μη σχετιζόμενα ανακτηθέντα, σχετιζόμενα μη ανακτηθέντα και μη σχετιζόμενα μη ανακτηθέντα.

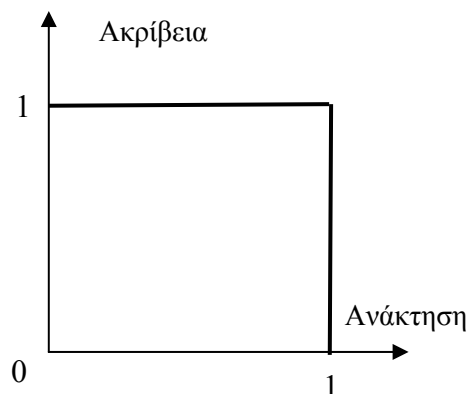
Η ακρίβεια είναι μετρική που υποδηλώνει την ποιότητα της απάντησης του χρήστη σε σχέση με το σύνολο των εγγράφων. Για παράδειγμα, μια ερώτηση που θα έφερνε 7 σχετιζόμενα έγγραφα σε σύνολο 10 ανακτηθέντων, θα είχε ακρίβεια 70%, ωστόσο παίζει ρόλο και ο συνολικός αριθμός των σχετιζόμενων εγγράφων. Αν στο συγκεκριμένο παράδειγμα ο αριθμός αυτός ήταν 7, τότε το σύστημα θα είχε πλήρη επιτυχία (η ανάκτηση θα

ήταν 100%), αν όμως ο αριθμός αυτός ήταν για παράδειγμα 200, τότε η ανάκτηση θα είχε τιμή 3,5%, θα περίμενε δηλαδή κανείς να έχει μεγαλύτερο αριθμό σχετιζόμενων ανακτηθέντων εγγράφων.

Η ανάκτηση εκφράζει την πιθανότητα με την οποία ένα κείμενο που εξυπηρετεί την πληροφοριακή ανάγκη του χρήστη ανήκει στα ανακτηθέντα κείμενα. Η μετρική της ανάκτησης εξαρτάται από το συνολικό αριθμό των σχετιζόμενων εγγράφων, κάτι που στην πραγματικότητα δεν μπορεί να υπολογιστεί με (ικανοποιητική) ακρίβεια. Αν στο προηγούμενο παράδειγμα το σύνολο των σχετιζόμενων εγγράφων ήταν 20, τότε για τα πρώτα 10 αποτελέσματα, η ανάκτηση θα είχε τιμή $7/20 = 0.35 = 35\%$ για ακρίβεια 70%.



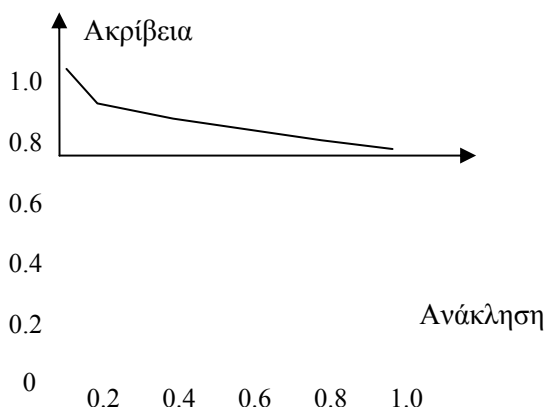
Σχήμα 1.2: Ιδανική καμπύλη ακρίβειας-ανάκτησης σε σχέση με τα ανακτηθέντα έγγραφα



Σχήμα 1.3: Ακρίβεια και ανάκτηση σε ΣΑΠ πλήρους ταύτισης

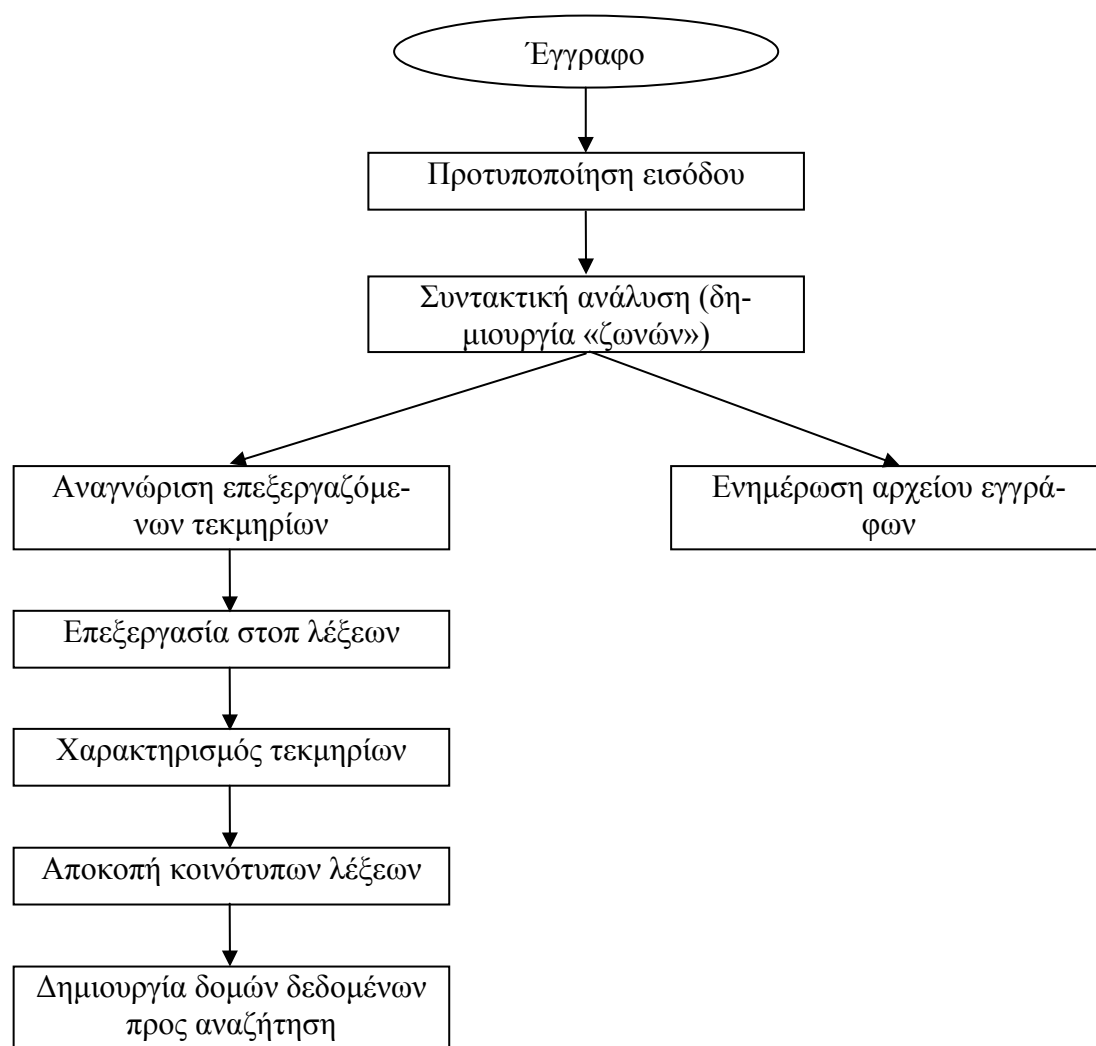
Η ακρίβεια εκφράζει την πιθανότητα με την οποία ένα κείμενο εξυπηρετεί την πληροφοριακή ανάγκη του χρήστη.

Η ανάκτηση προσεγγίζει τη μονάδα όσο ανακτώνται σχετιζόμενα έγγραφα και γίνεται ίση με αυτήν, από τη στιγμή που ανακτηθούν όλα. Αντίθετα, η ακρίβεια μπορεί να μεταβάλλεται συνέχεια από 0 έως και 1. Η ιδανική καμπύλη ακρίβειας και ανάκτησης σε σχέση με τα ανακτηθέντα έγγραφα φαίνεται στο σχήμα 1.2. Το σχήμα δείχνει την ιδανική



Σχήμα 1.4: Ακρίβεια και ανάκτηση σε ΣΑΠ μερικής ταύτισης

καμπύλη της ακρίβειας σε σχέση με την ανάκτηση για ΣΑΠ πλήρους ταύτισης, ενώ το σχήμα 1.4 δείχνει την αντίστοιχη καμπύλη για ΣΑΠ μερικής ταύτισης.



Σχήμα 1.5: Κανονικοποίηση όρων

1.4 Ανασκόπηση λειτουργίας ενός ΣΑΠ

Ένα σύστημα ανάκτησης πληροφορίας έχει τέσσερις κύριες λειτουργίες, την *κανονικοποίηση όρων* (*item normalization*), την *επιλεκτική διανομή της πληροφορίας* (*selective dissemination of information*), την *αναζήτηση εγγράφων στη βάση δεδομένων* (*document database search*) και την *ευρετηριοποίηση* (*indexing*).

1.4.1 Κανονικοποίηση όρων

Η διαδικασία της κανονικοποίησης φαίνεται στο σχήμα 1.5. Το πρώτο βήμα είναι η κανονικοποίηση των όρων σε μια πρότυπη μορφή.

Η *προτυποποίηση εισόδου* περιλαμβάνει τη μετατροπή της *μορφοποίησης* (*format*) του εγγράφου σε μια μορφή που είναι αναγνωρίσιμη για το σύστημα. Ένα σύστημα μπορεί να έχει ένα ή πολλά είδη μορφοποίησης για τα έγγραφα εισόδου. Ένα παράδειγμα της διαδικασίας της προτυποποίησης θα μπορούσε να είναι η μετατροπή της γλώσσας των εγγράφων εισόδου σε Unicode¹. Κάθε γλώσσα έχει διαφορετική εσωτερική κωδικοποίηση

¹ Το Unicode είναι σύστημα παράστασης χαρακτήρων βασισμένο σε δύο bytes ανά χαρακτήρα, που έχει τη δυνατότητα αναπαράστασης 65.536 χαρακτήρων, μπορεί δηλαδή να αποθηκεύσει τους χαρακτήρες όλων σχεδόν των γλωσσών.

για τους χαρακτήρες της. Ένα άλλο πρότυπο που καλύπτει τα Αγγλικά, Γαλλικά, Ισπανικά, κά είναι το ISO². Έχοντας λοιπόν μετατρέψει τα έγγραφα εισόδου σε παράσταση Unicode, η απεικόνιση μπορεί να γίνει από ένα πρόγραμμα πλοήγησης (*browser*) και η αναζήτηση να γίνει μία φορά από το σύστημα.

Η επόμενη διαδικασία είναι η *συντακτική ανάλυση* (*parsing*) των όρων εισόδου. Οι όροι υποδιαιρούνται σε λογικές υποδιαιρέσεις που έχουν νόημα για το χρήστη. Αυτή η διαδικασία είναι ορατή για το χρήστη και χρησιμοποιείται για να αυξήσει την ακρίβεια μιας αναζήτησης. Ο όρος υποδιαιρείται σε ζώνες (*zones*) που μπορούν να επικαλύπτονται και να είναι ιεραρχικές, όπως *τίτλος*, *συγγραφέας*, *περίληψη*, *κυρίως κείμενο*, *συμπέρασμα* και *παραπομπές*. Ο όρος «ζώνη» χρησιμοποιείται λόγω του μεταβλητού μεγέθους των δεδομένων που χαρακτηρίζει. Στη συνέχεια, οι πληροφορίες ζώνης χρησιμοποιούνται από τη διαδικασία αναγνώρισης των τεκμηρίων περιορίζοντας έτσι τις αναζητήσεις να γίνονται μόνο σε μια συγκεκριμένη ζώνη. Για παράδειγμα, αν ο χρήστης ενδιαφέρεται για άρθρα γραμμένα από τον Αϊνστάιν, η αναζήτηση δεν θα έπρεπε να περιλαμβάνει τη ζώνη *βιβλιογραφία* η οποία θα μπορούσε να έχει παραπομπές σε άρθρα του Αϊνστάιν.

Από τη στιγμή που γίνει η αναζήτηση, ο χρήστης θα θέλει να δει το αποτέλεσμα του, για παράδειγμα, σε μια οθόνη. Για να μεγιστοποιηθεί ο αριθμός των εγγράφων του αποτελέσματος στην οθόνη, θα πρέπει να εμφανίζεται ένας ελάχιστος αριθμός απαραίτητων δεδομένων για κάθε έγγραφο ώστε να μπορεί ο χρήστης να αποφασίσει κατά πόσο αυτό είναι σχετιζόμενο με την ερώτησή του. Πολύ συχνά, ο χρήστης θα βλέπει ζώνες όπως *τίτλος* και *συγγραφέας* ή και *περίληψη* επιτρέποντας έτσι την ανάγνωση πολλών εγγράφων. Ο χρήστης θα μπορεί επιλέγοντας ένα έγγραφο να το επεκτείνει και να δει το πλήρες κείμενο.

Μετά τις διαδικασίες της προτυποποίησης και συντακτικής ανάλυσης, πρέπει να αναγνωρισθεί σε κάθε όρο πληροφορία που χρησιμοποιείται κατά τη διαδικασία αναζήτησης. Ο όρος *τεκμήριο* (*token*) χρησιμοποιείται έναντι του όρου *λέξη* (*word*) διότι η λέξη δεν είναι η πιο αποτελεσματική μονάδα για να βασιστούν δομές αναζήτησης. Το πρώτο βήμα για την αναγνώριση των επεξεργαζόμενων τεκμηρίων είναι η αναγνώριση των λέξεων. Το σύστημα αναγνωρίζει τις λέξεις χωρίζοντας την είσοδο σε τρεις κατηγορίες : στα *έγκυρα σύμβολα λέξεων* (*valid word symbols*), στα *σύμβολα μεταξύ λέξεων* (*inter-word symbols*) και στα *ειδικά σύμβολα* (*special symbols*). Μια λέξη ορίζεται ως ένα συνεχές σει συμβόλων λέξεων που χωρίζονται από τα μεταξύ λέξεων σύμβολα. Σε πολλά συστήματα, τα μεταξύ λέξεων σύμβολα είναι σύμβολα που δεν συμμετέχουν σε αναζητήσεις. Σύμβολα λέξεων είναι οι αλφαβητικοί και αριθμητικοί χαρακτήρες. Παραδείγματα πιθανών μεταξύ λέξεων συμβόλων είναι το κενό, η τελεία και η άνω τελεία. Ο ακριβής ορισμός των μεταξύ λέξεων συμβόλων εξαρτάται από τη γλώσσα των όρων εισόδου. Για παράδειγμα, η απόστροφος μπορεί να είναι μικρής σημασίας αν χρησιμοποιείται για την περίπτωση κτητικού στα Αγγλικά, ωστόσο είναι σημαντική για την αναπαράσταση ονομάτων. Βάσει της απαιτούμενης ακρίβειας των αναζητήσεων και των χαρακτηριστικών της γλώσσας γίνεται ένας συγκερασμός στην επιλογή των μεταξύ λέξεων συμβόλων. Η τρίτη κατηγορία περιλαμβάνει ειδικά σύμβολα. Για παράδειγμα, η παύλα μπορεί να χρησιμοποιηθεί ανάλογα με τη διάθεση του συγγραφέα. Στο τέλος της γραμμής χρησιμοποιείται για να δείξει ότι μια λέξη συνεχίζει στην επόμενη γραμμή (συλλαβισμός), ενώ σε άλλη περίπτωση μπορεί να χρησιμοποιηθεί για να αποφευχθεί η σύγχυση στις λέξεις, όπως στην περίπτωση 'small business men', αντί του σωστού 'small-business men'. Όταν συ-

² ISO (International Standards Organization) : Το ISO πρότυπο για τις κωδικοσελίδες χρησιμοποιεί ένα byte ανά χαρακτήρα, με αποτέλεσμα να μπορούν να αναπαρασταθούν το πολύ 256 χαρακτήρες ανά κωδικοσελίδα. ISO 8859-7 είναι η επίσημη κωδικοσελίδα για τα ελληνικά, ενώ το ISO 8859-1 είναι η κωδικοσελίδα για την υποστήριξη των δυτικο-ευρωπαϊκών γλωσσών.

ναντάται ένα ειδικό σύμβολο, εκτελούνται κάποιοι αλγόριθμοι που αποφασίζουν ποιο θα είναι το επεξεργαζόμενο τεκμήριο.

Αφού αναγνωρισθούν τα τεκμήρια επεξεργασίας, εφαρμόζεται σε αυτά ο αλγόριθμος για τις *στοπ λέξεις* (*stop words*). Οι στοπ λέξεις είναι κοινότυπες λέξεις που εμφανίζονται πολύ συχνά σε ένα κείμενο, όπως το *the*, *my*, *play*, *a*, κτλ., και εξαιτίας ακριβώς αυτού του γεγονότος δεν πρέπει να ευρετηριοποιούνται. Εξαίρεση αποτελούν ορισμένες γλώσσες, όπως η εβραϊκή και η γαλλική, στις οποίες οι στοπ λέξεις παίζουν ρόλο και χωρίς αυτές αλλάζει το νόημα. Ο σκοπός αυτής της διαδικασίας είναι να εξοικονομηθούν πόροι του συστήματος ώστε το σετ των τεκμηρίων προς αναζήτηση να μην περιλαμβάνει αυτά που έχουν μικρή αξία. Οι λίστες με τις στοπ λέξεις υπάρχουν στα ΣΑΠ συνήθως με τη μορφή αρχείων. Οποιαδήποτε λέξη που υπάρχει σε όλα σχεδόν τα έγγραφα δεν έχει καμιά διακριτική αξία σε μια αναζήτηση. Μέρη του λόγου όπως τα άρθρα δεν έχουν καμιά αξία στην αναζήτηση και δεν αποτελούν χρήσιμα μέρη της ερώτησης του χρήστη. Παραλείποντας αυτές τις συχνά εμφανιζόμενες λέξεις το σύστημα εξοικονομεί τόσο σε επεξεργαστική όσο και σε αποθηκευτική ισχύ.

Ο Zipf διατύπωσε τον εξής κανόνα. Εξετάζοντας τη συχνότητα εμφάνισης των λέξεων σε μια συλλογή εγγράφων, η πλειοψηφία τους βρίσκεται να εμφανίζεται λίγες φορές. Ο νόμος της διαβάθμισης-συχνότητας του Zipf έχει ως εξής :

Συχνότητα * Διαβάθμιση = σταθερά,

όπου συχνότητα είναι ο αριθμός εμφάνισης μιας λέξης στο κείμενο και διαβάθμιση είναι η σειρά διαβάθμισής της, π.χ. η πλέον υψηλόσυχη έχει διαβάθμιση 1, η επόμενη 2, κτλ.

Ένας αριθμός λέξεων που έχουν εξαιρετικά μικρή συχνότητα εμφάνισης, για παράδειγμα μία ή δύο φορές, στη συλλογή είναι πολύ πιθανό να μην ανήκουν στο λεξιλόγιο του χρήστη, με αποτέλεσμα οι όροι αυτοί σπάνια να συμπεριλαμβάνονται στις αναζητήσεις. Η απαλοιφή τέτοιου είδους λέξεων εξοικονομεί αποθηκευτικό χώρο και δυσκολία στη δομή προσπέλασης. Η καλύτερη τεχνική για να απαλειφθεί η πλειοψηφία αυτών των λέξεων είναι η εφαρμογή ενός αλγορίθμου στοπ λέξεων. Παραδείγματα τέτοιου αλγορίθμου θα μπορούσαν να είναι τα εξής :

- Απαλοιφή όλων των αριθμών που είναι μεγαλύτεροι του «99999» επιτρέποντας έτσι την αναζήτηση ημερομηνιών
- Απαλοιφή τεκμηρίων που περιέχουν αναμειγμένους αλφαβητικούς και αριθμητικούς χαρακτήρες

Η επόμενη ενέργεια είναι ο *χαρακτηρισμός των τεκμηρίων*. Ο χαρακτηρισμός των τεκμηρίων βοηθά στην αποσαφήνιση μιας λέξης. Για παράδειγμα, *white* (λευκός) και *Mr. White* (ένα πρόσωπο). Ένα άλλο παράδειγμα είναι η διατήρηση των κεφαλαίων-πεζών. Σε πολλά συστήματα δεν υπάρχει διάκριση μεταξύ κεφαλαίων-πεζών για να αποφεύγεται έτσι η επαύξηση του όρου. Ωστόσο, για κύρια ονόματα, ακρωνύμια, οργανισμούς, κτλ. πρέπει να διατηρούνται τα κεφαλαία-πεζά, γιατί οι λέξεις αυτές μπορεί να έχουν τελείως διαφορετικό νόημα ανάλογα με το αν γράφονται με κεφαλαία ή πεζά. Άλλο είδος χαρακτηρισμού είναι οι αριθμοί και ημερομηνίες, τεκμήρια που επεξεργάζονται ξεχωριστά από το κείμενο.

Από τη στιγμή που εφαρμοστούν οι κύριοι κανόνες επεξεργασίας, αναγνωριστούν και χαρακτηριστούν τα τεκμήρια, το σύστημα εφαρμόζει τη διαδικασία της *αποκοπής κοινότυπων καταλήξεων* (*stemming*). Η διαδικασία αυτή κανονικοποιεί τα τεκμήρια σε μια στάνταρ μορφή. Η απόφαση για την εφαρμογή της τεχνικής αυτής αποτελεί ένα συγκεκριμένο μεταξύ της ακρίβειας της αναζήτησης έναντι της στάνταρ μορφής του τεκμηρίου που έχει ως αποτέλεσμα την ελάττωση του κόστους επεξεργασίας κατά την διαδικασία

επαύξησης του όρου σε παρόμοιες αναπαραστάσεις τεκμηρίων που αυξάνει το βαθμό ανάκτησης. Για παράδειγμα, το σύστημα θα πρέπει να διατηρεί ενικό, πληθυντικό, αόριστο χρόνο, κτλ. ως ξεχωριστά τεκμήρια με τα οποία μπορεί να γίνει αναζήτηση, η οποία γίνεται επαυξάνοντας τον όρο αναζήτησης με όλες τις δυνατές αναπαραστάσεις τεκμηρίων, ή πολύ απλά το σύστημα μπορεί να διατηρεί τον «κορμό» της λέξης, απαλείφοντας τις καταλήξεις. Ποσοτικά, η διαδικασία αυτή μπορεί να εφαρμοστεί σε τέτοιο βαθμό ώστε να γίνεται ανάκτηση μεγάλου σχετικά βαθμού μη σχετιζόμενων εγγράφων. Μερικά συστήματα αντί να εφαρμόζουν ένα εξελιγμένο αλγόριθμο χρησιμοποιούν λεξικά ή θησαυρούς για να εξακριβώσουν τον «κορμό» του τεκμηρίου. Παράδειγμα εφαρμογής της διαδικασίας αυτής θα αποτελούσε η εξαγωγή του *electric*- από τις λέξεις *electrical*, *electricity*, κτλ.

Τελευταία διαδικασία κατά την κανονικοποίηση είναι η χρησιμοποίηση των τεκμηρίων που έχουν προκύψει από την προηγούμενη διαδικασία της αποκοπής των κοινών καταλήξεων για την ενημέρωση της δομής αναζήτησης. Η δομή αυτή είναι η εσωτερική αναπαράσταση των όρων στους οποίους γίνεται αναζήτηση, περιέχει τη σημασιολογία των όρων της συλλογής και περιορίζει το χρήστη στο τι βρίσκει ως αποτέλεσμα μιας ερώτησης.

1.4.2 Επιλεκτική διανομή της πληροφορίας

Η επιλεκτική διανομή της πληροφορίας έχει τη δυνατότητα να συγκρίνει το περιεχόμενο εισερχόμενου κειμένου σε σχέση με ένα πεδίο ενδιαφερόντων των χρηστών στο σύστημα και να το «μοιράζει» στους χρήστες των οποίων το πεδίο ενδιαφερόντων συμπίπτει με αυτό του εισερχόμενου κειμένου. Αυτή η διαδικασία είναι και γνωστή ως ταχυδρομείο. Η διαδικασία του ταχυδρομείου αποτελείται από τη διαδικασία αναζήτησης, τις δηλώσεις ενδιαφερόντων των χρηστών (προφίλ) και τα αρχεία ταχυδρομείου των χρηστών. Για κάθε νέο κείμενο που εισέρχεται στο σύστημα γίνεται σύγκριση με το προφίλ κάθε χρήστη. Το προφίλ του χρήστη περιλαμβάνει μια ευρεία ερώτηση μαζί με ένα κατάλογο με τα αρχεία ταχυδρομείου του χρήστη. Αν το νέο κείμενο ικανοποιεί την ερώτηση στο προφίλ του χρήστη, τότε το κείμενο παραδίδεται στο ταχυδρομείο του χρήστη. Οι ερωτήσεις αναζήτησης στα προφίλ των χρηστών διαφέρουν από τις *ad-hoc* ερωτήσεις³ στο ότι περιέχουν ένα πολύ μεγαλύτερο αριθμό όρων (δέκα ή και εκατό φορές περισσότερους) προς αναζήτηση και ότι καλύπτουν ένα πιο ευρύ πεδίο ενδιαφερόντων. Τα προφίλ ορίζουν το πεδίο ενδιαφερόντων του χρήστη, ενώ οι *ad-hoc* ερωτήσεις αναφέρονται στην απάντηση μιας συγκεκριμένης ερώτησης μόνο.

Όταν η ερώτηση ικανοποιείται, το εισερχόμενο κείμενο μοιράζεται στα αρχεία του ταχυδρομείου του χρήστη που είναι συσχετισμένα με το προφίλ του. Τα νέα μηνύματα του χρήστη τοποθετούνται συνήθως σε σειρά χρόνου άφιξης και σβήνονται μετά το πέρας καθορισμένου χρονικού διαστήματος ή κατά εντολή του χρήστη.

1.4.3 Αναζήτηση εγγράφων στη βάση δεδομένων

Η διαδικασία αυτή προσφέρει την αναζήτηση έναντι μιας ερώτησης στη συλλογή όλων των καταχωρημένων κειμένων του συστήματος. Ο χρήστης εισάγει μια ερώτηση (*ad-hoc* ερώτηση) και του παραδίδεται το αποτέλεσμα όταν περατωθεί η αναζήτηση από το σύστημα. Αναζήτηση πληροφορίας που έχει ήδη επεξεργαστεί από το σύστημα θεωρείται αναζήτηση με αναδρομική ισχύ. Αυτό σημαίνει ότι η αναζήτηση μπορεί να γίνει σε περιορισμένο αριθμό κειμένων που εισήχθησαν στο σύστημα τις τελευταίες ώρες. Συνή-

³ *Ad-hoc* λέγονται οι ερωτήσεις που δίνει ο χρήστης με διαδραστικό (*interactive*) τρόπο.

θως όμως οι αναζητήσεις επεκτείνονται σε μεγαλύτερο χρονικό εύρος. Οι ad-hoc ερωτήσεις διαφέρουν από τα προφίλ στο ότι τυπικά είναι πολύ πιο μικρές και περιορίζονται σε συγκεκριμένο εύρος θεμάτων. Η συλλογή των εγγράφων μπορεί να είναι τεράστια, με εκατομμύρια έγγραφα. Από τη στιγμή που θα εισαχθούν στη βάση, τα έγγραφα είναι συνήθως στατικά. Έτσι η αξία της πληροφορίας μειώνεται με το χρόνο. Βάσει αυτού του γεγονότος, συχνά η βάση *διαμερίζεται (partitioned)* και ευρετηριοποιείται με βάση το χρόνο. Σε μερικά συστήματα, ο χρήστης αναγκάζεται να κάνει αναζήτηση με χρονικά περithώρια, κάνοντας έτσι χρήση των διαμερισμάτων χρόνου της βάσης.

1.4.4 Ευρετηριοποίηση

Όταν ένα κείμενο αποφασίζεται ότι ενδιαφέρει ένα χρήστη είναι στην κρίση του να αποφασίσει να το αποθηκεύσει για μελλοντική αναφορά. Αυτή η διαδικασία ονομάζεται αρχειοθέτηση. Σε ένα σύστημα ανάκτησης πληροφορίας, αυτό γίνεται με τη διαδικασία της ευρετηριοποίησης. Με αυτή τη διαδικασία, ο χρήστης μπορεί να αποθηκεύσει λογικά το έγγραφο στο αρχείο με πρόσθετους όρους και περιγραφικό κείμενο που θέλει να συσχετίσει με το έγγραφο. Είναι επίσης δυνατό να υπάρχουν εγγραφές στο *ευρετήριο (index)* που δεν αναφέρονται σε κάποιο έγγραφο, αλλά στο ίδιο το ευρετήριο αποθηκεύονται σημαντικές πληροφορίες του εγγράφου. Σε αυτή την περίπτωση, ο χρήστης διαβάσει άρθρα και εξάγει την πληροφορία που τον ενδιαφέρει, χωρίς να χρειάζεται να κοιτάξει στο αρχέτυπο έγγραφο. Από μια άλλη οπτική γωνία, το ευρετήριο μπορεί να θεωρηθεί ως δομημένη βάση, της οποίας οι εγγραφές μπορούν προαιρετικά να αναφέρονται σε εγγραφές της συλλογής των εγγράφων. Η διαδικασία της ευρετηριοποίησης έχει τη δυνατότητα δημιουργίας ευρετηρίων και αναζήτησης με τη βοήθειά τους. Ο χρήστης μπορεί να ψάχνει στο ευρετήριο και/ή τα έγγραφα στα οποία αναφέρεται.

Υπάρχουν δυο κατηγορίες ευρετηρίων : τα *δημόσια* και τα *ιδιωτικά* ευρετήρια. Κάθε χρήστης μπορεί να έχει ένα ή περισσότερα ιδιωτικά ευρετήρια που παραπέμπουν σε ένα μεγάλο αριθμό εγγράφων. Κάθε ιδιωτικό ευρετήριο αναφέρεται σε ένα μικρό υποσύνολο εγγράφων από το σύνολο στη βάση. Τα δημόσια ευρετήρια συντηρούνται από ειδικευμένο προσωπικό και τυπικά ευρετηριοποιούν όλα τα έγγραφα της συλλογής. Υπάρχει ένας μικρός αριθμός δημόσιων ευρετηρίων. Στα δημόσια ευρετήρια υπάρχει επίσης και *λίστα ελέγχου πρόσβασης (access control list)*, η οποία καθορίζει τους χρήστες και τα δικαιώματά τους στα ευρετήρια. Τα ιδιωτικά ευρετήρια συνήθως έχουν μικρές λίστες ελέγχου πρόσβασης.

Για να βοηθηθούν οι χρήστες στη δημιουργία των ευρετηρίων, το σύστημα προσφέρει τη διαδικασία της *αυτόματης δημιουργίας αρχείων (automatic file building)*. Η διαδικασία αυτή επεξεργάζεται τα εισερχόμενα έγγραφα και καθορίζει την ευρετηριοποίηση. Οι κανόνες που καθορίζουν ποια αρχεία θα επεξεργαστούν και πώς θα γίνει η ευρετηριοποίηση καθορίζονται από αρχεία-προφίλ που χρησιμοποιούνται από την παραπάνω διαδικασία. Κατά τη διαδικασία αυτή, δημιουργούνται υποψήφιες εγγραφές δεικτών, οι οποίες ωστόσο δεν καταχωρούνται στο κανονικό ευρετήριο. Αυτό γίνεται κατόπιν εντολής του χρήστη και αφού αναθεωρήσει τις εγγραφές αυτές. Παράλληλα, υπάρχουν δεδομένα που βοηθούν στη δημιουργία των δημοσίων ευρετηρίων, όπως είναι ο συγγραφέας του άρθρου, η ημερομηνία έκδοσης, ο εκδοτικός οίκος, κτλ.

1.5 Συσχέτιση με τα συστήματα διαχείρισης βάσεων δεδομένων

Υπάρχουν δυο μεγάλες κατηγορίες συστημάτων διαθέσιμα για την επεξεργασία άρθρων. Τα συστήματα ανάκτησης πληροφορίας και τα *συστήματα διαχείρισης βάσεων*

δεδομένων (*database management systems*). Η διαφορά μεταξύ των δυο έγκειται στον τύπο της πληροφορίας που το καθένα επεξεργάζεται. Το ΣΑΠ είναι το σύστημα εκείνο που έχει τις δυνατότητες διαχείρισης ημι-δομημένου τύπου πληροφορία, ενώ το σύστημα διαχείρισης βάσης δεδομένων (Σ.Δ.Β.Δ. χάριν συντομίας) διαχειρίζεται δομημένου τύπου πληροφορία.

Δομημένο κείμενο είναι καλώς ορισμένο κείμενο που αναπαρίσταται σε σχεσιακούς πίνακες. Υπάρχει μια περιγραφή που συσχετίζεται και ορίζει κάθε γνώρισμα (*attribute*) του πίνακα. Για παράδειγμα, δε δημιουργείται σύγχυση μεταξύ των χαρακτηριστικών «όνομα υπαλλήλου» και «μισθός υπαλλήλου» εφόσον κάθε εγγραφή στη βάση είναι καλώς ορισμένη σχετικά με το είδος των τιμών που μπορεί να δεχθεί. Το όνομα του υπαλλήλου είναι πεδίο αλφαριθμητικό, ενώ ο μισθός είναι πεδίο αριθμητικό.

Όταν ο χρήστης έχει να κάνει με δομημένη πληροφορία, εισάγει μια συγκεκριμένη ερώτηση και το σύστημα του επιστρέφει το αποτέλεσμα. Ο τρόπος εμφάνισης του αποτελέσματος ίσως να μην έχει και πολύ σημασία, αλλά συνήθως το βλέπει σε μορφή πίνακα για λόγους ευκολίας. Σε αντίθεση, αναζήτηση σε πληροφορία ενός ΣΑΠ μπορεί να μην επιστρέψει όλα τα έγγραφα που ικανοποιούν την ερώτηση του χρήστη. Είναι πολύ πιθανό, ο χρήστης να πρέπει να *τροποποιήσει* (*refine*) την ερώτησή του προκειμένου να εντοπίσει και άλλα έγγραφα. Ένα ΣΑΠ βοηθά το χρήστη να βρει αυτό που τον ενδιαφέρει με διάφορες μεθόδους, μία εκ των οποίων είναι η *επανατροφοδότηση αξίας καταλληλότητας* (*relevance feedback*). Τα έγγραφα παρουσιάζονται στο χρήστη ταξινομημένα με *σειρά διαβάθμισης* (*relevance rank*). Παρόλο που η υλοποίηση αποθήκευσης τέτοιου είδους πληροφορίας είναι εύκολη για ένα Σ.Δ.Β.Δ., ωστόσο ακριβώς επειδή το Σ.Δ.Β.Δ. είναι σχεδιασμένο για τη διαχείριση δομημένης πληροφορίας, δεν έχει τις εξελιγμένες δυνατότητες παρουσίασης του αποτελέσματος με σειρά διαβάθμισης, που είναι σημαντικές για ένα ΣΑΠ.

1.6 Ψηφιακές βιβλιοθήκες και αποθήκες δεδομένων

Δύο άλλες κατηγορίες συστημάτων που έχουν να κάνουν με την αποθήκευση και ανάκτηση της πληροφορίας είναι οι *ψηφιακές βιβλιοθήκες* (*digital libraries*) και οι *αποθήκες δεδομένων* (*data warehouses*). Τα ΣΑΠ, οι ψηφιακές βιβλιοθήκες και οι αποθήκες δεδομένων αποτελούν «αποθήκες» πληροφορίας και κύριος σκοπός τους είναι η ικανοποίηση πληροφοριακών αναγκών. Βιβλιοθήκες υπήρχαν από τότε που άρχισε ο άνθρωπος να γράφει και εξυπηρετούν ως αποθήκες του πνευματικού έργου της κοινωνίας. Ακριβώς για αυτό το λόγο, οι βιβλιοθήκες πάντοτε ενδιαφέρονταν για την αποθήκευση και την ανάκτηση δεδομένων. Όσο η ποσότητα της πληροφορίας αύξανε με εκθετικό τρόπο, οι βιβλιοθήκες αναγκάζονταν να μεγιστοποιούν τη χρήση ηλεκτρονικών μέσων για να εξυπηρετήσουν τις ανάγκες αποθήκευσης και ανάκτησης που είχαν. Παράλληλα, η ανάπτυξη των δικτύων και η διαδικτύωση των βιβλιοθηκών είχαν ως αποτέλεσμα να δοθεί περισσότερη προσοχή στην έννοια των ηλεκτρονικών βιβλιοθηκών, οι οποίες από ένα σημείο και μετά μετονομάστηκαν σε ψηφιακές βιβλιοθήκες.

Τι είναι όμως η ψηφιακή βιβλιοθήκη; Είναι σωστό να αρχίσει κανείς με την συσχέτιση με την κλασική βιβλιοθήκη. Το επόμενο βήμα είναι να δει πως εφαρμόζονται οι λειτουργίες μιας ψηφιακής βιβλιοθήκης σε μια ψηφιακή πλέον συλλογή κειμένων. Από τη στιγμή που τα κείμενα είναι σε ηλεκτρονική μορφή και είναι διαθέσιμα από όλες τις βιβλιοθήκες λόγω της δικτύωσής τους, η βιβλιοθήκη δε μπορεί πλέον να έχει την κυριότητα ενός κειμένου εφόσον το διαθέτει και σε άλλες βιβλιοθήκες. Η παρούσα τεράστια ποσότητα κειμένων σε χαρτί εγγυάται το γεγονός ότι δε θα υπάρχουν εξ' ολοκλήρου ψηφιακές βιβλιοθήκες για αρκετό καιρό ακόμη. Αλλά δεν υπάρχει ωστόσο αμφισβήτηση ότι οι βι-

βιβλιοθήκες έχουν αρχίσει προ πολλού να στρέφουν την προσοχή τους και σε τρόπους ηλεκτρονικής αποθήκευσης.

Η ευρετηριοποίηση είναι ένα από τα φλέγοντα ζητήματα που απασχολούν μια βιβλιοθήκη και γίνονται πολλές προσπάθειες για την καθιέρωση προτύπων ευρετηριοποίησης και δημιουργίας καταλόγων. Η μεταφορά πολλών υπηρεσιών βιβλιοθηκών σε ψηφιακή μορφή δημιουργεί προκλήσεις. Από τη στιγμή που το πλήρες κείμενο θα είναι σε ηλεκτρονική μορφή, η ευρετηριοποίηση αποκτά τεράστια σημασία. Μια από τις υπηρεσίες που προσφέρει μια βιβλιοθήκη είναι η ύπαρξη ανθρώπων που βοηθούν τους ενδιαφερόμενους να βρουν αυτό που ψάχνουν πιο εύκολα. Από τη στιγμή που τα κείμενα θα υπάρχουν σε ηλεκτρονική μορφή, ο ρόλος αυτός από ειδικός στην αναζήτηση θα μετατραπεί σε ειδικό στην ανάλυση του κειμένου. Η αναζήτηση πλέον θα επιστρέφει τόσα πολλά αποτελέσματα, που θα πρέπει να υπάρχει κάποιος που να αναγνωρίζει την αξία των πληροφοριών αυτών για το χρήστη. Αυτός θα είναι πλέον και ο ρόλος των μελλοντικών βιβλιοθηκών.

Ο τομέας της αποθήκευσης και ανάκτησης της πληροφορίας αντιμετωπίζει μέρος μόνο των θεμάτων που απασχολούν μια ψηφιακή βιβλιοθήκη. Το επίκεντρο της προσοχής είναι στην αναζήτηση και ανάκτηση δεδομένων κειμένου χωρίς να δοθεί προσοχή στην καθιέρωση προτύπων στα περιεχόμενα του συστήματος. Έχει επίσης αγνοηθεί το θέμα των πνευματικών δικαιωμάτων και το θέμα της μετατροπής της υπάρχουσας βιβλιογραφίας σε ηλεκτρονική μορφή, ενώ ούτε γίνεται λόγος για τη δημιουργία ενός προτύπου της μορφής αυτής έχοντας υπόψη και το γεγονός ότι τα πρότυπα εξελίσσονται.

Ο όρος αποθήκη δεδομένων προέρχεται κυρίως από τον εμπορικό τομέα και λιγότερο από τον ακαδημαϊκό. Μια αποθήκη δεδομένων χρησιμοποιεί τεχνολογία *OLAP* (*On Line Analytical Processing*), ενώ ένα Σ.Δ.Β.Δ. αποτελεί ένα *OLTP* (*On Line Transaction Processing*) σύστημα. Σκοπός των αποθηκών δεδομένων είναι να προσφέρουν κρίσιμες πληροφορίες στα άτομα που λαμβάνουν αποφάσεις για το μέλλον των επιχειρήσεων. Οι αποθήκες δεδομένων στηρίζονται σε μια *αρχιτεκτονική τριών σειρών* (*3-tier architecture*), που αποτελείται από ένα Σ.Δ.Β.Δ., μια πολυ-διάστατη μηχανή OLAP καθώς και εργαλεία ανάλυσης στον *πελάτη* (*client*). Διαπιστώνει κανείς ότι μια αποθήκη δεδομένων έχει στενή συσχέτιση με ένα Σ.Δ.Β.Δ.

Οι αποθήκες δεδομένων και τα ΣΑΠ έχουν κοινά σημεία στο ότι και τα δυο έχουν ανάγκη για αναζήτηση και ανάκτηση πληροφορίας. Ωστόσο, μια αποθήκη δεδομένων επικεντρώνεται πιο πολύ σε δομημένη πληροφορία και σε τεχνολογίες που βοηθούν στη λήψη αποφάσεων. Παράλληλα, υπάρχει και ο όρος *εξόρυξη δεδομένων* (*data mining*) που ονομάζεται η διαδικασία η οποία επεξεργάζεται τα δεδομένα, εξάγει συσχετίσεις και εξαρτήσεις που δεν ήταν αρχικά μέρος του σχεδιασμού. Το μεγαλύτερο μέρος της ερευνητικής προσπάθειας επικεντρώνεται στην στατιστική, στην αναγνώριση σχεδίων και σε αλγορίθμους τεχνητής νοημοσύνης για την ανακάλυψη κρυμμένων σχέσεων μεταξύ των δεδομένων.

ΚΕΦΑΛΑΙΟ 2

ΣΥΓΧΡΟΝΕΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ ΣΤΟ ΔΙΑΔΙΚΤΥΟ

2.1 Εισαγωγή

Το διαδίκτυο ξεκίνησε ως μια ακαδημαϊκή προσπάθεια επικοινωνίας των πανεπιστημίων όλου του κόσμου, ωστόσο κανείς ίσως δεν είχε φανταστεί τη τεράστια εξάπλωση που θα είχε. Στις μέρες μας, το διαδίκτυο χρησιμοποιείται από όλους, για όλων των ειδών τις ανάγκες, για ενημέρωση, για ψυχαγωγία, για δουλειά, κ.ά. Παραμένει φυσικά η δυνατότητα της επικοινωνίας, του πρωτεύοντα ρόλου δημιουργίας του διαδικτύου, ενώ έχει αρχίσει πλέον και η εισαγωγή του παράγοντα εμπορίου, μέσω των ηλεκτρονικών αγορών. Στην εξάπλωση του διαδικτύου, στην εισαγωγή του στα νοικοκυριά και στη χρησιμοποίησή του από χρήστες που λίγη σχέση έχουν με τους υπολογιστές και την πληροφορική, συνέβαλε και η εισαγωγή του *παγκόσμιου ιστού (world wide web)*, που απλοποίησε αρκετά τη χρήση του διαδικτύου μέσω των γραφικών *προγραμμάτων πλοήγησης (browsers)*.

Η δομή του διαδικτύου παραμένει μη ιεραρχική με αποτέλεσμα να είναι αρκετά δύσκολο για το χρήστη να εντοπίσει εύκολα την πληροφορία που τον ενδιαφέρει. Η δημιουργία νέων *τόπων (web sites)* και σελίδων με εκθετικό ρυθμό έκανε επιτακτική την ανάγκη ύπαρξης κάποιων τρόπων αναζήτησης μέσα στον κυκεώνα των πληροφοριών του διαδικτύου. Οι μηχανές αναζήτησης εξυπηρετούν αυτό το σκοπό. Διευκολύνουν το χρήστη να εντοπίσει την πληροφορία που τον ενδιαφέρει. Ίσως οι προσπάθειες να μην είναι πάντοτε επιτυχείς, μερικές φορές είναι πιθανό ο χρήστης να μην εντοπίσει την πληροφορία που θέλει, ωστόσο είναι σίγουρο ότι οι μηχανές αναζήτησης του διαδικτύου είναι πολύτιμες και χωρίς αυτές, η *πλοήγηση (browsing)* σε αυτό θα ήταν από αρκετά δύσκολη έως αδύνατη.

2.2 Ορισμός της μηχανής αναζήτησης

Η μηχανή αναζήτησης είναι ένας από τους κύριους τρόπους για την εύρεση τόπων στο διαδίκτυο. Είναι ίσως δύσκολο να δώσει κανείς ένα ακριβή ορισμό του τι είναι μια μηχανή αναζήτησης, μιας και οι σύγχρονες μηχανές αναζήτησης προσφέρουν πολλές πλέον υπηρεσίες, που ξεφεύγουν από τη στενή έννοια του τρόπου αναζήτησης πληροφοριών. Από μια πιο τεχνική σκοπιά, η μηχανή αναζήτησης είναι ένας τύπος που προσφέρει ένα πεδίο εισαγωγής κειμένου όπου ο χρήστης εισάγει την ερώτησή του, μια λέξη ή φράση, και αφού την αποστείλει, η μηχανή αναζήτησης την επεξεργάζεται και παρουσιάζει το αποτέλεσμα στο χρήστη. Στην ουσία η μηχανή αναζήτησης κοιτάει σε ένα *ευρετήριο (index)* στο οποίο υπάρχουν ευρετηριοποιημένοι τόποι και παρουσιάζει τους τόπους εκείνους που θεωρεί ότι ενδιαφέρουν το χρήστη.

2.3 Τρόπος λειτουργίας

Η μηχανή αναζήτησης δημιουργεί τις λίστες της αυτόματα. Αυτό το κάνει *έρποντας (crawl)* στο web και οι χρήστες ψάχνουν αυτά που έχει βρει η μηχανή. Εάν αλλάξουν μία ή περισσότερες σελίδες, η μηχανή εντοπίζει την αλλαγή και επαναευρετηριοποιεί τις νέες σελίδες.

Μια μηχανή αναζήτησης αποτελείται από τρία μέρη. Το πρώτο είναι η *αράχνη (spider)*, που ονομάζεται το μέρος εκείνο της μηχανής που «έρπει» στο web. Η «αράχνη» επισκέπτεται μια *ιστοσελίδα (web page)* και μετά ακολουθεί τους συνδέσμους που περιέχει η σελίδα. Αυτή η διαδικασία ονομάζεται *crawling*. Η «αράχνη» επιστρέφει στη σελίδα αυτή σε τακτά χρονικά διαστήματα, για παράδειγμα κάθε ένα ή δύο μήνες, για να εντοπίσει τυχόν αλλαγές.

Ότι βρει η «αράχνη» επιστρέφεται στο δεύτερο μέρος της μηχανής, το ευρετήριο. Το ευρετήριο, που πολλές φορές ονομάζεται και κατάλογος, μοιάζει με ένα τεράστιο βιβλίο που περιέχει κάθε σελίδα που έψαξε η «αράχνη». Εάν μια σελίδα αλλάξει, τότε το βιβλίο αυτό ενημερώνεται με τις νέες πληροφορίες. Αυτή η διαδικασία, η ενημέρωση του βιβλίου που στην ουσία είναι η ευρετηριοποίηση, μπορεί να αργήσει λίγο από τη λήψη των πληροφοριών της «αράχνης». Έτσι, ενώ μία σελίδα έχει επισκεφθεί από αυτήν, μπορεί να μην έχει συμπεριληφθεί ακόμη στο ευρετήριο και δεν είναι έτσι άμεσα διαθέσιμη από τους χρήστες που κάνουν αναζήτηση.

Το τρίτο μέρος είναι το λογισμικό που στηρίζει τη μηχανή. Είναι το πρόγραμμα που ψάχνει μεταξύ των εκατομμυρίων σελίδων που είναι ευρετηριοποιημένες για να βρει αυτές που ικανοποιούν μια συνθήκη και να τις διαβαθμίσει σε σχέση με αυτό που θεωρεί ότι είναι το πιο σχετιζόμενο.

2.4 Χαρακτηριστικά των μηχανών αναζήτησης

Παρόλο που όλες οι μηχανές αναζήτησης αποτελούνται από τα τρία μέρη που περιγράφονται παραπάνω, υπάρχουν διαφορές, που εξηγούν και το γεγονός ότι η ίδια αναζήτηση σε διαφορετικές μηχανές παράγει και διαφορετικά αποτελέσματα. Οι διαφορές αυτές έχουν σχέση με το ποια χαρακτηριστικά υποστηρίζει η κάθε μηχανή και με ποιο τρόπο αντιμετωπίζει τέσσερα κυρίως ζητήματα, που έχουν σχέση με το *crawling*, την *ευρετηριοποίηση*, τη *διαβάθμιση* και το *spawning*, που θα εξηγηθεί παρακάτω (§2.4.4). Παράλληλα, αναπτύσσεται και ο τρόπος με τον οποίο οι μηχανές κάνουν διάκριση μεταξύ κεφαλαίων-πεζών.

2.4.1 Crawling

Τα χαρακτηριστικά που υποστηρίζουν οι σύγχρονες μηχανές αναζήτησης¹ σε σχέση με το *crawling* έχουν να κάνουν με το πόσο καλά μια μηχανή ψάχνει τους τόπους.

Ένα χαρακτηριστικό που έχουν μηχανές όπως οι AltaVista, Google, Inktomi και Northern Light είναι το *deep crawl*, που έχει σχέση με το ότι οι μηχανές αυτές θα εμφανίσουν πολλές σελίδες από ένα τόπο, ακόμη κι αν οι σελίδες αυτές δεν αποσταλούν στις μηχανές για ευρετηριοποίηση. Μηχανές που δεν έχουν το χαρακτηριστικό αυτό, όπως οι Excite, Go και Lycos θα εμφανίσουν σαφώς λιγότερες σελίδες από τον τόπο. Σε γενικές γραμμές, όσο πιο μεγάλο είναι το μέγεθος του ευρετηρίου μιας μηχανής, τόσο περισσότερες σελίδες μπορεί να συμπεριλάβει.

Η *στιγμιαία ευρετηριοποίηση* έχει σχέση με το αν μία σελίδα που θα αποσταλεί στη μηχανή για ευρετηριοποίηση, θα ευρετηριοποιηθεί και θα είναι διαθέσιμη για αναζήτηση σε σύντομο σχετικά χρονικό διάστημα, για παράδειγμα μετά από μία ή δύο μέρες.

Η συμπεριφορά των μηχανών σε σχέση με τα *πλαίσια*² φαίνεται από το αν τα υποστηρίζουν. Σελίδες που περιέχουν πλαίσια είναι πιθανό να μην ευρετηριοποιηθούν από μία μηχανή.

Η υποστήριξη των *χαρτογραφημένων εικόνων (image maps)*³ έχει σχέση με το αν οι μηχανές μπορούν να ακολουθήσουν τους συνδέσμους μέσω αυτών.

¹ Τα χαρακτηριστικά που θα αναφερθούν για τις μηχανές αναζήτησης ισχύουν όπως έχουν μέχρι τις 14 Ιανουαρίου 2000.

² Τα πλαίσια (frames) είναι χαρακτηριστικό που εμφανίστηκε με το πρότυπο HTML έκδοση 3.2. Επιτρέπει τον χωρισμό της σελίδας σε περισσότερες, με τη βοήθεια πλαισίων. Αυτό δηλαδή που βλέπει ο χρήστης στην οθόνη είναι πολλές σελίδες ταυτόχρονα.

Η ύπαρξη ενός αρχείου με όνομα `robots.txt` έχει σχέση με το αν το άτομο που συντηρεί ένα τόπο (*webmaster*) θέλει να μην επιτρέψει μια μηχανή αναζήτησης να ψάξει τον τόπο του. Παράλληλα, αν μια σελίδα περιέχει την *meta παραπομπή* (*tag*) *robot* στον πηγαίο κώδικα της HTML, τότε δεν θα ευρετηριοποιηθεί.

Ένας ακόμη τρόπος για να διαπιστώσει μια μηχανή ποιες σελίδες πρέπει να ευρετηριοποιηθεί είναι με το να εξετάσει πόσο «δημοφιλής» είναι μια σελίδα από το να μετρήσει πόσοι σύνδεσμοι από άλλες σελίδες οδηγούν στη σελίδα αυτή.

Χαρακτηριστικό που έχει σχέση με τη συχνότητα ανανέωσης των σελίδων είναι το αν μια μηχανή «μαθαίνει» πόσο συχνά αλλάζουν οι σελίδες και ανάλογα με τη συχνότητα αυτή, η μηχανή αλλάζει και τη δική της συχνότητα με την οποία επισκέπτεται τις σελίδες αυτές.

Ένα ομολογουμένως χρήσιμο χαρακτηριστικό για τον σχεδιαστή ενός τόπου είναι να του επιτρέπει η μηχανή να μάθει αν ο τόπος του χρήστη υπάρχει ήδη στους καταλόγους της μηχανής. Για παράδειγμα, η μηχανή Go μπορεί να εμφανίσει την πληροφορία αυτή, αν ο χρήστης εισάγει στο πεδίο εισαγωγής κειμένου τη λέξη κλειδί ‘url:’, ακολουθούμενη από τη διεύθυνση του τόπου που ενδιαφέρει το χρήστη.

2.4.2 Ευρετηριοποίηση

Τα χαρακτηριστικά των μηχανών αναζήτησης σε σχέση με την ευρετηριοποίηση έχουν να κάνουν με το πώς ακριβώς γίνεται αυτή.

Η πρώτη δυνατότητα έχει να κάνει με την ευρετηριοποίηση όλου του ορατού κειμένου της σελίδας, αν και μερικές σελίδες μπορεί να μην ευρετηριοποιούν στοπ λέξεις ή λέξεις που θεωρούνται *spam*.

Ένα δεύτερο χαρακτηριστικό είναι ο αποκλεισμός των στοπ λέξεων, είτε εκτός ευρετηρίου, είτε αποκλεισμός τους κατά τη διάρκεια επεξεργασίας μιας ερώτησης. Λόγοι για τους οποίους αποκλείονται αυτές οι λέξεις έχουν να κάνουν με την ελάττωση σε αποθηκευτικό χώρο του ευρετηρίου και με την αύξηση της ταχύτητας των αναζητήσεων.

Εκτός από την *meta παραπομπή robot* που αναφέρθηκε παραπάνω, υπάρχει και η δυνατότητα υποστήριξης επιπλέον παραπομπών που βοηθούν στην εξαγωγή περίληψης της σελίδας από κάποιες μηχανές (*παραπομπή description*) και στη συμπερίληψη λέξεων κλειδιών που έχουν σχέση με τη σελίδα (*παραπομπή keywords*). Αυτές οι δύο παραπομπές, που είναι και οι κυριότερες, στην περίπτωση που υποστηρίζονται από τη μηχανή, μπορεί να παίξουν κάποιο ρόλο στη σειρά διαβάθμισης της σελίδας.

Υποστήριξη της παραπομπής ALT του HTML προτύπου και του κειμένου εντός σχολίων της σελίδας, που έχουν να κάνουν με την ευρετηριοποίηση ή όχι του εναλλακτικού κειμένου στη θέση μιας εικόνας, που εμφανίζεται αντί των εικόνων για προγράμματα πλοήγησης σε περιβάλλον χαρακτήρων όπως το δημοφιλές *lynx* του Unix, καθώς και την ευρετηριοποίηση του κειμένου εντός των σχολίων της σελίδας, τα οποία δεν εμφανίζονται στο χρήστη κατά τη διάρκεια εμφάνισής της.

Ένα τελευταίο χαρακτηριστικό σε σχέση με την ευρετηριοποίηση, είναι η υποστήριξη της *αποκοπής κοινότυπων καταλήξεων* (*stemming*), αν η μηχανή δηλαδή θα υποστηρίζει αναζήτηση σε σχέση με τη «ρίζα» της λέξης, για παράδειγμα αναζήτηση για ‘swim’ θα επιστρέψει και τα ‘swims’ ή ‘swimming’.

2.4.3 Διαβάθμιση

³ Image maps είναι εικόνες που περιέχουν πολλούς συνδέσμους. Ανάλογα με το που θα κάνει κλικ ο χρήστης, θα ακολουθήσει και διαφορετικό σύνδεσμο.

Οι περισσότερες μηχανές χρησιμοποιούν τη θέση και τη συχνότητα εμφάνισης μιας λέξης κλειδί σε μια σελίδα για να την κατατάξουν στη σειρά διαβάθμισης. Ωστόσο, ο ακριβής μηχανισμός είναι λίγο διαφορετικός για κάθε μηχανή. Υπάρχει μια σειρά χαρακτηριστικών με τα οποία η μηχανή τροποποιεί λίγο τη θέση κατάταξης της σελίδας στη σειρά διαβάθμισης.

Μερικές μηχανές χρησιμοποιούν τις *meta παραπομπές* για να δώσουν μια μικρή ώθηση στη σειρά διαβάθμισης της σελίδας. Αυτό όμως δε σημαίνει ότι όλες οι μηχανές που υποστηρίζουν τις *meta παραπομπές* θα δώσουν αυτή την ώθηση. Αυτό εξαρτάται και πάλι από τον τρόπο λειτουργίας της μηχανής.

Εάν ένας τόπος υπάρχει ήδη σε ένα κατάλογο, το γεγονός αυτό μπορεί να ωθήσει τον τόπο ψηλότερα στη σειρά διαβάθμισης.

Όπως περιγράφηκε πιο πάνω, το πόσο «δημοφιλής» είναι μια σελίδα μπορεί να τη βοηθήσει σε κάποιο βαθμό στη σειρά διαβάθμισης.

Επίσης, οι μηχανές είναι αρκετά έξυπνες στο να αλλάζουν τη σειρά διαβάθμισης των σελίδων ανάλογα με το σύνδεσμο που θα κάνει κλικ ο χρήστης στη λίστα αποτελέσματος.

2.4.4 Spamming

Οι περισσότερες μηχανές αναζήτησης υποβαθμίζουν ένα τόπο αν θεωρηθεί ότι δημιουργούν *spam* για να βελτιώσει τη θέση του στη σειρά διαβάθμισης. Το *spam* θεωρείται «θόρυβος», ένας «φτηνός» τρόπος με τον οποίο ο σχεδιαστής ενός τόπου μπορεί να αυξήσει τη δημοτικότητα των σελίδων του. Μια συνηθισμένη τακτική είναι η διαρκής επανάληψη των ίδιων λέξεων στη σελίδα. Αν η μηχανή εντοπίσει κάτι τέτοιο, τότε μπορεί να τοποθετήσει τον τόπο χαμηλότερα στη σειρά διαβάθμισης. Υπάρχουν χαρακτηριστικά που βοηθούν στον εντοπισμό του *spam*.

Μερικοί σχεδιαστές τόπων δημιουργούν σελίδες που οδηγούν αυτόματα τους χρήστες σε διαφορετικές σελίδες μέσα στον τόπο. Ένας τρόπος για να γίνει αυτό είναι η *meta παραπομπή refresh*. Μερικές μηχανές δε θα ευρετηριοποιήσουν τέτοιες σελίδες.

Μια άλλη τεχνική είναι να υπάρχει κείμενο του οποίου το χρώμα είναι ίδιο με το χρώμα του υποβάθρου ώστε να μην είναι ορατό στο χρήστη. Το κείμενο αυτό ονομάζεται *αόρατο κείμενο*. Πολλές μηχανές είτε δεν ευρετηριοποιούν κείμενο τέτοιου είδους, είτε όλη τη σελίδα που περιέχει το κείμενο.

Άλλη τεχνική παρόμοια με εκείνη του αόρατου κειμένου είναι κείμενο με πολύ μικρό μέγεθος γραμματοσειράς ώστε να μην είναι εύκολα ορατό στο χρήστη. Τέτοιες σελίδες μπορεί να θεωρηθεί ότι προκαλούν *spam* και είτε δεν ευρετηριοποιείται το συγκεκριμένο τμήμα κειμένου, είτε δεν ευρετηριοποιείται ολόκληρη η σελίδα.

2.4.5 Πώς αντιμετωπίζονται τα κεφαλαία και πεζά

Οι άνθρωποι ξοδεύουν αρκετό χρόνο σχετικά με τα κεφαλαία και πεζά, ειδικά σε ότι αφορά στις *meta παραπομπές*. Αυτό συμβαίνει επειδή μερικές μηχανές αναζήτησης κάνουν διάκριση μεταξύ κεφαλαίων-πεζών, ενώ άλλες όχι. Επειδή όμως ένας σχεδιαστής τόπου δεν θέλει να χάσει κίνηση, στις *meta παραπομπές* εισάγει παραλλαγές λέξεων, όπως :

entertainment, Entertainment, ENTERTAINMENT

Το πρόβλημα με τα παραπάνω είναι ότι αυτές οι επαναλήψεις της ίδιας λέξης μπορεί να προκαλέσουν το μηχανισμό για την αποφυγή *spamming* της μηχανής, ειδικά όταν

υπάρχουν φράσεις. Για παράδειγμα, μπορεί να φανταστεί κανείς τις παραλλαγές της φράσης ‘golf courses’ :

golf courses, Golf courses, Golf Courses, golf Courses, GOLF COURSES, GOLF courses, golf COURSES

που είναι λίγοι μόνο από τους συνδυασμούς που προκύπτουν, αν ο σχεδιαστής του τόπου αποφασίσει να συμπεριλάβει ενικό και πληθυντικό μαζί.

Αυτό που πρέπει λοιπόν να γίνει στην περίπτωση αυτή είναι να χρησιμοποιούνται τα πεζά, εφόσον το μεγαλύτερο ποσοστό των χρηστών χρησιμοποιεί πεζά και προς το παρόν τουλάχιστον είναι ελάχιστες οι μηχανές που κάνουν διάκριση μεταξύ κεφαλαίων-πεζών, μεταξύ των οποίων είναι το AltaVista και Go. Παραδείγματα μηχανών που δεν κάνουν διάκριση είναι το AOL Search, το Excite, το Google, το Lycos, κá. Στο AltaVista για παράδειγμα, το ‘buenavista’ θα επιστρέψει αποτελέσματα για το ‘buenavista’ ασχέτως όμως των κεφαλαίων-πεζών, αλλά αναζήτηση για το ‘BUENAVISTA’, θα επιστρέψει αποτελέσματα για το ‘BUENAVISTA’ και μόνο, ενώ το Excite θα επέστρεφε αποτελέσματα ασχέτως των κεφαλαίων. Βλέπει κανείς ότι αυτή η διάκριση γίνεται υπέρ των ερωτήσεων που περιέχουν κεφαλαίους χαρακτήρες.

2.5 Μηχανές αναζήτησης και θεματικοί κατάλογοι

Ο όρος «μηχανή αναζήτησης» χρησιμοποιείται συχνά για να περιγράψει τόσο μια πραγματική μηχανή αναζήτησης, όσο και ένα θεματικό κατάλογο. Ωστόσο, δεν είναι το ίδιο. Η διαφορά έγκειται στο πως δημιουργούνται οι κατάλογοι σε κάθε περίπτωση.

2.5.1 Μηχανές αναζήτησης

Οι μηχανές αναζήτησης όπως το AltaVista και το HotBot, δημιουργούν τις λίστες τους αυτόματα. Ψάχνουν το δίκτυο και οι χρήστες στη συνέχεια κάνουν αναζήτηση σε αυτά που έχουν βρει οι «αράχνες» των μηχανών. Αν αλλαχθεί μία σελίδα, η μηχανή αυτόματα εντοπίζει την αλλαγή. Οι τίτλοι της σελίδας, το «σώμα» της και άλλα στοιχεία, όλα παίζουν ρόλο στο πως θα κατατάσσεται.

Μια ειδική κατηγορία των μηχανών αναζήτησης είναι οι *metacrawlers*. Οι *metacrawlers* δεν ψάχνουν στο διαδίκτυο όπως οι κλασικές μηχανές, δεν διαθέτουν δηλαδή δική τους «αράχνη», αλλά την ερώτηση που υποβάλλει ο χρήστης, την στέλνουν σε πολλές διαφορετικές μηχανές από τις οποίες λαμβάνουν τα αποτελέσματα και τα παρουσιάζουν συγκεντρωμένα στο χρήστη.

2.5.2 Θεματικοί κατάλογοι

Ο θεματικός κατάλογος βασίζεται πάντα στην ανθρώπινη παρέμβαση για τη δημιουργία των λιστών του. Ένας σχεδιαστής τόπου αποστέλλει μια μικρή περιγραφή στο θεματικό κατάλογο για ολόκληρο τον τόπο του ή οι σχεδιαστές του θεματικού καταλόγου γράφουν μια δική τους περίληψη για τους τόπους που επιθεωρούν. Μια αναζήτηση ψάχνει μόνο σε αυτές τις περιλήψεις. Αυτό ίσως φαίνεται αδυναμία του θεματικού καταλόγου, αλλά πρέπει να σκεφτεί κανείς ότι οι θεματικοί κατάλογοι περιέχουν οργάνωση των τόπων σε κατηγορίες και είναι έτσι πιο εύκολο για το χρήστη να εντοπίσει την πληροφορία που θέλει, ξεκινώντας από μια ευρεία κατηγορία και περιορίζοντας όσο προχωρά, το πεδίο κίνησής του μέχρι να βρει σελίδες σχετιζόμενες με το θέμα του.

Αλλαγή σε μια σελίδα του τόπου δεν έχει κανένα αποτέλεσμα στην κατηγοριοποίησή του. Στοιχεία που είναι χρήσιμα για την βελτίωση της θέσης ενός τόπου σε μια μηχανή αναζήτησης δεν έχουν κανένα αποτέλεσμα στο θεματικό κατάλογο. Η μόνη εξαίρεση είναι ότι ένας τόπος πλούσιος σε περιεχόμενο έχει μεγαλύτερες πιθανότητες να επιθεωρηθεί σε σχέση με ένα τόπο με φτωχότερο περιεχόμενο.

Ο γνωστότερος και μεγαλύτερος θεματικός κατάλογος του διαδικτύου είναι το δημοφιλές Yahoo! (<http://www.yahoo.com>).

2.6 Κατηγορίες μηχανών αναζήτησης

Λόγω του τεράστιου όγκου πληροφοριών που υπάρχει στο διαδίκτυο, πολλές μηχανές αναζήτησης απευθύνονται σε συγκεκριμένες θεματικές κατηγορίες. Δε λείπουν βέβαια και οι μηχανές αναζήτησης γενικού περιεχομένου.

2.6.1 Γενικού περιεχομένου

Οι μηχανές αναζήτησης γενικού περιεχομένου είναι και οι πιο δημοφιλείς. Καθένας που διατηρεί ένα τόπο στο διαδίκτυο θέλει ο τόπος του να συμπεριλαμβάνεται στις λίστες των μηχανών αυτών, λόγω της μεγάλης τους κίνησης. Στην κατηγορία αυτή περιλαμβάνονται επίσης και μηχανές οι οποίες πουλάνε την τοποθέτηση ενός τόπου στις λίστες τους. Το γεγονός αυτό δεν θα έπρεπε να κάνει ένα χρήστη να αποφεύγει τη χρησιμοποίηση των μηχανών αυτών, διότι όταν ψάχνει για υπηρεσίες και προϊόντα, θα μπορούσε να βρει εταιρίες που μπορούν να πληρώσουν για τη διαφήμισή τους και οι οποίες τυχαίνει να τον ενδιαφέρουν. Παραδείγματα τέτοιων μηχανών είναι το GoTo (<http://www.goto.com>), του οποίου τα αποτελέσματα από τη δωρεάν λίστα προέρχονται από το Inktomi (<http://www.inktomi.com>), που ανήκει στις μηχανές αναζήτησης γενικού μη-πληρωμένου περιεχομένου, το FindWhat (<http://www.findwhat.com>), κá. Οι πιο γνωστές από τις μηχανές γενικού περιεχομένου είναι το AltaVista (<http://www.altavista.com>), το Excite (<http://www.excite.com>), το Lycos (<http://www.lycos.com>), το Northern Light (<http://www.northernlight.com>), κá.

2.6.2 Εξυπηρέτησης παιδικών αναγκών

Οι μηχανές αυτές επικεντρώνουν σε τόπους που είναι κατάλληλοι για παιδιά και αποκλείουν εκείνους που περιέχουν υλικό ακατάλληλο. Έτσι, υπάρχουν μηχανές αναζήτησης άρθρων, γενικού περιεχομένου, όπως το AOL NetFind Kids Only (<http://www.aol.com/netfind/kids/>), το Yahoooligans (<http://www.yahoooligans.com>), κá., ενώ μερικές μηχανές αναζήτησης γενικού περιεχομένου περιέχουν φίλτρα που προσπαθούν να αποκλείσουν υλικό ακατάλληλο για τα παιδιά, όπως το Family Filter του AltaVista, το Excite's Magellan (<http://magellan.excite.com>) που προσφέρει τη δυνατότητα αναζήτησης σε τόπους που έχουν «πράσινο φως» από μια επιλογή στην κεντρική σελίδα αναζήτησης. Μάλιστα υπάρχουν μηχανές ειδικά για γονείς, με άρθρα σχετικά με την εγκυμοσύνη, τη μόρφωση των παιδιών, κτλ. όπως το ABC's of Parenting (<http://www.abcparenting.com>) και το The Labor Of Love (<http://www.thelaboroflove.com/websearch/links/>).

2.6.3 Metacrawlers

Οι metacrawlers, από τους οποίους υπάρχει μια κατηγορία στην οποία ο χρήστης επιλέγει από μια πληθώρα τις μηχανές στις οποίες θα σταλεί η ερώτησή του. Παραδείγ-

ματα τέτοιων είναι το Search-It-All (<http://www.search-it-all.com>), το Skworm (<http://www.skworm.com>), το OneSeek (<http://www.oneseek.com>), κά. Παραδείγματα γενικών metacrawlers είναι το Go2Net/Metacrawler (<http://www.go2net.com>), το Dogpile (<http://www.dogpile.com>), το Inference Find (<http://www.infind.com>), κά.

2.6.4 Εύρεσης αρχείων πολυμέσων

Οι μηχανές αυτές αναζητούν αρχεία πολυμέσων, δηλαδή εικόνες, αρχεία ήχων, κτλ. Παραδείγματα είναι το AltaVista Photo Finder (<http://image.altavista.com/cgi-bin/avncgi>), το Ditto (<http://www.ditto.com>), το Lycos Pictures and Sounds (<http://www.lycos.com/picturethis/>), το Scour.net (<http://scour.net>), ενώ υπάρχουν πολλές μηχανές αναζήτησης αρχείων MP3⁴.

2.6.5 Αναζήτησης ειδήσεων και νέων

Μηχανές αναζήτησης ειδήσεων, όπως το Excite NewsTracker (<http://nt.excite.com>), το News Index (<http://www.newsindex.com>), τα νέα από το Northern Light (<http://www.northernlight.com/news.html>), που προέρχονται από διάφορες πηγές, όπως το Associated Press και το Business Wire, κά.

2.6.6 Αναζήτησης συγκεκριμένου γεωγραφικού χώρου

Μηχανές αναζήτησης που έχουν συγκεκριμένο γεωγραφικό χώρο δράσης. Οι μηχανές αυτές συνήθως είναι γενικού περιεχομένου. Έτσι, υπάρχουν μηχανές που αναζητούν σε τόπους της Αφρικής, όπως το Mosaique (<http://www.mosaique.net>), ο θεματικός κατάλογος Orientation Africa (<http://af.orientation.com>), το Zebra (<http://www.zebra.co.za>), κά., της Αμερικής, για την οποία υπάρχουν ειδικά για τον Καναδά και τη λατινική Αμερική, της Ασίας όπως το AltaVista Asia (<http://altavista.skali.com.my>), της Αυστραλίας και Νέας Ζηλανδίας, όπως το AltaVista Australia (<http://www.altavista.yellowpages.com.au>), το Anzwers (<http://www.anzwers.com.au>), κά., της Ευρώπης, η οποία έχει αρκετά μεγάλο πλήθος μηχανών, πχ. για την Ελλάδα υπάρχει το iBoom (<http://www.iboom.gr>), ενώ για όλη την Ευρώπη υπάρχει το EuroSeek (<http://www.euroseek.net>), της Κεντρικής Ανατολής, όπως το Arabist (<http://www.arabist.com>), και μηχανές που κάνουν αναζήτηση βάσει της γλώσσας, όπως το WorldBlaze Search (<http://www.worldblaze.com/search.html>), το οποίο δέχεται ερώτηση σε Αγγλικά, το μεταφράζει σε μια από τις γλώσσες που επιλέγει ο χρήστης, το στέλνει στο Inktomi και παρουσιάζει τα αποτελέσματα μεταφρασμένα στα Αγγλικά, κτλ.

2.6.7 Εξειδικευμένων θεμάτων

Η κατηγορία αυτή των μηχανών έχουν συγκεκριμένο θεματικό χώρο δράσης. Έτσι, υπάρχουν μηχανές που ειδικεύονται στην απάντηση ερωτήσεων, όπως το Ask Jeeves (<http://www.askjeeves.com>), η οποία δουλεύει με τη βοήθεια ανθρώπων και προσπαθεί να οδηγήσει στη σελίδα που εξηγεί ακριβώς τι είναι αυτό που ρώτησε ο χρήστης. Αν αποτύχει να βρει τέτοια σελίδα, τότε αποστέλλει την ερώτηση σε διάφορες μηχανές αναζήτησης

⁴ Το αρχείο MP3 είναι μουσικό αρχείο με μορφοποίηση MPEG-1 Layer III, το οποίο χρησιμοποιείται εκτενώς στο διαδίκτυο για τη διάδοση τραγουδιών, μιας και το μέγεθος αυτών των αρχείων είναι σχετικά μικρό, ενώ η ποιότητά τους είναι παραπλήσια εκείνης του μουσικού CD.

γενικού περιεχομένου και παρουσιάζει τα αποτελέσματα στο χρήστη. Το KnowPost (<http://www.knowpost.com>) απαντά στην ερώτηση του χρήστη, ο οποίος δεσμεύεται να απαντήσει και σε άλλες ερωτήσεις ο ίδιος. Άλλη υποκατηγορία των μηχανών αυτών είναι μηχανές που εξειδικεύονται σε θέματα υπολογιστών και πληροφορικής, όπως το Filez (<http://www.filez.com>), μηχανές αναζήτησης *domains*⁵ καθώς και των κατόχων τους, όπως το Domain Notes (<http://wdvl.internet.com/Internet/Domains/>), μηχανές για την αναζήτηση οικονομικών στοιχείων, όπως το Daily Stocks (<http://www.dailystocks.com>) για την παρακολούθηση των μετοχών, μηχανές αναζήτησης πληροφοριών σχετικά με τα νομικά, όπως το FindLaw (<http://www.findlaw.com>), σχετικά με ιατρικά θέματα, όπως το MedHunt (<http://www.hon.ch/MedHunt/>) που χρησιμοποιεί εκτός από την «αράχνη» και τον ανθρώπινο παράγοντα για τη δημιουργία των ευρετηρίων του, μηχανές αναζήτησης ηλεκτρονικών λιστών (*mailing lists*) και ομάδων συζητήσεων (*newsgroups*) όπως το Liszt (<http://www.liszt.com>) και το ForumOne (<http://www.forumone.com>) αντίστοιχα και μηχανές που έχουν να κάνουν με την αγορά προϊόντων μέσω του διαδικτύου, όπως το Buyer's Index (<http://www.buyersindex.com>). Παράλληλα, υπάρχουν μηχανές εξειδικευμένες στην αναζήτηση επιστημονικών θεμάτων, όπως το BioCrawler (<http://www.biocrawler.com>) για την αναζήτηση πληροφοριών σχετικά με τη βιολογία, το Chemie.de (<http://www.chemie.de>) που ειδικεύεται στη χημεία, κ.ά. Τέλος υπάρχουν μηχανές αναζήτησης που ειδικεύονται στην αναζήτηση στο «αόρατο» web, σε τύπους δηλαδή που δεν ευρετηριοποιούνται από τις άλλες μηχανές αναζήτησης. Παράδειγμα τέτοιας μηχανής είναι το Invisible Web (<http://www.invisibleweb.com>) και το Web Data (<http://www.webdata.com>).

2.7 Βασική σύνταξη ερωτήσεων

Όπως ειπώθηκε προηγουμένως, οι μηχανές αναζήτησης δεν είναι ίδιες. Για το λόγο αυτό η σύνταξη των ερωτήσεων που μπορεί να δεχθεί η καθεμία είναι διαφορετική, ωστόσο υπάρχουν κάποιοι γενικοί κανόνες που πρακτικά ισχύουν για όλες τις μηχανές.

Ένας βασικός κανόνας είναι ότι ο χρήστης πρέπει να είναι συγκεκριμένος αναφορικά με αυτό που ψάχνει. Όσο πιο συγκεκριμένος είναι, τόσο πιο πιθανό είναι να βρει αυτό για το οποίο ψάχνει. Όταν για παράδειγμα θέλει να βρει οτιδήποτε σχετικό με τα λάθη (*bugs*) των Windows 98 πρέπει να δώσει 'Windows 98 bugs' κι όχι ένα απλό 'Windows'.

Αρκετές φορές, ο χρήστης θέλει τα έγγραφα που του επιστρέφονται από μια αναζήτηση να περιέχουν όλες τις λέξεις της ερώτησής του. Το σύμβολο '+' το επιτρέπει αυτό. Για παράδειγμα, αναζήτηση για το 'formula+1+Ferrari' θα επιστρέψει έγγραφα που περιέχουν και τις τρεις λέξεις.

Άλλες φορές, είναι επιθυμητό να αποκλεισθούν κάποιες λέξεις, ενώ είναι θεμιτό να περιέχονται άλλες. Το σύμβολο '-' εξυπηρετεί το σκοπό αυτό. Για παράδειγμα, αναζήτηση για το 'Windows+98-NT' θα επιστρέψει έγγραφα που αναφέρονται στα Windows 98 και όχι στα Windows NT.

Το σύμβολο '+' χρησιμοποιείται για να εξασφαλίσει την ύπαρξη όλων των λέξεων της ερώτησης, ωστόσο δεν εξασφαλίζει την εγγύτητα των λέξεων αυτών, δηλαδή το αν θα είναι κοντά ή πιο μακριά η καθεμία λέξη από τις άλλες στο κείμενο. Είναι έτσι πιθανό ένα έγγραφο τελικά να μη σχετίζεται με αυτό που θέλει ο χρήστης, παρόλο που περιέχει όλες τις λέξεις που έθεσε. Κάνοντας μια αναζήτηση φράσης λύνει αυτό το πρόβλημα. Στην περίπτωση αυτή η μηχανή αναζήτησης αναζητά έγγραφα που περιέχουν τη φράση που έδωσε ο χρήστης ακριβώς όπως είναι. Αυτό γίνεται περικλείοντας τη φράση μέσα σε διπλά

⁵ Domain ονομάζεται το μνημονικό όνομα που χρησιμοποιείται αντί μιας περιοχής IP διευθύνσεων, για παράδειγμα το domain του διακομιστή aetos.it.teithe.gr είναι το it.teithe.gr.

εισαγωγικά. Για παράδειγμα, αναζήτηση για το “information retrieval systems” θα επιστρέψει έγγραφα που περιέχουν τη φράση όπως δίνεται.

Φυσικά, όλα αυτά τα σύμβολα μπορούν να συνδυαστούν μεταξύ τους για ακόμη πιο πολύπλοκα σχήματα ερωτήσεων.

2.7.1 Λογικοί τελεστές

Οι λογικοί τελεστές AND και NOT προσφέρουν την ίδια λειτουργικότητα με τα σύμβολα ‘+’ και ‘-’. Ο τελεστής OR που χρησιμοποιείται για την ύπαρξη ενός τουλάχιστον από τους όρους που συνδέει, δεν έχει κάποιο σύμβολο. Χρησιμοποιείται όπως κατά τα γνωστά :

book OR chapter

Εξ’ ορισμού, οι περισσότερες μηχανές χρησιμοποιούν το OR, παρόλο που τείνουν πρώτα να διαβαθμίζουν σελίδες που περιέχουν όλους τους όρους.

Οι τελεστές AND και NOT χρησιμοποιούνται όπως και τα σύμβολα ‘+’ και ‘-’ όσον αφορά στη σύνταξη.

Καλό είναι οι λογικοί τελεστές να γράφονται με κεφαλαίους χαρακτήρες, γιατί σε μερικές μηχανές είναι απαραίτητο, όπως το Excite.

2.7.2 Τελεστής NEAR

Ο τελεστής αυτός χρησιμοποιείται για να καθορίσει πόσο κοντά πρέπει να βρίσκονται οι όροι που συνδέει. Χρησιμοποιείται ως εξής :

message NEAR bottle

Οι μηχανές που υποστηρίζουν τον τελεστή αυτό είναι το AOL Search (πρέπει να δοθεί αριθμός απόστασης), το AltaVista (εξ’ ορισμού 10 λέξεις), το Lycos (εξ’ ορισμού 25 λέξεις) και το WebCrawler (σε απόσταση 2 λέξεων). Σε αυτές τις μηχανές, εκτός του AltaVista, μπορεί να γίνει υπέρβαση της εξ’ ορισμού τιμής απόστασης. Αυτό γίνεται καθορίζοντας ένα αριθμό στο τέλος του τελεστή, αμέσως μετά από το χαρακτήρα ‘/’. Για παράδειγμα :

message NEAR/15 bottle

καθορίζει την απόσταση στις 15 λέξεις.

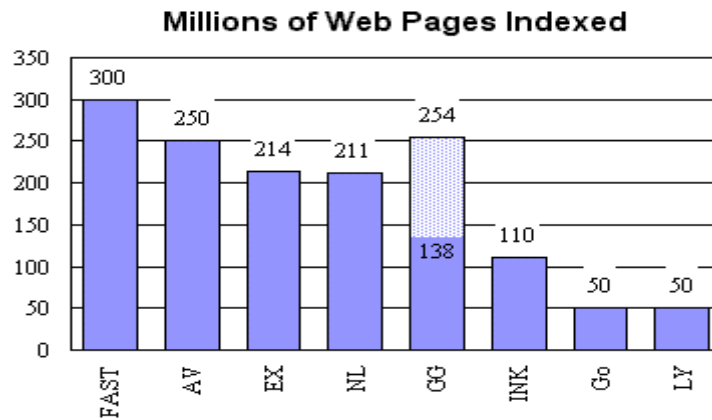
2.8 Μεγέθη των μηχανών αναζήτησης και των θεματικών καταλόγων

2.8.1 Μεγέθη των μηχανών αναζήτησης

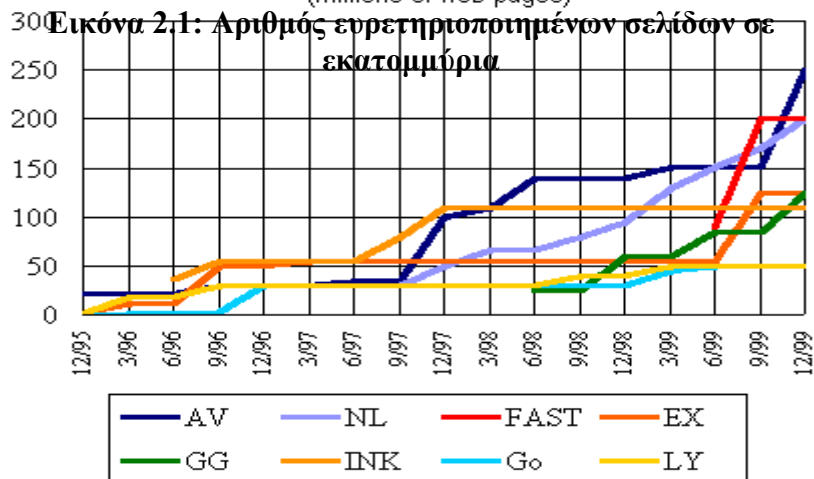
Η εικόνα 2.1 δείχνει τον αριθμό των ιστοσελίδων που έχουν ευρετηριοποιηθεί από τις πιο γνωστές μηχανές αναζήτησης (τα στοιχεία έχουν ημερομηνία 3 Φεβρουαρίου 2000). Ένας χρήστης που ψάχνει για ασυνήθιστη ή δύσκολη στην εύρεση πληροφορία θα δοκιμάσει να ψάξει σε μηχανή με μεγάλο μέγεθος ευρετηρίου, επειδή αυτό σημαίνει ότι καλύπτει μεγάλο μέρος του web. Συνεπώς, υπάρχει μεγαλύτερη πιθανότητα να βρει αυτό που ψάχνει. Ωστόσο, για πιο γενικές αναζητήσεις ή όταν γίνεται αναζήτηση για συγκεκριμένα θέματα, μεγάλο μέγεθος ευρετηρίου δε σημαίνει αυτόματα και καλύτερα αποτελέσματα. Η εκτεταμένη μπάρα για το Google οφείλεται στον τρόπο που χειρίζεται τους συνδέσμους, παρόλο που έχει γύρω στα 138 εκατομμύρια ευρετηριοποιημένες σελίδες,

μπορεί να επιστρέψει ακόμα και σελίδες που δεν έχει επισκεφθεί και έτσι ο αριθμός υπολογίζεται στα 254 εκατομμύρια.

Η εικόνα 2.2 δείχνει το μέγεθος των μηχανών σε σχέση με το χρόνο. Όπως φαίνεται και από το σχήμα, το AltaVista εμφανίστηκε το Δεκέμβριο του 1995 και το μέγεθός



Κλειδί: FAST=FAST, AV=AltaVista, EX=Excite,
NL=Northern Light, GG=Google, INK=Inktomi, Go=Go
(InfoSeek), LY=Lycos
(millions of web pages)

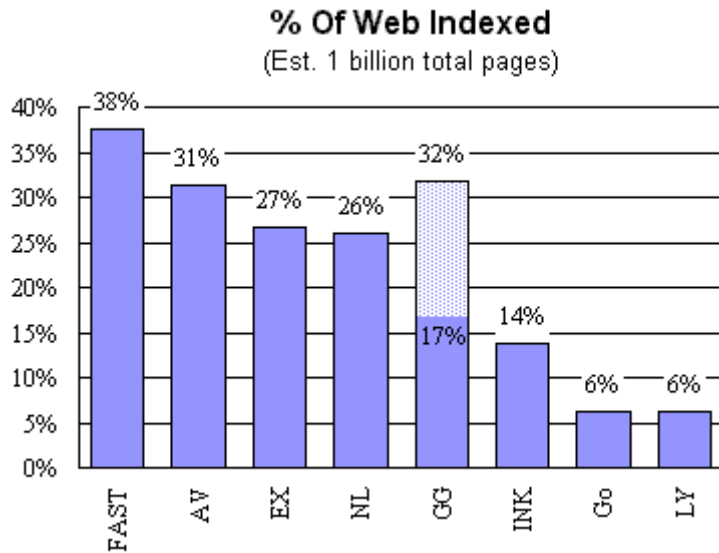


Εικόνα 2.2: Μέγεθος μηχανών σε σχέση με το χρόνο

του ήταν κατά πολύ μεγαλύτερο από τις άλλες μηχανές. Έτσι, ο συναγωνισμός ανάγκασε τους ανταγωνιστές στις αρχές του 1996 να αυξήσουν κι εκείνοι το μέγεθός τους όπως φαίνεται και από την εικόνα. Από το Σεπτέμβριο του 1996 μέχρι και το Σεπτέμβριο του 1997, δεν υπήρξε καμία δραστική αλλαγή στο μέγεθος των μηχανών, παρόλο που το μέγεθος του web συνέχισε να αυξάνει. Από το Σεπτέμβριο του 1997 μέχρι και το τέλος του 1998, το AltaVista και το Inktomi συναγωνίζονταν μεταξύ τους για το μέγεθος. Αλλά μέ-

σα στο 1999, η μάχη αυτή μεταφέρθηκε μεταξύ AltaVista, Northern Light και FAST Search. Τον Ιανουάριο του 2000, το FAST Search ανακοίνωσε ότι έσπασε το φράγμα των 300 εκατομμυρίων σελίδων, δίνοντάς του το στέμμα του μεγαλύτερου ευρετηρίου στο web. Αυτό δεν φαίνεται στην εικόνα, επειδή αυτό ανανεώνεται κάθε τετράμηνο.

Η εικόνα 2.3 δείχνει την κάλυψη του web από τις μηχανές αναζήτησης. Το Φε-



Εικόνα 2.3: Κάλυψη του web ανά μηχανή

βρουάριο του 2000, μια πρόβλεψη από το Inktomi και το NEC Research Institute υπολογίζει ότι υπάρχει περίπου 1 δισεκατομμύριο σελίδες του web που μπορούν να ευρετηριοποιηθούν. Το σχήμα δείχνει για κάθε μηχανή αναζήτησης ποιο ποσοστό του web καλύπτει. Όπως φαίνεται από το σχήμα, τα πρωτεία έχει το FAST Search με 38% και δεύτερο ακολουθεί το AltaVista με 31%. Το Google έχει ένα εικονικό ποσοστό 32% που το φέρνει δεύτερο μπροστά από το AltaVista, αλλά το πραγματικό του ποσοστό είναι 17%.

Είναι παράξενο το γεγονός ότι το πολύ ένα 40% του web καλύπτεται από τις μηχανές αναζήτησης, αλλά πρέπει να σκεφτεί την ύπαρξη του «αόρατου» web και ότι ίσως πρέπει να είναι ικανοποιημένος με ένα τέτοιο ποσοστό, εφόσον το διαδίκτυο αναπτύσσεται με εκθετικούς ρυθμούς. Φυσικά, λόγω και του ανταγωνισμού και της γρήγορης ανάπτυξης του ηλεκτρονικού εμπορίου, τα ποσοστά αυτά θα αλλάζουν συνεχώς, ώστε κάθε μηχανή αναζήτησης να προσπαθεί να παραμένει ανταγωνιστική σε σχέση με τις άλλες.

2.8.2 Μεγέθη των θεματικών καταλόγων

Ο πίνακας 2.1 δείχνει κάποια μεγέθη μερικών θεματικών καταλόγων.

Κατάλογος	Άτομα	Κατηγορίες	Σύνδεσμοι	Δεδομένα στις
Yahoo!	100+	δ/δ	1,2 εκατ.+	11/99
Open Directory	15.400	153.000	950.000	9/99
LookSmart	200	60.000	1 εκατ.	9/99
Snap	30 – 50	64.000	600.000	11/99
Ask Jeeves	30	δ/δ	7 εκατ.	11/98

Πίνακας 2.1: Μεγέθη θεματικών καταλόγων (δ/δ σημαίνει ότι δεν υπάρχουν δεδομένα για το συγκεκριμένο στοιχείο)

Το Ask Jeeves δεν είναι ακριβώς κατάλογος, είναι υπηρεσία που παρέχει απαντήσεις σε ερωτήσεις, ωστόσο η δομή του μοιάζει με εκείνη ενός καταλόγου.

Ο πίνακας δείχνει τον αριθμό των ατόμων που είναι απασχολημένα σε κάθε κατάλογο για τη δημιουργία των λιστών, τον αριθμό των κατηγοριών και τον αριθμό των συνδέσμων κάθε καταλόγου. Περισσότερα άτομα δεν σημαίνουν απαραίτητα και καλύτερη υπηρεσία, ωστόσο ικανός αριθμός δείχνει ότι γίνεται ποιοτική κατηγοριοποίηση.

2.9 Το μέλλον των μηχανών αναζήτησης

Ο ρόλος των μηχανών αναζήτησης έχει αρχίσει εδώ και καιρό να αλλάζει από απλές μηχανές αναζήτησης. Οι σύγχρονες μηχανές προσφέρουν ήδη κι άλλες υπηρεσίες εκτός των κλασικών υπηρεσιών εύρεσης πληροφοριών. Μια επίσκεψη σε μια από τις σελίδες των μηχανών αναζήτησης θα δείξει ότι οι σελίδες αυτές προσφέρουν πια διάφορες υπηρεσίες, όπως ενημέρωση είτε με τη μορφή διαφημίσεων, είτε με τη μορφή τίτλων ειδήσεων, υπηρεσίες ηλεκτρονικού ταχυδρομείου και φιλοξενίας προσωπικών ιστοσελίδων, προσφορές για την αγορά προϊόντων μέσω του διαδικτύου (το λεγόμενο *ηλεκτρονικό εμπόριο e-commerce*), υπηρεσίες εύρεσης ατόμων (οι λεγόμενες *yellow pages*), υπηρεσίες επικοινωνίας μέσω των δωματίων συνομιλίας (*chat rooms*) και πολλές άλλες (όπως μετάφραση που προσφέρει το AltaVista). Και όλα αυτά προσφέρονται μέσα από ένα προσωποποιημένο περιβάλλον το οποίο μπορεί να αλλάξει και να έρθει στα μέτρα του κάθε χρήστη. Ο χρήστης είναι αυτός που αποφασίζει πλέον τι θα βλέπει και πώς θα το βλέπει. Εκτός αυτού, οι μηχανές αναζήτησης θέλουν να ολοκληρωθούν με το περιβάλλον εργασίας του χρήστη στον ηλεκτρονικό υπολογιστή μέσω των *βοηθημάτων (utilities)*, τα οποία ενσωματώνονται συνήθως με τα προγράμματα πλοήγησης του χρήστη και προσφέρουν πιο γρήγορη προσπέλαση στις υπηρεσίες τους.

Αυτό που θα συνεχίσει ο κόσμος να βλέπει στο μέλλον είναι τη γέννηση των *πυλών (web portals)*, που προσφέρουν ό,τι προσφέρει μια σύγχρονη μηχανή αναζήτησης, υπηρεσίες ευρετηρίου, κτλ. και θέλουν να αποτελέσουν το σημείο εκκίνησης του χρήστη όταν συνδέεται με το διαδίκτυο. Γνωστή πύλη του ελληνικού διαδικτύου που δημιουργήθηκε και σχετικά πρόσφατα είναι το In.gr (<http://www.in.gr>).

ΚΕΦΑΛΑΙΟ 3

IBM DB2 TEXT EXTENDER

3.1 Εισαγωγή στον IBM DB2 Text Extender

Ο Text Extender αποτελεί *προσθήκη (add-on)* για τη σχεσιακή βάση DB2 της IBM. Προσθέτει τη δυνατότητα ανάκλησης κειμένου μέσω SQL ερωτήσεων σε κείμενο που έχει *ευρετηριοποιηθεί (indexed)* στη βάση. Με τον Text Extender, η DB2 μπορεί να αποθηκεύσει μη δομημένο κείμενο μεγέθους μέχρι και 2GB στη βάση. Το κείμενο μπορεί να υπάρχει και εκτός της βάσης σε αρχεία του λειτουργικού συστήματος, μιας και ο Text Extender επιτρέπει την ευρετηριοποίηση εξωτερικών αρχείων. Με λίγα λόγια, οι δυνατότητες που προσφέρει ο Text Extender είναι οι εξής :

- Αναζήτηση σε κείμενα για συγκεκριμένες λέξεις, συνώνυμα μιας λέξης ή φράσης.
- Αναζήτηση με *χαρακτήρες μάσκες (wildcard characters)*.
- Αναζήτηση σε κείμενα διαφόρων γλωσσών, σε διάφορες μορφοποιήσεις.
- Αναζήτηση τύπου «fuzzy» για λέξεις που έχουν ορθογραφία παρόμοια με αυτή του όρου αναζήτησης, διαδικασία που είναι χρήσιμη για την εύρεση λέξεων με λάθος ορθογραφία.
- Αναζήτηση με βάση ερώτηση που είναι εκφρασμένη σε φυσική γλώσσα.
- Αναζήτηση ονομάτων ανθρώπων, τοποθεσιών και οργανισμών.
- Αναζήτηση λέξεων που «ακούγονται» παρόμοια με τον όρο αναζήτησης.

Με την εγκατάσταση, ο Text Extender δημιουργεί στη βάση *διαδικασίες και συναρτήσεις (procedures and functions)*, που χρησιμοποιούνται για την αναζήτηση κειμένου με SQL ερωτήσεις. Ακολουθεί παράδειγμα :

```
SELECT * FROM MyTextTable
WHERE version = '2'
AND DB2TX.CONTAINS (
    DB2BOOKS,
    "authorization"
    IN SAME PARAGRAPH AS "table"
    AND SYNONYM FORM OF "delete") = 1
```

Αυτή η SQL ερώτηση ανακτά όλες τις εγγραφές του πίνακα MyTextTable, όπου η στήλη DB2BOOKS περιέχει τη λέξη 'authorization', στην ίδια παράγραφο με τη λέξη 'table', 'delete' ή συνώνυμο της 'delete'.

3.2 Ο Text Extender σε περιβάλλον πελάτη - διακομιστή

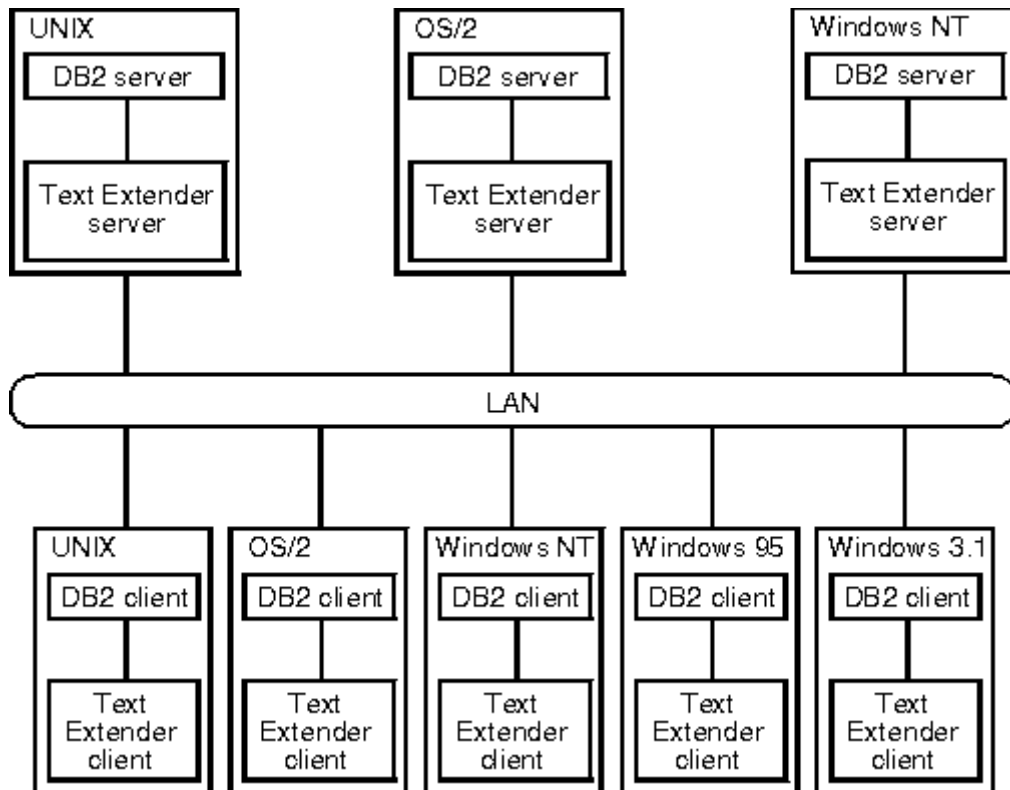
Το σχήμα 3.1 δείχνει πως ο Text Extender υπάρχει σε ένα περιβάλλον *πελάτη - διακομιστή (client - server)*. Το κύριο μέρος του Text Extender εγκαθίσταται στον ίδιο διακομιστή με εκείνον που είναι εγκατεστημένη και η βάση DB2. Ένα *στιγμιότυπο (instance)* του Text Extender εγκαθίσταται με ένα στιγμιότυπο βάσης.

Μια τυπική εγκατάσταση του Text Extender σε *τοπικό δίκτυο (LAN)* μπορεί να αποτελείται :

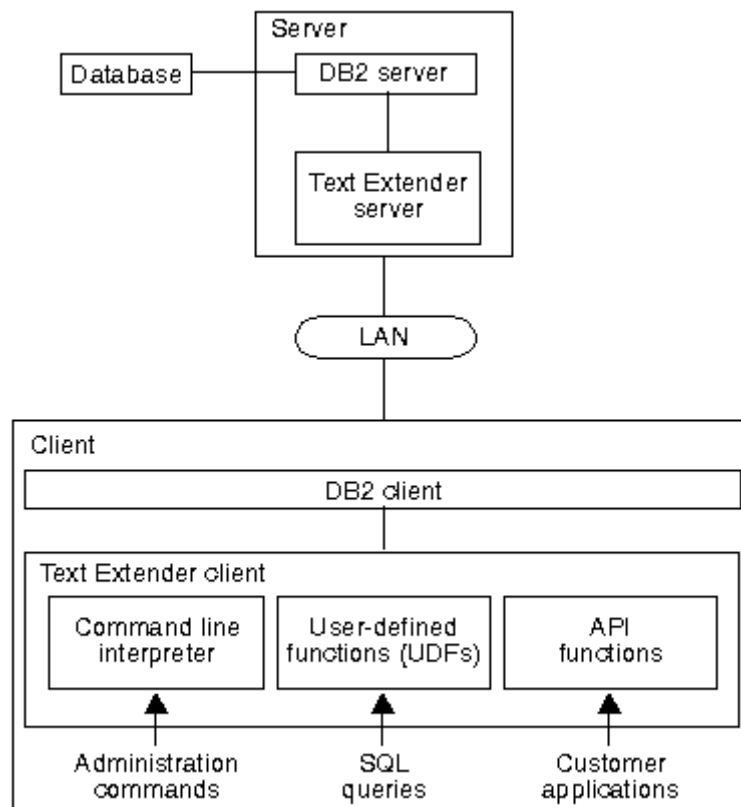
- Από έναν ή περισσότερους Text Extender διακομιστές (σε περιβάλλοντα AIX, Windows NT, Solaris κá.).
- Πελάτες σε περιβάλλοντα Windows NT/9x/3.xx, OS/2, AIX, Solaris, HP-UX κá. με πρόσβαση σε έναν ή περισσότερους από τους παραπάνω Text Extender διακομιστές.

Το σχήμα 3.2 δείχνει μια τυπική *παραμετροποίηση (configuration)* του Text Extender. Προκειμένου ένας πελάτης να τον χρησιμοποιήσει, πρέπει να εγκαταστήσει ένα πελάτη DB2 και μερικά εργαλεία του Text Extender. Τα εργαλεία αυτά αποτελούν τον

πελάτη του Text Extender, ο οποίος επικοινωνεί με τον διακομιστή Text Extender μέσω του πελάτη DB2.



Σχήμα 3.1: Ο Text Extender σε περιβάλλον πελάτη – διακομιστή



Σχήμα 3.2: Τυπική παραμετροποίηση του Text Extender

Ο Text Extender αποτελείται από τα εξής :

- Ένα μεταφραστή γραμμής εντολών (*command line interpreter*), παρόμοιο με αυτόν που χρησιμοποιεί η DB2, για διαχειριστικές εργασίες.
- Συναρτήσεις ορισμένες από το χρήστη (*user defined functions*), για την χρησιμοποίησή τους σε SQL ερωτήσεις. Στην παραπάνω εικόνα, φαίνεται ότι οι συναρτήσεις αυτές είναι μέρος του πελάτη, γιατί είναι διαθέσιμες σε αυτόν. Παρόλα αυτά, είναι μέρος του διακομιστή και εκτελούνται εκεί, ωστόσο μπορούν να χρησιμοποιηθούν από οποιοδήποτε πελάτη DB2, χωρίς να έχει εγκατασταθεί ο πελάτης του Text Extender.
- Ένα API¹, αποτελούμενο από συναρτήσεις που μπορούν να κληθούν από C προγράμματα για αναζήτηση κειμένου και εμφάνιση του αποτελέσματος.

3.3 Είδη ευρετηρίων του Text Extender

Ο Text Extender υποστηρίζει τέσσερα είδη ευρετηρίων (*indexes*) : το γλωσσολογικό (*linguistic*), ακριβές (*precise*), διπλό (*dual*) και το *ngram*.

3.3.1 Γλωσσολογικό ευρετήριο (*linguistic index*)

Στην περίπτωση που χρησιμοποιείται αυτό το είδος ευρετηρίου, η γλωσσολογική επεξεργασία του κειμένου εφαρμόζεται κατά τη διάρκεια ανάλυσης του προς ευρετηριοποίηση κειμένου. Αυτό σημαίνει ότι οι λέξεις ανάγονται στη βασική τους μορφή όταν ευρετηριοποιούνται. Για παράδειγμα η λέξη ‘industries’ ευρετηριοποιείται ως ‘industry’.

Η διαδικασία της αναγωγής εφαρμόζεται και κατά τη διάρκεια επεξεργασίας της ερώτησης. Οι όροι αναζήτησης ανάγονται στη βασική τους μορφή πριν γίνει η αναζήτηση στο ευρετηριοποιημένο κείμενο.

Κύριο πλεονέκτημα αυτού του τρόπου ευρετηριοποίησης είναι ότι οποιαδήποτε παραλλαγή του όρου αναζήτησης μπορεί να ταιριάζει με οποιαδήποτε παραλλαγή ενός όρου στο ευρετηριοποιημένο κείμενο. Αυτό έχει και ως αποτέλεσμα ότι η αναζήτηση δεν είναι βασισμένη σε κεφαλαία-πεζά (*case insensitive*).

Αυτό το είδος ευρετηριοποίησης απαιτεί το λιγότερο αποθηκευτικό χώρο, ωστόσο η διαδικασία της ευρετηριοποίησης και της αναζήτησης διαρκεί περισσότερο χρόνο σε σχέση με άλλους τρόπους ευρετηριοποίησης, λόγω του κόστους της επεξεργασίας των όρων αναζήτησης.

Υπάρχουν διάφοροι τρόποι γλωσσολογικής επεξεργασίας οι οποίοι εξαρτώνται από τη γλώσσα του κειμένου. Οι τρόποι αυτοί αναφορικά είναι οι εξής :

- Διαχωρισμός λέξης και πρότασης.
- Επεξεργασία αρχής πρότασης.
- Η αντίστροφη διαδικασία του συλλαβισμού.
- Κανονικοποίηση των όρων σε μια στάνταρτ μορφή κατά την οποία δεν υπάρχουν κεφαλαίοι χαρακτήρες, καθώς και χαρακτήρες με τόνους, όπως το ‘ü’.
- Αναγωγή των όρων στη βασική τους μορφή, για παράδειγμα η λέξη ‘bought’ ανάγεται στην ‘buy’.
- Απλοποίηση σύνθετων λέξεων, όπου σύνθετες λέξεις, όπως για παράδειγμα η γερμανική λέξη ‘wetterbericht’, απλοποιείται και ευρετηριοποιείται όχι μόνο ως ‘wetterbericht’, αλλά και ως ‘wetter’ και ‘bericht’.
- Φιλτράρισμα των στοπ λέξεων, όπως είναι τα άρθρα, οι αντωνυμίες, κά.

¹ API (Application Programming Interface) : Σύνολο εντολών που χρησιμοποιεί ένα πρόγραμμα προκειμένου να εκτελέσει χαμηλού επιπέδου εργασίες, οι οποίες εκτελούνται από το λειτουργικό σύστημα.

- Φιλτράρισμα μέρους του λόγου (*part-of-speech filtering*), τεχνική παρόμοια με το φιλτράρισμα των στοπ λέξεων. Ευρετηριοποιούνται μόνο ουσιαστικά, ρήματα και επίθετα.
- *Εξαγωγή χαρακτηριστικών (feature extraction)*, κατά την οποία εξάγονται από το κείμενο ειδικοί όροι, όπως ονόματα, τοποθεσίες, πρόσωπα, συντομογραφίες, κ.ά.

3.3.2 Ακριβές ευρετήριο (precise index)

Όταν επιλεγεί αυτό το είδος ευρετηρίου, οι όροι των κειμένων ευρετηριοποιούνται ακριβώς όπως εμφανίζονται σε αυτά. Για παράδειγμα, οι λέξεις 'cars' και 'car' ευρετηριοποιούνται ξεχωριστά. Εδώ πλέον, η αναζήτηση βασίζεται σε κεφαλαία-πεζά (*case sensitive*).

Η ίδια διαδικασία ακολουθείται και στους όρους αναζήτησης, συγκρίνονται δηλαδή όπως είναι με τους ευρετηριοποιημένους όρους. Μπορεί να χρησιμοποιηθούν και χαρακτήρες μάσκες, για να επαυξηθεί η αναζήτηση. Για παράδειγμα, με όρο αναζήτησης το 'car*' θα επιστρέψει 'cars' και 'car'.

Κύριο πλεονέκτημα αυτού του τρόπου ευρετηριοποίησης είναι ότι η αναζήτηση είναι περισσότερο ακριβής και παράλληλα οι διαδικασίες ευρετηριοποίησης και ανάκτησης είναι γρήγορες, ωστόσο το κυριότερο μειονέκτημά του είναι η απαίτηση περισσότερου αποθηκευτικού χώρου για το ευρετηριοποιημένο κείμενο.

Οι τρόποι γλωσσολογικής επεξεργασίας που χρησιμοποιούνται για ένα ακριβές ευρετήριο είναι ο χωρισμός λέξεων και προτάσεων και το φιλτράρισμα στοπ λέξεων.

3.3.3 Διπλό ευρετήριο (dual index)

Το διπλό ευρετήριο είναι συνδυασμός των γλωσσολογικού και ακριβούς. Ευρετηριοποιεί τόσο την κανονική μορφή (στάνταρ μορφή του όρου σε πεζά και χωρίς τόνους), τη βασική μορφή, όπως τα απαρέμφατα, όσο και την μορφή κάθε όρου έτσι όπως εμφανίζεται στο κείμενο. Για παράδειγμα, τα 'horses', 'Horses' και 'horse' θα ευρετηριοποιηθούν ξεχωριστά.

Αυτό το είδος ευρετηρίου επιτρέπει στο χρήστη να επιλέξει το αν η αναζήτηση θα γίνει με γλωσσολογικό ή ακριβές τρόπο. Στην περίπτωση SQL ερώτησης, αυτό γίνεται με δυο παραμέτρους. Για γλωσσολογική αναζήτηση χρησιμοποιείται η παράμετρος `STEMMED FORM OF`, ενώ για αναζήτηση ακρίβειας η `PRECISE FORM OF`.

Το διπλό ευρετήριο χρησιμοποιεί και τον περισσότερο αποθηκευτικό χώρο για την ευρετηριοποίηση. Παράλληλα, οι διαδικασίες της ευρετηριοποίησης και αναζήτησης διαρκούν περισσότερο σε σχέση με το γλωσσολογικό ευρετήριο.

3.3.4 Ngram ευρετήριο

Το ngram ευρετήριο βασίζεται στην ανάλυση κειμένου με τμηματοποίηση σετ χαρακτήρων. Η ανάλυση δεν βασίζεται σε λεξικό. Όταν το κείμενο χρησιμοποιεί DBCS² χαρακτήρες, πρέπει να χρησιμοποιηθεί αυτό το είδος ευρετηρίου. Κανένα άλλο δεν υποστηρίζει σύνολο DBCS χαρακτήρων.

Η ευρετηριοποίηση με ngram υποστηρίζει «fuzzy» αναζήτηση, μπορεί δηλαδή να βρεθούν συμβολοσειρές που «ακούγονται» παρόμοια με τους όρους αναζήτησης. Για πα-

² DBCS (Double Byte Character Set) : σετ χαρακτήρων που χρησιμοποιεί ένα ή δύο bytes για την παράσταση ενός χαρακτήρα. Μπορούν έτσι να παρασταθούν παραπάνω από 256 χαρακτήρες, επιτρέποντας την παράσταση χαρακτήρων που χρησιμοποιούνται από ιδιογραφικά συστήματα, όπως κινέζικα, κ.ά.

ράδειγμα, αναζήτηση για 'Extender' μπορεί να επιστρέψει και την ορθογραφικά λάθος λέξη 'Extendrrs'. Επίσης ορίζεται συντελεστής ομοιότητας, κατά πόσο θέλει ο χρήστης να είναι αυστηρός ο Extender στη σύγκριση.

Για να γίνει αναζήτηση με διάκριση κεφαλαίων-πεζών σε ευρετηριοποιημένο κείμενο κατά ngram, δεν είναι αρκετό να δοθεί μόνο η παράμετρος `PRECISE FORM OF` (όπως και στην περίπτωση του διπλού ευρετηρίου). Αυτό συμβαίνει διότι συνήθως ένα τέτοιο ευρετήριο δεν μπορεί να ξεχωρίσει μεταξύ κεφαλαίων-πεζών στο ευρετηριοποιημένο κείμενο. Αυτό μπορεί να γίνει αν δοθεί και η παράμετρος `CASE_ENABLED` κατά τη διάρκεια δημιουργίας του ευρετηρίου. Στη συνέχεια, η παράμετρος `PRECISE FORM OF` μπορεί να χρησιμοποιηθεί κατά τα γνωστά.

Όπως θα περίμενε κανείς, χρησιμοποιώντας την παράμετρο `CASE_ENABLED` κατά τη δημιουργία του ευρετηρίου, απαιτείται περισσότερος αποθηκευτικός χώρος και παράλληλα η αναζήτηση διαρκεί περισσότερο.

3.4 Γλωσσολογική επεξεργασία για την ευρετηριοποίηση

Κατά τη διάρκεια ευρετηριοποίησης, ο Text Extender κάνει μια γλωσσολογική ανάλυση του κειμένου. Η ανάλυση αποτελείται από :

- Βασική ανάλυση κειμένου
 - Αναγνώριση τεκμηρίων (*tokenization*)
 - Κανονικοποίηση όρων σε στάνταρ μορφή
 - Αναγνώριση προτάσεων
- Αναγωγή όρων στη βασική τους μορφή
- Φιλτράρισμα των στοπ λέξεων
- Απλοποίηση των σύνθετων όρων

Οι διαδικασίες αυτές περιγράφονται αναλυτικά παρακάτω.

Ο πίνακας 3.1 δείχνει τους τρόπους ανάλυσης του κειμένου όταν το είδος του ευρετηρίου που χρησιμοποιείται είναι το γλωσσολογικό.

Κείμενο	Ευρετηριοποιημένος όρος	Γλωσσολογική επεξεργασία
mouse, käfer	mouse, kaefer	Βασική ανάλυση (κανονικοποίηση)
bought, cars	buy, car	Αναγωγή στη βασική μορφή
system-based	system-based, system, base	Αποσύνθεση και αναγωγή
a report on animals	report, animal	Φιλτράρισμα στοπ λέξεων και αναγωγή. Οι στοπ λέξεις είναι : a, on

Πίνακας 3.1: Ανάλυση κειμένου με το γλωσσολογικό ευρετήριο

Ο πίνακας 3.2 δείχνει τους τρόπους ανάλυσης για το ακριβές ευρετήριο.

Κείμενο	Ευρετηριοποιημένος όρος	Γλωσσολογική επεξεργασία
mouse, käfer	mouse, käfer	Δεν γίνεται κανονικοποίηση
bought, cars	bought, cars	Δεν γίνεται αναγωγή

system-based	system-based	Δεν γίνεται αποσύνθεση
a report on animals	report, animals	Φιλτράρισμα στοπ λέξεων. Οι στοπ λέξεις είναι : a, on

Πίνακας 3.2: Ανάλυση κειμένου με το ακριβές ευρετήριο

Περαιτέρω επεξεργασία γίνεται όταν το ευρετήριο έχει δημιουργηθεί χρησιμοποιώντας την επιλογή της εξαγωγής χαρακτηριστικών (*feature extraction*) και όταν το κείμενο είναι μόνο στην αγγλική γλώσσα. Η εξαγωγή χαρακτηριστικών περιγράφεται αναλυτικά παρακάτω (§3.4.5).

3.4.1 Ανάλυση κειμένου

Ο Text Extender κάνει τη βασική ανάλυση του κειμένου χωρίς να χρησιμοποιεί λεξικό.

3.4.1.1 Αναγνώριση τεκμηρίων

Όταν ευρετηριοποιείται ένα κείμενο, οι όροι αναγνωρίζονται ακόμα κι όταν περιέχουν μη αλφαβητικούς χαρακτήρες, όπως '14.000δρχ.' ή '20/10/1998'. Τα παρακάτω θεωρούνται μέρη ενός όρου :

- Τόνοι
- Συντομογραφίες που συμβολίζουν νομίσματα (όπως 'δρχ.')
- Χαρακτήρες που διαχωρίζουν αριθμούς (όπως '/' ή '.')
- Ο χαρακτήρας '@' που χρησιμοποιείται στις διευθύνσεις ηλεκτρονικού ταχυδρομείου
- Ο χαρακτήρας '+'.

Παράλληλα, υπάρχουν κανόνες ανά γλώσσα που χρησιμοποιούνται στην αναγνώριση όρων που περιέχουν :

- Προθέματα με τόνους στις γλώσσες ρωμαϊκού τύπου, όπως *l'aventure* στα γαλλικά
- Μορφοποιήσεις ημερομηνιών, ώρας και αριθμών ανά χώρα
- Εναλλακτικούς όρους, όπως *mission/responsibility*, χρησιμοποιώντας στα αγγλικά τον χαρακτήρα '/'
- Ακολουθούμενες αποστροφές, όπως σε ιταλικές λέξεις, για παράδειγμα στη λέξη *securita'*. Στα ιταλικά, είναι συνηθισμένο όταν δεν υπάρχει σετ χαρακτήρων που έχει τονούμενους χαρακτήρες, ο τόνος να τοποθετείται μετά τον χαρακτήρα, όπως στο «a'» στην προηγούμενη λέξη, που στην ουσία είναι το «à»

3.4.1.2 Κανονικοποίηση όρων σε στάνταρ μορφή

Η κανονικοποίηση μειώνει όρους που έχουν μικτούς κεφαλαίους-πεζούς χαρακτήρες, καθώς και όρους με τονισμένους ή ειδικούς χαρακτήρες σε μια στάνταρ μορφή. Αυτό γίνεται εξ' ορισμού όταν το είδος του ευρετηρίου είναι το γλωσσολογικό, ή το διπλό, στο οποίο όταν γίνεται η αναζήτηση χρησιμοποιείται και η παράμετρος `STEMMED FORM OF` (στο ακριβές ευρετήριο, κεφαλαίοι-πεζοί χαρακτήρες μένουν όπως είναι και οι αναζητήσεις βασίζονται σε κεφαλαία-πεζά).

Για παράδειγμα, ο όρος 'Computers' ευρετηριοποιείται ως 'computer', το πρώτο κεφαλαίο γράμμα γίνεται πεζό και ο πληθυντικός μετατρέπεται σε ενικό. Αναζήτηση για τον όρο 'computer' βρίσκει όχι μόνο το 'computer', αλλά και το 'Computer'. Το αποτέ-

λεσµα της κανονικοποίησης κατά την ευρετηριοποίηση είναι ότι οι όροι ευρετηριοποιούνται µε τον ίδιο τρόπο, ανεξάρτητα από τα κεφαλαία-πεζά των όρων.

Η κανονικοποίηση δεν εφαρµόζεται µόνο κατά την ευρετηριοποίηση, αλλά και κατά την ανάκτηση. Κεφαλαίοι χαρακτήρες σε όρο αναζήτησης αλλάζουν σε πεζούς πριν την αναζήτηση. Όταν για παράδειγμα, ο όρος αναζήτησης είναι το 'Computer', ο όρος που χρησιµοποιείται για την αναζήτηση είναι το 'computer'.

Τονούµενοι και ειδικοί χαρακτήρες κανονικοποιούνται µε τον ίδιο τρόπο. Οποιαδήποτε παραλλαγή του école, όπως École, βρίσκει το école, Ecole, κ.ο.κ.

Αν ο όρος αναζήτησης περιλαµβάνει και χαρακτήρες µάσκες, τότε η κανονικοποίηση εφαρµόζεται πριν γίνει η επεξεργασία αυτών των χαρακτήρων.

3.4.1.3 Αναγνώριση προτάσεων

Η αναζήτηση όρων µέσα στην ίδια πρόταση είναι δυνατή. Για να υπάρξει αυτή η δυνατότητα, το κείµενο αναλύεται πριν την ευρετηριοποίηση για να βρεθεί το πού αρχίζει και τελειώνει κάθε πρόταση. Το τέλος µιας πρότασης συµβολίζεται µε τελεία, θαυµαστικό ή ερωτηµατικό, ακολουθούµενο από ένα κενό χαρακτήρα. Συντοµογραφίες που τελειώνουν σε τελεία αγνοούνται.

3.4.2 Αναγωγή των όρων στη βασική τους µορφή

Σε ένα γλωσσολογικό ή διπλό ευρετήριο, µπορεί να γίνει αναζήτηση για το 'window' και να βρεθεί το 'windows'. Οι όροι ανάγονται στη βασική τους µορφή πριν ευρετηριοποιηθούν. Ο όρος 'windows' ευρετηριοποιείται ως 'window'. Αργότερα, όταν γίνει αναζήτηση για το 'window', θα βρεθεί το έγγραφο που περιέχει τον ευρετηριοποιηµένο όρο. Το ίδιο έγγραφο θα βρεθεί αν γίνει αναζήτηση για το 'windows'.

Με παρόµοιο τρόπο, τα ρήµατα ανάγονται στα αντίστοιχα απαρέµφατά τους, για παράδειγμα, το 'fell' γίνεται 'fall'.

3.4.3 Φιλτράρισμα των στοπ λέξεων

Οι στοπ λέξεις είναι λέξεις όπως προθέσεις και αντωνυµίες που εμφανίζονται πολύ συχνά στα κείµενα και έχουν λίγη αξία ώστε να αναζητηθούν. Τέτοιου είδους λέξεις βρίσκονται σε µια λίστα µε στοπ λέξεις η οποία χρησιµοποιείται κατά τη διάρκεια της ευρετηριοποίησης, ώστε οι λέξεις αυτές να µη συµπεριληφθούν στο ευρετήριο.

Το φιλτράρισμα των στοπ λέξεων δεν βασίζεται σε κεφαλαία-πεζά. Η λίστα των στοπ λέξεων ανά γλώσσα µπορεί να αλλαχθεί.

3.4.4 Απλοποίηση των σύνθετων όρων

Γλώσσες όπως τα γερµανικά ή τα ολλανδικά είναι πλούσιες σε σύνθετους όρους, όπως η λέξη 'Versandetiketten', που σηµαίνει ταχυδροµική (Versand) ετικέτα (Etiketten). Τέτοιοι όροι αποσυντίθενται στα συνθετικά τους.

Για ένα ακριβές ευρετήριο, οι σύνθετοι όροι ευρετηριοποιούνται ακριβώς ως µια λέξη. Για ένα γλωσσολογικό ή διπλό ευρετήριο, απλοποιούνται στα συνθετικά τους. Στην αναζήτηση, οι σύνθετοι όροι αναζήτησης απλοποιούνται αν χρησιµοποιείται γλωσσολογικό ευρετήριο ή αν χρησιµοποιείται διπλό ευρετήριο µε την παράµετρο `STEMMED FORM OF`.

Τα συνθετικά βρίσκονται αν υπάρχουν σε οποιαδήποτε σειρά στο κείµενο αρκεί να βρίσκονται στην ίδια πρόταση. Για παράδειγμα, όταν γίνεται αναζήτηση για τη γερµανική

λέξη ‘Wetterbericht’, θα βρεθεί και το έγγραφο που περιέχει τη φράση ‘Bericht über das Wetter’.

Προσπάθεια να απλοποιηθεί στα συνθετικά του ένας σύνθετος όρος γίνεται εάν :

- Η γλώσσα στην οποία είναι ανήκει χρησιμοποιεί σύνθετες λέξεις
- Ο όρος έχει ένα ελάχιστο μήκος σε χαρακτήρες
- Ο όρος από μόνος του δεν υπάρχει σε λεξικό. Υπάρχουν κοινές σύνθετες λέξεις που υπάρχουν σε λεξικά.

Αν η απλοποίηση είναι δυνατή, τα συνθετικά μέρη του όρου ανάγονται στη βασική τους μορφή.

3.4.5 Εξαγωγή χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών περιλαμβάνει διαδικασίες σχεδιασμένες για την εξαγωγή σετ όρων από το κείμενο που μπορούν να αποθηκευτούν και να επεκτείνουν τον όρο αναζήτησης. Για παράδειγμα, μια αναζήτηση για το ‘Clinton’, θα μπορούσε να επαυξηθεί ώστε να περιλαμβάνει και το ‘president’. Η σχέση μεταξύ αυτών των όρων δεν είναι δυνατόν να είναι γνωστή από πριν, όπως συμβαίνει στα συνώνυμα για παράδειγμα. Μόνο μετά από ανάλυση των κειμένων, μπορούν να βρεθούν τέτοιες σχέσεις και να αποθηκευτούν για επαύξηση αναζητήσεων.

Οι διαδικασίες εξαγωγής χαρακτηριστικών είναι :

- Εξαγωγή ονομάτων
- Εξαγωγή domain όρων
- Εξαγωγή συντομογραφιών

3.4.5.1 Εξαγωγή ονομάτων

Η εξαγωγή ονομάτων ισχύει μόνο για τα κείμενα στην αγγλική γλώσσα. Είναι μια διαδικασία δυο σταδίων :

1. Εφαρμόζονται συγκεκριμένα *σχέδια (patterns)* στην εσωτερική παράσταση του κειμένου όπως υπάρχει μετά την αναγνώριση τεκμηρίων. Υπάρχουν διάφορα είδη ονομάτων, όπου κάθε είδος αντιστοιχεί και σε ένα σετ προτύπων που αναγνωρίζει το είδος. Τα παρακάτω είδη υποστηρίζονται :
 - Ονόματα προσώπων, όπως ‘President George Washington’, ‘Prof. John Henry’, κά.
 - Ονόματα οργανισμών, όπως ‘International Business Machines Corporation’, κά.
 - Ονόματα τοποθεσιών, όπως ‘New York (NY)’, κά.
2. Κατά τη διάρκεια της ευρετηριοποίησης, αναγνωρίζονται τα σετ ονομάτων που αναφέρονται στο ίδιο πρόσωπο ή αντικείμενο. Για παράδειγμα, το ‘President Bill Cates’ και ‘Mr. William Cates’, αναφέρονται στο ίδιο πρόσωπο και ομαδοποιούνται. Αναζήτηση για έναν από αυτούς τους όρους μπορεί να επεκταθεί και με τους συνώνυμους όρους.

3.4.5.2 Εξαγωγή όρων domain

Οι όροι domain είναι πολυ-λεξικοί όροι (τουλάχιστον δύο λέξεις) που εμφανίζονται στο κείμενο με μια ελάχιστη συχνότητα εμφάνισης. Ένα σετ από παρόμοιους όρους που εμφανίζεται στο κείμενο αποθηκεύεται και χρησιμοποιείται για επαύξηση όρων σε μια αναζήτηση. Συντομογραφίες ως μέρη ενός όρου λαμβάνονται επίσης υπόψη.

Παρακάτω ακολουθούν μερικά παραδείγματα domain όρων :

```
system board
enhanced mode
```

```

free reader
release candidate
UNIX platform
operating system
document editor
error message

```

Η επαύξηση μπορεί να συσχετίσει αυτούς τους αποθηκευμένους όρους με κάθε μέρος της ερώτησης που συμφωνεί με μέρος του εξαγόμενου domain όρου. Για παράδειγμα, ερώτηση με τον όρο 'systems' επαυξάνεται στο 'system system board operating system', που μπορεί να ανακτήσει κείμενα που δεν είναι προσπελάσιμα από τη μη επαυξημένη ερώτηση.

3.4.5.3 Εξαγωγή συντομογραφιών

Σε αυτή τη διαδικασία, γίνεται αναζήτηση στο κείμενο για όλες τις συντομογραφίες με κεφαλαίους χαρακτήρες και γίνεται προσπάθεια να αναλυθούν με βάση το περιεχόμενο του κειμένου. Ο Text Extender υποστηρίζει μια πληθώρα από πρότυπα συντομογραφιών κατά την εξαγωγή. Για παράδειγμα, όλες οι παρακάτω φράσεις αναφέρονται στο 'International Business Machines', άσχετα με τη συχνότητα των εμφανίσεών τους στο κείμενο :

```

I.B.M.
IBM
IBM Corp.
IBM Corporation
I.B.M. Corporation

```

3.5 Γλωσσολογική επεξεργασία για την ανάκτηση

Η επεξεργασία της ερώτησης έχει σκοπό να κάνει τους όρους αναζήτησης ασθενέστερους ώστε ο βαθμός ανάκτησης να αυξηθεί, να βρεθούν δηλαδή περισσότερα σχετιζόμενα κείμενα. Υπάρχουν δύο βασικές λειτουργίες πάνω στους όρους αναζήτησης που επιτυγχάνουν αυτό το σκοπό. Αυτές είναι η *επαύξηση* και η *μείωση*. Επιπρόσθετα, μερικές λειτουργίες στους όρους αναζήτησης περιλαμβάνουν και επαύξηση και μείωση.

Κατά τη διαδικασία της επαύξησης, λαμβάνεται ένας απλός ή πολυ-λεξικός όρος από ένα όρο αναζήτησης και συσχετίζεται με ένα σετ εναλλακτικών όρων, καθένας από τους οποίους μπορεί να είναι ένας πολυ-λεξικός όρος από μόνος του. Αυτοί οι όροι εφαρμόζουν ένα λογικό OR στη γλώσσα ερωτήσεων του Text Extender. Από τη στιγμή που οι επαυξήσεις δεν αλλάζουν τους όρους αναζήτησης, είναι σε κάποιο βαθμό ανεξάρτητες από το είδος του ευρετηρίου. Παραδείγματα επαυξήσεων είναι :

- Επαύξηση συνωνύμων
- Επαύξηση θησαυρού
- Επαύξηση domain όρων και συντομογραφιών

Οι μειώσεις αλλάζουν τον όρο αναζήτησης σε μια μορφή που είναι γενικότερη από αυτή που καθόρισε ο χρήστης. Λόγω του γεγονότος ότι αλλάζει τον αρχικό όρο αναζήτησης, οι μειώσεις εξαρτώνται από το είδος του ευρετηρίου ώστε να εξασφαλίσουν ότι ο αλλαγμένος όρος ταιριάζει. Έτσι, ο Text Extender λαμβάνει τα δεδομένα για τις μειώσεις από το είδος του ευρετηρίου στο οποίο απευθύνεται η ερώτηση. Παραδείγματα μειώσεων αποτελούν τα :

- *Lemmatization*
- Κανονικοποίηση

- Στοπ λέξεις

Μερικές λειτουργίες αλλάζουν και επαυξάνουν τον όρο αναζήτησης με ένα σετ εναλλακτικών όρων. Εξαιτίας της μείωσης, αυτές οι λειτουργίες εξαρτώνται και πάλι από πληροφορίες που περιέχονται στο ευρετήριο. Λειτουργίες που ανήκουν σε αυτή την κατηγορία είναι :

- Χαρακτήρες και λέξεις μάσκες
- Επαύξηση ονομάτων
- Ηχητική επαύξηση

3.5.1 Επαύξηση συνωνύμων

Τα συνώνυμα είναι νοηματικά σχετιζόμενες λέξεις. Συνήθως, αυτές οι λέξεις ανήκουν στην ίδια κατηγορία (όπως ουσιαστικά, ρήματα, κτλ.) με αυτή του αρχικού όρου. Τα συνώνυμα λαμβάνονται από ένα ξεχωριστό αρχείο για κάθε γλώσσα. Χρησιμοποιείται πάντα η βασική τους μορφή και, εκτός μερικών εξαιρέσεων, δεν είναι πολυ-λεξικοί όροι. Οι όροι αναζήτησης πάντα ανάγονται στη βασική τους μορφή όταν γίνεται αναζήτηση συνωνύμων τους.

3.5.2 Επαύξηση θησαυρού

Ένας όρος αναζήτησης μπορεί να επαυξηθεί χρησιμοποιώντας όρους από *θησαυρό* (*thesaurus*) που μπορούν να προσπελαστούν μέσω μιας σχέσης. Αυτές οι σχέσεις μπορεί να είναι ιεραρχικές, συσχετιστικές ή σχέσεις συνωνύμων (βλέπε §3.8.2). Ένας όρος θησαυρού μπορεί να είναι, και συχνά είναι, ένας πολυ-λεξικός όρος.

3.5.3 Επαύξηση domain όρων και συντομογραφιών

Ένας domain όρος ή μια συντομογραφία ως μέρος των όρων αναζήτησης επαυξάνονται με όλους τους domain όρους ή συντομογραφίες που είναι αποθηκευμένες σε ένα *ευρετήριο χαρακτηριστικών* (*feature index*) που έχουν την ίδια κανονική μορφή. Και οι αρχικοί και οι επαυξημένοι όροι μπορεί να είναι πολυ-λεξικοί όροι.

3.5.4 Χαρακτήρες και λέξεις μάσκες

Η διαδικασία εφαρμογής των χαρακτήρων και λέξεων μασκών είναι διαδικασία μη γλωσσολογική, κατά την οποία μια *κανονική έκφραση* (*regular expression*) αντικαθίσταται με τη διάζευξη όλων των ευρετηριοποιημένων λέξεων που την ικανοποιούν. Καμία από τις εκφράσεις μάσκες ή από τις επαυξημένες μορφές τους υπόκεινται σε lemmatization, εξαγωγή στοπ λέξεων ή σε οποιαδήποτε από τις άλλες τεχνικές επαύξησης. Αυτό μπορεί να έχει το αποτέλεσμα ότι, για παράδειγμα, ένα ανώμαλο ρήμα στα αγγλικά όπως το *swum*, όταν γίνεται αναζήτηση με χαρακτήρα μάσκα *swu**, εφαρμόζεται σε ένα ακριβές ευρετήριο, αλλά όχι σε γλωσσολογικό, στο οποίο η μορφή αυτή έχει ευρετηριοποιηθεί ως *swim*.

3.5.5 Επαύξηση ονομάτων

Μια παραλλαγή ενός ονόματος από τους όρους αναζήτησης μπορεί να επαυξηθεί χρησιμοποιώντας όλα τα κανονικοποιημένα ονόματα που έχουν συμπεριληφθεί στο ευρε-

τήριο. Ο αρχικός όρος, μια παραλλαγή ενός ονόματος, μπορεί να είναι πολυ-λεξικός όρος και μπορεί να υπάρχει οποιοσδήποτε αριθμός παραλλαγών στον όρο αναζήτησης.

Για λόγους βελτιστοποίησης χώρου, ένα κανονικοποιημένο όνομα που εμφανίζεται πολύ σπάνια στο κείμενο, μπορεί να διαγραφεί από το ευρετήριο μελλοντικά. Αυτό έχει το αποτέλεσμα ότι αυτή η επαύξηση δεν συμβαίνει, αλλά δεν επηρεάζει την αναζήτηση.

Η εξαγωγή των στοπ λέξεων, το lemmatization και η κανονικοποίηση δεν εφαρμόζονται σε παραλλαγές ή κανονικοποιημένα ονόματα.

3.5.6 Ηχητική επαύξηση

Αυτή η διαδικασία επαυξάνει λέξεις από ένα σύνολο ηχητικά παρόμοιων λέξεων. Χρησιμοποιείται σε περιπτώσεις που η ακριβής ορθογραφία του όρου αναζήτησης δεν είναι γνωστή.

3.6 Γλωσσολογική επεξεργασία για την πλοήγηση

Γλωσσολογική επεξεργασία γίνεται και όταν γίνεται *πλοήγηση (browsing)* των αποτελεσμάτων. Η διαδικασία αυτή γίνεται σε δυο στάδια :

- Βασική ανάλυση κειμένου : κανονικοποίηση και επαύξηση όρων
- *Εκτεταμένη ταύτιση (extended matching)*

3.6.1 Πρώτο στάδιο

Το πρώτο στάδιο γίνεται χωρίς τη βοήθεια ηλεκτρονικού λεξικού.

Η κανονικοποίηση γίνεται όπως περιγράφεται παραπάνω.

Η επαύξηση όρων είναι η αντίστροφη διαδικασία της μείωσης του όρου στη βασική του μορφή. Αν το είδος του ευρετηρίου είναι το γλωσσολογικό ή το διπλό και η αναζήτηση γίνεται με παράμετρο το `STEMMED FORM OF`, τότε οι όροι αναζήτησης ανάγονται στη βασική τους μορφή πριν αρχίσει η αναζήτηση.

Με παρόμοιο τρόπο, αν χρησιμοποιείται γλωσσολογικό ή διπλό ευρετήριο, οι όροι του κειμένου ανάγονται στη βασική τους μορφή πριν ευρετηριοποιηθούν. Έτσι, τα κείμενα βρίσκονται με βάση τη βασική μορφή των όρων.

Ωστόσο, όταν ο χρήστης βλέπει τα αποτελέσματα της ερώτησής του, περιμένει να δει όλες τις παραλλαγές των βασικών όρων. Για να ικανοποιηθεί αυτή η απαίτηση, οι βασικοί όροι επαυξάνονται.

Όλες οι παραλλαγές που υπάρχουν στα λεξικά μπορούν να παραχθούν.

3.6.2 Δεύτερο στάδιο

Το δεύτερο στάδιο είναι η εκτεταμένη ταύτιση, που χρησιμοποιείται στις σπάνιες περιπτώσεις κατά τις οποίες η βασική ανάλυση του κειμένου και η κανονικοποίηση δεν μπορούν να *τονίσουν (highlight)* έναν όρο που βρέθηκε. Αυτό το στάδιο, βρίσκει τις πιο «παράξενες» ταυτίσεις.

Η εκτεταμένη ταύτιση ενεργοποιείται καθορίζοντας το `DES_EXTENDED` ως παράμετρο στη συνάρτηση `DesOpenDocument` του API.

Η διαδικασία χρησιμοποιεί την ίδια γλωσσολογική επεξεργασία που γίνεται και στη διάρκεια της ευρετηριοποίησης.

Η εκτεταμένη ταύτιση μπορεί να βρει επιπλέον περιπτώσεις όταν :

- Ο όρος αναζήτησης χρησιμοποιεί χαρακτήρες μάσκες και είναι λέξη που κλίνεται. Οι χαρακτήρες μάσκες αναλύονται, γίνεται αποκοπή των κοινότυπων καταλήξεων (*stemming*) και γίνεται αναζήτηση. Χωρίς την εκτεταμένη ταύτιση, κείμενο που συμφωνεί με τον δοθέντα όρο αναζήτησης δεν θα ανακτώνταν.
Παράδειγμα : Έγγραφο περιέχει την κλιμένη λέξη 'swam'. Κατά τη διάρκεια της ευρετηριοποίησης ο όρος ευρετηριοποιείται ως swim. Αν ο όρος αναζήτησης είναι το 'swi%', το παραπάνω έγγραφο ανακτάται διότι από την αποκοπή προκύπτει το 'swim'. Χωρίς την εκτεταμένη ταύτιση, μόνο οι λέξεις που συμφωνούν με το 'swi%' τονίζονται, ενώ με αυτή και η κλιμένη λέξη 'swam' τονίζεται.
- Ευρετηριοποιούνται σύνθετες λέξεις. Όταν ένα κείμενο για παράδειγμα στη γερμανική γλώσσα περιέχει σύνθετες λέξεις και ευρετηριοποιείται χρησιμοποιώντας γλωσσολογικό ευρετήριο, το κείμενο διατηρεί τα μέρη της λέξης και την ίδια τη σύνθετη λέξη. Όταν γίνεται αναζήτηση για ένα συνθετικό της σύνθετης λέξης, τα έγγραφα που περιέχουν τη σύνθετη λέξη βρίσκονται, αλλά χωρίς την εκτεταμένη ταύτιση, η λέξη δεν τονίζεται.
Παράδειγμα : Κείμενο περιέχει τη γερμανική λέξη 'Apfelbaum' (apple tree). Κατά την ευρετηριοποίηση με γλωσσολογικό ευρετήριο, η λέξη μειώνεται στις 'apfel' και 'baum'. Όταν γίνει αναζήτηση για τον όρο 'baum', βρίσκεται μέσω του ευρετηρίου ο όρος 'baum', καθώς και το έγγραφο που τον περιέχει. Με την εκτεταμένη ταύτιση, ο σύνθετος όρος 'Apfelbaum' απλοποιείται στα συνθετικά του και ο όρος 'baum' τονίζεται.
- Γίνεται συλλαβισμός των λέξεων στο τέλος της γραμμής. Αν η παύλα (-) εισάγεται αυτόματα από ένα πρόγραμμα επεξεργασίας κειμένου, η συλλαβισμένη λέξη μπορεί να βρεθεί και να τονιστεί. Αν ωστόσο ο χρήστης εισάγει την παύλα, τα έγγραφα που περιέχουν τη λέξη βρίσκονται, αλλά χωρίς την εκτεταμένη ταύτιση, η λέξη δεν τονίζεται.
Παράδειγμα : Κείμενο περιέχει τη λέξη 'container', συλλαβισμένη στο τέλος της γραμμής ως εξής :
Another name for a folder is a con-
tainer.

Κατά την ευρετηριοποίηση η λέξη κανονικοποιείται στο 'container'. Όταν γίνεται αναζήτηση για το 'container' βρίσκονται και ο όρος και το κείμενο που τον περιέχει. Γίνεται προσπάθεια για να τονιστούν οι λέξεις που συμφωνούν με το 'container'. Χωρίς την εκτεταμένη ταύτιση, προσαρμογή γίνεται μόνο αν η παύλα είχε εισαχθεί από επεξεργαστή κειμένου κι όχι από το χρήστη.

3.7 Γλωσσολογικές λειτουργίες που υποστηρίζει ο Text Extender ανά γλώσσα

Ο πίνακας 3.3 δείχνει τις γλωσσολογικές λειτουργίες ανά γλώσσα που υποστηρίζει ο Text Extender. Δυστυχώς, στην έκδοση 5.2 του Text Extender δεν υποστηρίζεται η ελληνική γλώσσα.

Γλώσσα	Γλωσσολογικές λειτουργίες
Βραζιλιάνικα	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Καναδέζικα γαλλικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Καταλανικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Δανέζικα	Όλες οι λειτουργίες

Ολλανδικά	Όλες οι λειτουργίες
Φινλανδικά	Βασική ανάλυση κειμένου, κανονικοποίηση, αποσυλλαβισμός και φιλτράρισμα στοπ λέξεων
Γαλλικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Γερμανικά	Όλες οι λειτουργίες
Ισλανδικά	Αναγωγή στη βασική μορφή, αποσυλλαβισμός, φιλτράρισμα στοπ λέξεων και φιλτράρισμα ως μέρος του λόγου δεν υποστηρίζονται
Ιταλικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Πορτογαλικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Ισπανικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Σουηδικά	Όλες οι λειτουργίες
Ελβετικά γερμανικά	Όλες οι λειτουργίες
Βρετανικά αγγλικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων
Αμερικάνικα αγγλικά	Δεν υποστηρίζεται η απλοποίηση των σύνθετων λέξεων

Πίνακας 3.3: Γλωσσολογικές λειτουργίες ανά γλώσσα

3.8 Αρχές και έννοιες θησαυρού

Ο Text Extender επιτρέπει την επαύξηση όρου με την προσθήκη επιπλέον όρων από θησαυρό. Ο θησαυρός είναι ελεγχόμενο λεξικό από νοηματικά σχετιζόμενους όρους που συνήθως καλύπτουν ένα συγκεκριμένο τομέα θεμάτων.

Οι βασικές έννοιες ενός θησαυρού του Text Extender είναι οι *όροι* και οι *σχέσεις*.

3.8.1 Όροι

Όρος ονομάζεται μια λέξη ή έκφραση που δηλώνει μια έννοια σε μια περιοχή ενδιαφέροντος του θησαυρού. Για παράδειγμα, οι επόμενες φράσεις θα μπορούσαν να αποτελούν όρους του θησαυρού :

```
data processing
helicopter
automobile industry
```

Οι όροι διακρίνονται στους *περιγραφικούς* και *μη-περιγραφικούς*. *Περιγραφικός (descriptor)* είναι ένας όρος σε μια κλάση συνωνύμων που προτιμάται για ευρετηριοποίηση και αναζήτηση. Οι άλλοι όροι της κλάσης ονομάζονται *μη-περιγραφικοί (non-descriptors)*. Για παράδειγμα, η λέξη 'δεελφίνι' είναι μη-περιγραφικός όρος, που ανήκει στην κλάση του περιγραφικού όρου 'θηλαστικά'.

3.8.2 Σχέσεις

Μια *σχέση (relation)* είναι η νοηματική συσχέτιση μεταξύ δύο όρων. Οι σχέσεις έχουν τις εξής ιδιότητες :

- Το *βάθος* της σχέσης είναι ο αριθμός των επιπέδων στα οποία αυτή εκτείνεται. Αυτό καθορίζεται στη σύνταξη του ερωτήματος αναζήτησης με τη λέξη-κλειδί COUNT στην ερώτηση που δίνεται στον Extender.
- Η κατευθυντικότητα της σχέσης καθορίζει εάν η σχέση εκτείνεται από τον ένα όρο στον άλλο και αντίστροφα (δι-κατευθυντική), ή μόνο από τον ένα όρο στον άλλο.

Η επαύξηση του θησαυρού μπορεί να χρησιμοποιήσει κάθε σχέση που είναι ορισμένη σε αυτό. Μπορεί επίσης να καθοριστεί το βάθος της επαύξησης. Αυτό είναι ο μέγι-

στος αριθμός μεταβάσεων από τον πηγαίο όρο στον όρο-στόχο. Πρέπει να σημειωθεί ότι καθώς το βάθος αυξάνει, αυξάνει εκθετικά και το μέγεθος του σετ των όρων.

Το επόμενο παράδειγμα δείχνει την αύξηση του αριθμού των όρων καθώς αυξάνει και το βάθος. Το παράδειγμα δείχνει ότι η λέξη 'medicine' είναι ένα επίπεδο κάτω από την 'health', καθώς και το 'allergology' είναι ένα επίπεδο κάτω από τα 'health service', 'paramedical', 'medicine' και 'illness' και δύο επίπεδα κάτω από το 'health'.

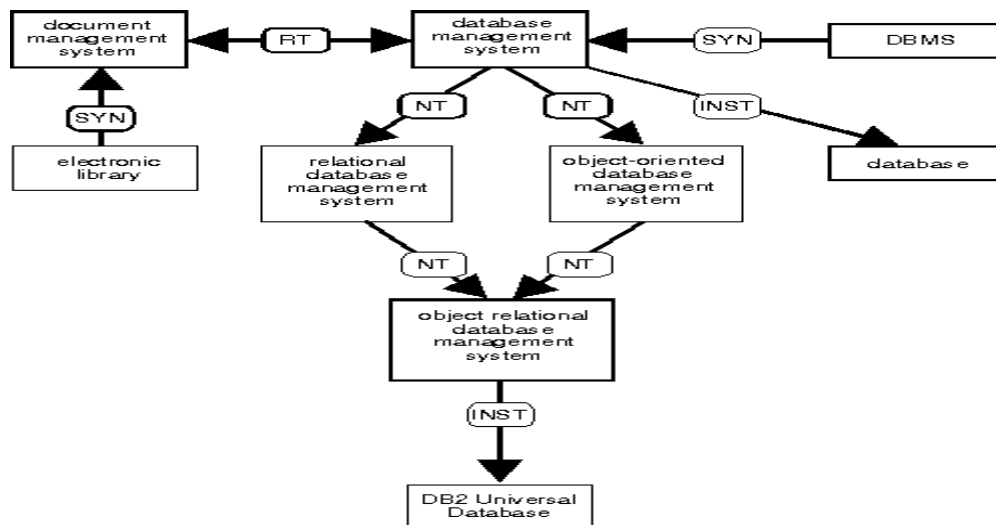
health

health service, paramedical, medicine, illness

allergology, virology, veterinary medicine, toxicology, surgery, stomatology, rheumatology, radiotherapy, psychiatry, preventive medicine, pathology, odontology, nutrition, nuclear medicine, neurology, nephrology, medical check up, industrial medicine, hematology, general medicine, epidemiology, clinical trial, cardiology, cancerology

Τύποι σχέσεων είναι :

- Συσχετιστική (*associative*)
- Σχέση συνωνύμων
- Ιεραρχική σχέση
- Οποιοσδήποτε άλλος τύπος σχέσης



Σχήμα 3.3: Παράδειγμα δομής θησαυρού

Για διευκόλυνση, μπορεί να δοθεί σε κάθε σχέση ένα μνημονικό όνομα, όπως BROADER TERM (ευρύτερος όρος), που μπορεί να έχει τη συντομογραφία BT. Τα μνημονικά ονόματα που χρησιμοποιούνται στο λεξικό είναι :

- BT = Broader Term (ευρύτερος όρος)
- NT = Narrower Term (στενότερος όρος)
- RT = Related Term (σχετιζόμενος όρος)

- SYN = Synonym (συνώνυμο)
- USE (χρησιμοποιεί)
- UF = USE FOR (χρησιμοποιείται για)

Το σχήμα 3.3 δείχνει τη δομή ενός απλού και μικρού θησαυρού. Οι όροι συμβολίζονται με τα μεγάλα ορθογώνια, ενώ οι σχέσεις με μικρά ορθογώνια με στρογγυλεμένες άκρες.

3.8.2.1 Συσχετιστικές σχέσεις

Μια συσχετιστική σχέση είναι μια σχέση δυο κατευθύνσεων μεταξύ περιγραφικών όρων, εκτεινόμενη σε οποιοδήποτε βάθος. Δεσμεύει δυο όρους που δεν είναι ούτε ισότιμοι ούτε ιεραρχικοί, ωστόσο συσχετίζονται νοηματικά σε τέτοιο βαθμό, ώστε ο σύνδεσμος μεταξύ τους να επιτρέπει και άλλους όρους για ευρετηριοποίηση και αναζήτηση.

Οι συσχετιστικές σχέσεις συμβολίζονται με RT (related term).

Παράδειγμα : dog RT security

3.8.2.2 Συνώνυμα

Η σχέση μεταξύ συνωνύμων είναι σχέση μιας κατευθύνσεως μεταξύ δύο όρων που έχουν ίδιο ή παρόμοιο νόημα. Σε μια κλάση συνωνύμων, ένας από τους όρους προσδιορίζεται ως περιγραφικός. Οι υπόλοιποι προσδιορίζονται ως μη-περιγραφικοί.

Η περιγραφή USE καθοδηγεί από ένα μη-περιγραφικό όρο στον αντίστοιχο περιγραφικό του. Η περιγραφή USE FOR καθοδηγεί από τον περιγραφικό σε κάθε μη-περιγραφικό όρο.

Παράδειγμα : feline³ USE cat, lawyer UF advocate

3.8.2.3 Ιεραρχικές σχέσεις

Ιεραρχική είναι μια σχέση μιας κατευθύνσεως μεταξύ περιγραφικών όρων που υποδηλώνει ότι ένας από τους όρους είναι πιο ειδικός ή πιο γενικός από τον άλλο. Αυτή η διάκριση έχει ως αποτέλεσμα την αναπαράσταση των όρων σε μια ιεραρχία, όπου ένας όρος αντιπροσωπεύει μια κλάση και οι όροι συνιστώσες αναφέρονται στα μέλη. Για παράδειγμα, ο όρος 'ποντίκι' ανήκει στην κλάση 'τρωκτικό'.

Οι περιγραφές BROADER TERM και NARROWER TERM αντιπροσωπεύουν ιεραρχικές σχέσεις.

Παράδειγμα : car NT limousine, equine BT horse

3.8.2.4 Άλλες σχέσεις

Μια σχέση που ανήκει στην κατηγορία αυτή είναι η πιο γενική. Αντιπροσωπεύει σχέση που δεν μπορεί εύκολα να συμπεριληφθεί σε μια από τις παραπάνω κατηγορίες. Μια τέτοια σχέση μπορεί να είναι μιας ή και δυο κατευθύνσεων, δεν υπάρχει περιορισμός βάθους και σχέσεις μπορούν να υπάρξουν μεταξύ περιγραφικών και μη όρων.

Αυτό το είδος της σχέσης χρησιμοποιείται όταν προστίθενται νέοι όροι στο θησαυρό μέχρι να καθοριστεί η σωστή σχέση με τους άλλους όρους.

3.9 Διαχειριστικές εργασίες του Text Extender

Σε αυτή την ενότητα θα περιγραφεί πώς γίνεται η διαχείριση και η διαδικασία ευρετηριοποίησης. Οι οδηγίες που περιγράφονται παρακάτω έχουν δοκιμαστεί σε λειτουργία.

³ Feline : αιλουροειδές

γικό σύστημα Windows NT 4 Server (service pack 6a), IBM DB2 UDB v5.2 (fixpak 12) και DB2 Extenders v5.2.

Υπάρχουν ενέργειες που πρέπει να εκτελεστούν τόσο στον διακομιστή όσο και στον πελάτη.

Ο Text Extender έχει την ίδια αρχή λειτουργίας όπως και η DB2. Λειτουργεί με στιγμιότυπα. Κάθε στιγμιότυπο προσφέρει ένα απομονωμένο περιβάλλον διαχείρισης στο οποίο διατηρούνται ευρετήρια που αποθηκεύονται ξεχωριστά σε σχέση με τα άλλα στιγμιότυπα.

Η μεταβλητή περιβάλλοντος `DB2INSTANCE` του λειτουργικού συστήματος, καθορίζει με ποιο στιγμιότυπο θα δουλέψει η DB2. Στη συνέχεια μπορεί να ξεκινήσει η DB2, καθώς και ο Text Extender. Αυτή είναι η διαδικασία που ακολουθείται σε γενικές γραμμές.

Οι διαχειριστικές εντολές του Text Extender δίνονται από τη γραμμή εντολών των Windows NT (κονσόλα στα UNIX συστήματα).

3.9.1 Εργασίες στον διακομιστή Text Extender

Στο παράρτημα (σελ. 115) περιγράφονται σε λίγες γραμμές οι διαχειριστικές εντολές του Text Extender τόσο από την πλευρά του πελάτη, όσο και από την πλευρά του διακομιστή.

3.9.1.1 Δημιουργία στιγμιότυπου Text Extender

Αυτό γίνεται με το εργαλείο γραμμής εντολών (*command line tool*) `txicrt`. Σε περιβάλλον Windows NT, η γενική σύνταξη της εντολής είναι :

```
txicrt db2-instance-name
```

όπου *db2-instance-name* είναι το όνομα του στιγμιότυπου της DB2.

Η αντίστοιχη εντολή για την *κατάργηση (drop)* ενός στιγμιότυπου είναι η `txidrop`. Στα Windows NT, η γενική σύνταξή της είναι :

```
txidrop instance-name
```

όπου *instance-name* είναι το όνομα του στιγμιότυπου του Text Extender.

3.9.1.2 Διαχείριση του Text Extender διακομιστή

Η εκκίνηση του διακομιστή γίνεται με την εντολή `txstart`. Ο τερματισμός του γίνεται με την εντολή `txstop`. Την κατάσταση του μπορούμε να τη δούμε με την εντολή `txstatus`.

Οι παραπάνω εργασίες, η διαχείριση των στιγμιότυπων και το ξεκίνημα - σταμάτημα του Text Extender διακομιστή αποτελούν εργασίες που πρέπει να γίνουν στον διακομιστή που είναι εγκατεστημένη η DB2 με τον Text Extender.

3.9.2 Εργασίες στον πελάτη Text Extender

Προκειμένου να αρχίσουν οι εργασίες στον πελάτη Text Extender, πρέπει ο διακομιστής του Text Extender να έχει ξεκινήσει στον διακομιστή DB2. Οι εργασίες στον πελάτη προϋποθέτουν την ύπαρξη ενός στιγμιότυπου βάσης DB2. Οι εργασίες που πρέπει να γίνουν σε γενικές γραμμές είναι :

1. Εκκίνηση της γραμμής εντολών του Text Extender και προαιρετικά σύνδεση με τη βάση.

2. Προετοιμασία εγγράφων για αναζήτηση. Η διαδικασία αυτή περιέχει ένα αριθμό υπο-διαδικασιών που πρέπει να γίνουν πριν δοθούν ερωτήσεις για αναζήτηση στα έγγραφα.
 - 2.1. Τροποποίηση της παραμετροποίησης του κειμένου (προαιρετικό)
 - 2.2. Τροποποίηση της λίστας των στοπ λέξεων και συντομογραφιών (προαιρετικό)
 - 2.3. Ενεργοποίησης της βάσης για χρήση από τον Text Extender
 - 2.4. Δημιουργία πινάκων επίδειξης (*demo*) (προαιρετικό)
 - 2.5. Ενεργοποίηση πίνακα για χρήση από τον Text Extender (προαιρετικό)
 - 2.6. Ενεργοποίηση στήλης πίνακα που έχει οριστεί για το κείμενο
 - 2.7. Ενεργοποίηση εξωτερικών αρχείων κειμένου για χρήση από τον Text Extender
3. Αντιστροφή της διαδικασίας προετοιμασίας που μπορεί να περιλαμβάνει απενεργοποίηση της στήλης πίνακα που ορίστηκε για το κείμενο, απενεργοποίηση των εξωτερικών αρχείων, απενεργοποίηση των πινάκων και τέλος απενεργοποίηση της βάσης για χρήση από τον Text Extender.
4. Συντήρηση των ευρετηρίων που περιλαμβάνει :
 - 4.1. Ενημέρωση ενός ευρετηρίου
 - 4.2. Αλλαγή των ρυθμίσεων ενός ευρετηρίου
 - 4.3. Επαναφορά της κατάστασης του ευρετηρίου
 - 4.4. Διαγραφή γεγονότων των ευρετηρίων
5. Εμφάνιση πληροφοριών όπως :
 - 5.1. Η κατάσταση της βάσης ή πίνακα ή στήλης, αν είναι ενεργοποιημένα ή όχι
 - 5.2. Οι τιμές των μεταβλητών περιβάλλοντος
 - 5.3. Οι τιμές των ρυθμίσεων παραμετροποίησης
 - 5.4. Η κατάσταση των ευρετηρίων
 - 5.5. Λάθη
 - 5.6. Οι ρυθμίσεις των ευρετηρίων
 - 5.7. Οι ρυθμίσεις του κειμένου για μια στήλη

3.9.2.1 Αναλυτική διαδικασία των εργασιών στον πελάτη

Το παράθυρο γραμμής εντολών του Text Extender ανοίγει μέσω του εικονιδίου DB2TX Command Line Processor. Αυτό γίνεται κάθε φορά που πρέπει να γίνουν διαχειριστικές εργασίες στον Text Extender. Το επόμενο βήμα είναι να ανοιχτεί μια *σύνοδος* (*session*) με τη βάση.

Στη συνέχεια, πρέπει να οριστούν οι παράμετροι του κειμένου που θα ευρετηριοποιηθεί. Οι παράμετροι έχουν να κάνουν σχέση με το κωδικοποιημένο σετ χαρακτήρων⁴, τη γλώσσα των κειμένων, τη μορφοποίησή τους, το είδος του ευρετηρίου που θα χρησιμοποιηθεί, κάθε πότε θα ενημερώνεται το ευρετήριο και που θα αποθηκεύονται τα αρχεία του (ένας κατάλογος σε σύστημα αρχείων), μια παράμετρο που καθορίζει αν η ευρετηριοποίηση των κειμένων θα αρχίσει απευθείας μετά την εκτέλεση της εντολής και τέλος με ένα αριθμό που έχει σχέση με τον αριθμό των εισαγωγών ή ενημερώσεων στη βάση για τον οποίο η βάση θα *οριστικοποιεί* (*commit*) τις αλλαγές. Η εντολή που αλλάζει τις παραμέτρους του κειμένου είναι η `CHANGE TEXT CONFIGURATION`. Επειδή αυτές οι παράμετροι έχουν εξ' ορισμού τιμές, η εκτέλεση της εντολής αυτής από τη γραμμή εντολών δεν είναι υποχρεωτική⁵. Οι εξ' ορισμού τιμές είναι : CCSID είναι το 850, γλώσσα είναι η αγγλική

⁴ Coded Character Set Id (CCSID) : Είναι η κωδικοσελίδα που χρησιμοποιεί η DB2 για να αποθηκεύει τα δεδομένα. Ο Text Extender εφόσον αποτελεί μια εφαρμογή που δουλεύει για τη DB2, πρέπει να χρησιμοποιεί το ίδιο CCSID με εκείνο της DB2.

⁵ Αν η βάση δεν είναι ενεργοποιημένη, η εκτέλεση της εντολής θα αποτύχει, για αυτό το λόγο στο στάδιο αυτό η εντολή δεν είναι υποχρεωτική.

(Αμερικής), μορφοποίηση είναι το TDS (ASCII κείμενο), ευρετήριο είναι το ngram το οποίο δεν ενημερώνεται αυτόματα, κατάλογος αποθήκευσης των αρχείων του ευρετηρίου είναι ο `drive:\dmb\instance\db2\db2tx\indexes` (όπου *drive* είναι το γράμμα δίσκου στο οποίο έχουν εγκατασταθεί οι Extenders της DB2), η ευρετηριοποίηση αρχίζει μετά την εκτέλεση της εντολής και ο αριθμός για την οριστικοποίηση των αλλαγών είναι 0 (γίνεται δηλαδή απευθείας commit).

Το επόμενο (προαιρετικό) βήμα είναι η αλλαγή της λίστας των στοπ λέξεων και των αρχείων συντομογραφιών. Οι λίστες των στοπ λέξεων και των συντομογραφιών είναι σε αρχεία. Σε OS/2 και Windows NT περιβάλλοντα, τα αρχεία αυτά βρίσκονται στον κατάλογο `drive:\dmb\db2tx\dict`.

Αυτό που πρέπει να γίνει υποχρεωτικά στη συνέχεια είναι η ενεργοποίηση της βάσης για χρήση από τον Text Extender. Αυτό γίνεται με την εντολή `ENABLE DATABASE`. Η εντολή αυτή δηλώνει συναρτήσεις και τύπους ορισμένους από το χρήστη στη βάση. Οι συναρτήσεις είναι αυτές που θα χρησιμοποιηθούν σε SQL ερωτήσεις για αναζήτηση σε κείμενο. Επίσης, δημιουργούνται δύο όψεις με ονόματα `TEXTCOLUMNS` και `TEXTINDEXES`, που κρατούν τις στήλες και τα ονόματα των ευρετηρίων. Η βάση παραμένει ενεργοποιημένη μέχρι να δοθεί η εντολή απενεργοποίησης `DISABLE DATABASE`.

Σε αυτό το στάδιο θα μπορούσαν να δημιουργηθούν οι πίνακες επίδειξης στο σχήμα⁶ `DB2TX` της βάσης (που δημιουργείται με την ενεργοποίησή της). Για το σκοπό αυτό υπάρχουν δύο *σενάρια (script)* που δημιουργούν μια βάση επίδειξης, καθώς και κάποιους πίνακες. Τα σενάρια αυτά υπάρχουν στον κατάλογο `DMBMPATH\samples`, όπου `DMBMPATH` είναι μια μεταβλητή περιβάλλοντος που δείχνει σε ποιο κατάλογο εγκαταστάθηκαν οι Extenders της DB2. Το σενάριο `txverify.cmd` δημιουργεί μια βάση επίδειξης και την ενεργοποιεί, ενώ το `txsample.cmd` δημιουργεί τους πίνακες. Οι πίνακες αυτοί χρησιμοποιούν ευρετήρια διαφόρων τύπων και γεμίζουν με δεδομένα. Και τα δύο σενάρια πρέπει να τρέξουν από ένα *παράθυρο εντολών (command window)* της DB2.

Κατά τη δημιουργία των πινάκων, πρέπει να σημειωθεί ότι οι υποστηριζόμενοι τύποι δεδομένων προς ευρετηριοποίηση από τον Text Extender είναι οι εξής : `CHAR`, `GRAPHIC`, `VARGRAPHIC`, `LONG VARGRAPHIC`, `DBCLOB`, `VARCHAR`, `LONG VARCHAR` και `CLOB`.

Σε περίπτωση που δεν εκτελεστούν αυτά τα σενάρια πρέπει να ενεργοποιηθεί ένα νέο ή υπάρχων στιγμιότυπο βάσης και να δημιουργηθούν κατά τα γνωστά πίνακες οι οποίοι προορίζονται για ευρετηριοποίηση κειμένου. Σε αυτό το σημείο, όσον αφορά την ευρετηριοποίηση ή όχι του πίνακα, πρέπει να ξεκαθαριστεί το γεγονός αν θα χρησιμοποιηθεί ένα κοινό ευρετήριο για όλες τις στήλες του πίνακα που προορίζονται για ευρετηριοποίηση ή αν θα χρησιμοποιηθεί ξεχωριστό για καθεμιά στήλη. Η διαφορά μεταξύ των δύο είναι ότι στην περίπτωση των ξεχωριστών ευρετηρίων, υπάρχει περισσότερη ευελιξία σε σχέση με την περίπτωση του κοινού ευρετηρίου, διότι εφόσον υπάρχει ξεχωριστό ευρετήριο για κάθε στήλη του πίνακα, μπορεί έτσι να έχει διαφορετικές ρυθμίσεις από τα ευρετήρια των άλλων στηλών, ενώ στο κοινό ευρετήριο, οι ρυθμίσεις του ισχύουν για όλες τις στήλες του πίνακα που ευρετηριοποιεί. Η ξεχωριστή ευρετηριοποίηση κάθε στήλης έχει και περισσότερο διαχειριστικό φόρτο. Επομένως, αν είναι επιθυμητό να υπάρχει κοινό ευρετήριο, τότε σε αυτό το στάδιο πρέπει να δοθεί η εντολή `ENABLE TEXT TABLE`, που ενεργοποιεί τον πίνακα. Η εντολή δέχεται ως παραμέτρους το είδος του ευρετηρίου που θα χρησιμοποιηθεί, κάθε πότε θα γίνεται ενημέρωσή του, παράμετρος για το αν θα υπάρξει εξαγωγή χαρακτηριστικών (μόνο για το γλωσσολογικό ευρετήριο) ή αν θα γίνεται

⁶ Σχήμα (schema) είναι μια ομάδα αντικειμένων της βάσης που μπορεί να αποτελείται από πίνακες, συναρτήσεις, κτλ.. Το σχήμα ανήκει σε ένα χρήστη ο οποίος μπορεί να ελέγχει την πρόσβαση στα δεδομένα και στα αντικείμενα εντός του σχήματος.

διάκριση μεταξύ κεφαλαίων–πεζών (μόνο για το ngram ευρετήριο), καθώς και τον κατάλογο που θα αποθηκεύονται τα αρχεία των ευρετηρίων. Η διαδικασία απενεργοποίησης του πίνακα γίνεται με την `DISABLE TEXT TABLE`.

Το επόμενο στάδιο είναι η διαδικασία κατά την οποία γίνεται ενεργοποίηση των στηλών πινάκων που προορίζονται για ευρετηριοποίηση. Η εντολή που δίνεται είναι η `ENABLE TEXT COLUMN`. Ακόμη κι αν χρησιμοποιείται κοινό ευρετήριο, πρέπει να δοθεί ρητά η εντολή αυτή για κάθε στήλη κειμένου (στήλη του πίνακα που είναι του τύπου «κειμένου» και στην οποία αποθηκεύεται το προς ευρετηριοποίηση κείμενο). Οι παράμετροι που δέχεται η εντολή είναι αυτές που περιγράφηκαν στην εντολή `CHANGE TEXT CONFIGURATION` παραπάνω. Κατά τη διάρκεια εκτέλεσης της εντολής, συμβαίνουν διάφορα γεγονότα. Στον πίνακα όπου εκτελείται η `ENABLE TEXT COLUMN`, προστίθεται μια στήλη τύπου `VARCHAR(60)`, που λέγεται *στήλη χειριστής (handle column)*, που περιέχει πληροφορίες σχετικά με τη στήλη που ενεργοποιείται. Οι πληροφορίες αυτές είναι ένας αριθμός που χαρακτηρίζει μονοσήμαντα το κείμενο που ευρετηριοποιείται, η γλώσσα του κειμένου, η μορφοποίησή του, το CCSID, καθώς και το όνομα ενός ευρετηρίου. Στην περίπτωση που η στήλη κειμένου είναι ένα δείκτης σε αρχείο (δηλαδή γίνεται ευρετηριοποίηση αρχείου εκτός βάσης), τότε η στήλη χειριστής είναι του τύπου `VARCHAR(210)`. Επίσης, δημιουργείται ένας πίνακας με όνομα `TLGIXXXXXXX`, όπου `IXXXXXXX` είναι το όνομα του ευρετηρίου. Δημιουργούνται επίσης *σκανδάλες (triggers⁷)* στον πίνακα αυτό που εκτελούνται κάθε φορά που γίνονται αλλαγές στη στήλη κειμένου για να ενημερώνουν τον πίνακα κατάλληλα. Τέλος, δημιουργείται το ίδιο το ευρετήριο με όνομα `IXXXXXXX` που ευρετηριοποιεί τη στήλη κειμένου. Η απενεργοποίηση της στήλης γίνεται με την `DISABLE TEXT COLUMN`.

Όπως αναφέρθηκε παραπάνω, υπάρχει δυνατότητα ευρετηριοποίησης αρχείων συστήματος, εκτός βάσης. Η εντολή για την ενεργοποίηση αρχείων είναι η `ENABLE TEXT FILES` και παίρνει τις γνωστές παραμέτρους που παίρνει και η `ENABLE TEXT COLUMN`. Σε αυτή την περίπτωση, δεν είναι αναγκαία η ύπαρξη στήλης του τύπου κειμένου του πίνακα που ευρετηριοποιείται, εφόσον η εντολή θα προσθέσει το χειριστή για τα εξωτερικά αρχεία. Αφού γίνει η ευρετηριοποίηση, το αρχείο του κειμένου μπορεί να διαγραφεί ή να αλλάξει τοποθεσία, ωστόσο οι ερωτήσεις θα εκτελούνται κανονικά. Η απενεργοποίηση γίνεται με την αντίστοιχη εντολή `DISABLE TEXT FILES`.

Σε αυτό το σημείο, έχει τελειώσει το πρώτο μέρος της διαδικασίας για την ευρετηριοποίηση. Στην περίπτωση που η ενεργοποίηση των πινάκων, στηλών ή αρχείων γίνει με την παράμετρο να μην γίνει απευθείας η ευρετηριοποίηση, το επόμενο βήμα θα μπορούσε να είναι αυτή η διαδικασία. Η εντολή που εκτελεί αυτή την εργασία είναι η `UPDATE INDEX`. Οι παράμετροι που δέχεται η εντολή είναι το όνομα του πίνακα του οποίου τα ευρετήρια θα ενημερωθούν, το όνομα του χειριστή που είναι προαιρετικό στην περίπτωση χρήσης κοινού ευρετηρίου στον πίνακα, υποχρεωτικό στην περίπτωση ξεχωριστών, και ο αριθμός των εγγραφών για οριστικοποίηση. Η εντολή αυτή εκτελεί την ευρετηριοποίηση κατευθείαν με το που θα δοθεί από το χρήστη, χωρίς αυτός να περιμένει το επόμενο χρονικό διάστημα που είναι προγραμματισμένο το ευρετήριο να ενημερωθεί.

Η ρύθμιση του ευρετηρίου όσον αφορά το διάστημα του κάθε πότε θα ενημερώνεται μπορεί να αλλάχθει μέσω της εντολής `CHANGE INDEX SETTINGS`.

Υπάρχουν περιπτώσεις κατά τις οποίες δεν είναι δυνατή η αναζήτηση σε ένα ευρετήριο. Η κατάσταση του ευρετηρίου μπορεί να ληφθεί με `GET INDEX STATUS`. Αν η εντολή αυτή εμφανίσει κάποιο λάθος, τότε η `RESET INDEX STATUS` επαναφέρει το ευρετήριο

⁷ Σκανδάλη (trigger) ονομάζεται μια διαδικασία που εκτελείται κάθε φορά που γίνεται ένα INSERT, DELETE ή UPDATE σε πίνακα της βάσης. Χρησιμοποιούνται συνήθως για έλεγχο δεδομένων, για διατήρηση της ακεραιότητας δεδομένων της βάσης, για υποστήριξη πολύπλοκων κανόνων, κτλ.

σε κατάσταση όπου μπορεί να χρησιμοποιηθεί και πάλι. Οι περιπτώσεις αυτές γράφονται ως γεγονότα στον πίνακα ημερολογίου του ευρετηρίου και μπορούν να βοηθήσουν στη διάγνωση προβλημάτων. Όταν δεν χρειάζονται πλέον οι εγγραφές των γεγονότων αυτών, μπορούν να διαγραφούν με την εντολή `DELETE INDEX EVENTS` που δέχεται παράμετρο το όνομα του πίνακα και το όνομα του χειριστή αν πρόκειται για μεμονωμένο ευρετήριο.

Πολλές φορές είναι επιθυμητό ο χρήστης να παίρνει πληροφορίες σχετικά την κατάσταση της βάσης, των ευρετηρίων, κτλ. Η εντολή `GET STATUS` επιστρέφει πληροφορίες σχετικά με την κατάσταση της βάσης, αν είναι ενεργοποιημένη ή όχι, των πινάκων, των στηλών και των εξωτερικών αρχείων. Στην περίπτωση που μια βάση είναι ενεργοποιημένη, τότε με την εντολή αυτή ο χρήστης παίρνει πληροφορίες σχετικά με τους πίνακες που είναι ενεργοποιημένοι, αν χρησιμοποιούν κοινά ή ξεχωριστά ευρετήρια και παράλληλα το όνομα του χειριστή κάθε ευρετηρίου.

Η `GET ENVIRONMENT` επιστρέφει τις τιμές των μεταβλητών περιβάλλοντος που έχουν σχέση με ποιο στιγμιότυπο βάσης είναι ενεργό, το όνομα της βάσης, καθώς και ο κατάλογος του στιγμιότυπου του Text Extender.

Η `GET TEXT CONFIGURATION` επιστρέφει τις τρέχουσες ρυθμίσεις κειμένου, αυτές που χρησιμοποιούνται ως οι εξ' ορισμού.

Η `GET INDEX STATUS` με παράμετρο το όνομα πίνακα στην περίπτωση του κοινού ευρετηρίου ή το όνομα του χειριστή στην περίπτωση των ξεχωριστών ευρετηρίων επιστρέφει την κατάσταση του ευρετηρίου, αν μπορεί να γίνει αναζήτηση σε αυτό, τον αριθμό των προγραμματισμένων εγγράφων προς ευρετηριοποίηση καθώς και τον αριθμό των εγγράφων που έχουν ήδη ευρετηριοποιηθεί και τον αριθμό πιθανών μηνυμάτων λάθους.

Οι ρυθμίσεις των ευρετηρίων φαίνονται με την εντολή `GET INDEX SETTINGS`, που επιστρέφει τις ρυθμίσεις του ευρετηρίου σχετικά με τον τύπο του, τη συχνότητα ενημέρωσής του, καθώς και τον κατάλογο στον οποίο αποθηκεύονται τα αρχεία του ευρετηρίου.

Οι ρυθμίσεις του κειμένου κάθε στήλης κειμένου πίνακα λαμβάνονται με την εντολή `GET TEXT INFO` με παράμετρο το όνομα του πίνακα για την περίπτωση κοινού ευρετηρίου ή το όνομα του χειριστή για την περίπτωση ξεχωριστών.

3.10 SQL ερωτήσεις

Όπως αναφέρθηκε παραπάνω, με την εγκατάστασή του, ο Text Extender εγκαθιστά συναρτήσεις ορισμένες από το χρήστη προκειμένου ο χρήστης να εκμεταλλευτεί τις δυνατότητες του Extender μέσω SQL. Οι συναρτήσεις αυτές εκτός των ορισμάτων που δέχονται και που είναι ειδικά για καθεμιά, δέχονται και την ερώτηση βάσει της οποίας γίνεται η αναζήτηση.

3.10.1 Σύνταξη ερώτησης

Το όρισμα της ερώτησης είναι η συνθήκη που καθορίζει ο χρήστης όταν κάνει αναζήτηση σε κείμενο. Αποτελείται από μια *συμβολοσειρά (string)* του τύπου `LONG VARCHAR` που περιέχει διάφορους όρους αναζήτησης και παραμέτρους.

Η σύνταξή της είναι :

```
>>+-----+----->
' -THESAURUS--"thesaurus-name"--+-----+-- '
                                     '-COUNT--"depth"-- '
>-----+----->
```

```
("passenger" & "vehicle") | ("transport" & "public")
```

Πρέπει να χρησιμοποιούνται παρενθέσεις στην περίπτωση που ο χρήστης θέλει να διατυπώσει κάτι διαφορετικό.

```
search-factor
|---++-----+---| search-primary |-----+-->
|   '-NOT-'      |
|               |
|               |               .-AND-----+---|
|               |               V               |
|   '-| s.-primary |---+-IN SAME PARAGRAPH AS-+---| s.-primary |---+-'
|               |               '-IN SAME SENTENCE AS-'
|
>-----|
```

```
search-primary
|---++| search-atom |-----+-----|
|   .-,'-----+---|
|   V               |
|   '-(-----| search-atom |---+---)---'
```

```
search-atom
|---+-----+-----+-----+-----+----->
|   +-PRECISE FORM OF-----+   '- language--'
|   +-STEMMED FORM OF-----+
|   +-FUZZY FORM OF--match-level--+
|   +-SYNONYM FORM OF-----+
|   +-BOUND-----+
|   +-SOUNDS LIKE-----+
|   +-feature-----+
|   '-| thesaurus |-----'
|
>----"word-or-phrase"-----+-----+-----|
|               '-ESCAPE--"escape-character"--'
```

```
thesaurus (αν καθοριστεί στην ερώτηση)
|---+-----+---TERM OF-----|
|   '-EXPAND--"relation"--'
```

Χαρακτηριστικό (feature)

Το χαρακτηριστικό είναι μια επιλογή που καθορίζει την επαύξηση του όρου αναζήτησης. Αυτή η επιλογή χρησιμοποιείται μόνο όταν χρησιμοποιείται γλωσσολογικό ευρετήριο που χρησιμοποιεί εξαγωγή χαρακτηριστικών. Χρησιμοποιείται μόνο για κείμενα στην αγγλική. Οι επαυξήσεις που μπορούν να γίνουν αναφορικά είναι : Για πρόσωπα (PERSON), οργανισμούς (ORGANIZATION), ονόματα τοποθεσιών (PLACE) και domain όρους (DOMAIN). Αν καθοριστεί το ANY, τότε περιλαμβάνει όλες τις δυνατές επαυξήσεις.

THESAURUS thesaurus-name

Η λέξη κλειδί thesaurus thesaurus-name καθορίζει το όνομα του θησαυρού που θα χρησιμοποιηθεί για τον εμπλουτισμό στη σύνταξη της ερώτησης. Το όνομα του θησαυρού είναι το όνομα του αρχείου που έχει δημιουργηθεί από το μεταγλωττιστή θησαυρού. Υπάρχει ένα εξ' ορισμού αρχείο θησαυρού με όνομα desthes. Το όνομα του θησαυρού μπορεί να είναι το πλήρες όνομα του αρχείου.

COUNT αριθμός

Αυτή η λέξη κλειδί καθορίζει τον αριθμό επιπέδων των όρων που θα χρησιμοποιηθούν για την επαύξηση του όρου αναζήτησης για την δοθέντα σχέση. Αν δεν καθοριστεί αυτή η παράμετρος, ο αριθμός επιπέδων είναι εξ' ορισμού ένα.

RESULT LIMIT αριθμός

Η παράμετρος αυτή καθορίζει το μέγιστο αριθμό των αποτελεσμάτων ανά ερώτηση. Ο αριθμός μπορεί να κυμαίνεται μεταξύ 1 και 32.767.

EXPAND σχέση

Αυτή η λέξη κλειδί καθορίζει τη σχέση μεταξύ του όρου που καθορίζεται με τη λέξη κλειδί **TERM OF** και των όρων από το θησαυρό που χρησιμοποιείται για την επαύξηση του όρου αναζήτησης. Η σχέση πρέπει να αντιστοιχεί σε μια από τις σχέσεις που χρησιμοποιούνται στο θησαυρό.

TERM OF «λέξη ή φράση»

Ο όρος αναζήτησης ή φράση στον οποίο άλλοι όροι αναζήτησης θα προστεθούν από το θησαυρό.

NOT search primary

Με αυτή την παράμετρο ο χρήστης μπορεί να αποκλείσει από την αναζήτηση έγγραφα που περιέχουν συγκεκριμένο όρο. Όταν χρησιμοποιείται ο τελεστής **NOT**, δεν μπορεί να χρησιμοποιηθεί η λέξη κλειδί **SYNONYM FORM OF**.

search primary IN SAME PARAGRAPH AS search primary

Αυτή η λέξη κλειδί επιτρέπει την αναζήτηση για συνδυασμούς όρων που υπάρχουν στην ίδια παράγραφο. Η λέξη κλειδί δεν μπορεί να χρησιμοποιηθεί όταν υπάρχει ο τελεστής **NOT** στον παράγοντα αναζήτησης.

search primary IN SAME SENTENCE AS search primary

Λέξη κλειδί που έχει παρόμοια λειτουργία με την προηγούμενη, μόνο που ισχύει για πρόταση.

AND search primary

Το **AND** μπορεί να συνδυάσει πολλά **search-primaries** για αναζήτηση στην ίδια πρόταση ή παράγραφο. Για παράδειγμα, η επόμενη πρόταση ψάχνει για τα 'forest', 'rain' και 'land' στην ίδια παράγραφο :

```
"forest" IN SAME SENTENCE AS "rain" AND "erosion" AND "land"
```

search atom

Κάθε **search atom** περιέχει τουλάχιστον μια λέξη ή φράση. Αν συνδυαστούν πολλά από αυτά χωρισμένα με κόμμα, τότε μια αναζήτηση θεωρείται επιτυχής αν ένας τουλάχιστον από τους όρους βρεθεί.

PRECISE FORM OF, STEMMED FORM OF, FUZZY FORM OF, SYNONYM FORM OF, BOUND

Ο πίνακας 3.4 δείχνει τις επιλογές που αντιστοιχούν στα διάφορα είδη ευρετηρίων.

Λέξη κλειδί search atom	Είδος ευρετηρίου					
	Γλωσσολογικό		Ακριβές	Διπλό	Ngram	
	--	FEATURE_EXTRACTION			--	CASE_ENABLED
PRECISE FORM OF			X	O		O
STEMMED FORM OF	X	X		X	O	O
FUZZY FORM OF					O	O
IS ABOUT	O	O	O	O		
SYNONYM FORM OF	O	O	O	O		
EXPAND	O	O	O	O		
Χαρακτηριστικό		O		O		
SOUNDS LIKE	O	O	O	O		
IN SAME SENTENCE AS	O	O	O	O	O	O
IN SAME PARAGRAPH AS	O	O	O	O	O	O
BOUND					O	O
Με X συμβολίζονται οι εξ' ορισμού επιλογές, με O αυτές που είναι διαθέσιμες						

Πίνακας 3.4: Διαθέσιμες επιλογές για τα είδη ευρετηρίων

Ο πίνακας 3.5 δείχνει τη διαδικασία επεξεργασίας των όρων αναζήτησης.

Λέξη κλειδί search atom	Κεφαλαία-πεζά				
	Κεφαλαία-πεζά		Αποκοπή κοινότων καταλήξεων	Ταύτιση	
	Διάκριση	Μη διάκριση		Ακριβής	Fuzzy
PRECISE FORM OF	Μόνο αν είναι ενεργοποιημένη η επιλογή CASE_ENABLED	X		X	
STEMMED FORM OF		X	X		
FUZZY FORM OF		X			X
Με X συμβολίζονται οι εξ' ορισμού επιλογές					

Πίνακας 3.5: Επεξεργασία όρων αναζήτησης

PRECISE FORM OF

Η λέξη που ακολουθεί τη λέξη κλειδί PRECISE FORM OF αναζητείται ακριβώς όπως δίνεται, αντί πρώτα να αναχθεί στη βασική της μορφή. Αυτή η μορφή αναζήτησης είναι βασισμένη σε κεφαλαία-πεζά.

Αυτή είναι και η εξ' ορισμού επιλογή για το ακριβές και διπλό ευρετήριο. Αν η λέξη κλειδί δοθεί για γλωσσολογικό ευρετήριο αγνοείται και χρησιμοποιείται η επιλογή STEMMED FORM OF.

STEMMED FORM OF

Με αυτή τη λέξη κλειδί, γίνεται αναγωγή στη βασική μορφή της λέξης που ακολουθεί τη δήλωση **STEMMED FORM OF** πριν γίνει η αναζήτηση. Αυτός ο τρόπος αναζήτησης δεν βασίζεται σε κεφαλαία-πεζά. Η αναγωγή στη βασική μορφή εξαρτάται από τη γλώσσα του κειμένου.

Η επιλογή αυτή είναι και η εξ' ορισμού για τα γλωσσολογικά ευρετήρια. Αν η επιλογή αυτή καθοριστεί για ακριβές ευρετήριο αγνοείται και χρησιμοποιείται η επιλογή **PRECISE FORM OF**.

FUZZY FORM OF

Η επιλογή αυτή δίνεται για fuzzy αναζήτηση, που είναι αναζήτηση για όρους που έχουν παρόμοια ορθογραφία με εκείνη του όρου αναζήτησης. Αυτό είναι χρήσιμο για έγγραφα που έχουν δημιουργηθεί από πρόγραμμα *οπτικής αναγνώρισης χαρακτήρων* (*Optical Character Recognition*). Τέτοια έγγραφα περιέχουν συνήθως αρκετές λάθος ορθογραφικά λέξεις.

Μαζί με αυτή τη λέξη κλειδί, δίνεται και μια αριθμητική παράμετρος (από 1 έως 5) που δείχνει το βαθμό ομοιότητας (το 5 είναι το αυστηρότερο).

SYNONYM FORM OF

Αυτή η λέξη κλειδί προκαλεί την αναζήτηση εκτός του όρου αναζήτησης και των συνωνύμων της που δίνονται από το λεξικό της γλώσσας.

Τα συνώνυμα φράσεων είναι εναλλακτικές φράσεις που περιέχουν όλους τους πιθανούς συνδυασμούς συνωνύμων που μπορούν να προκύψουν αν αντικατασταθεί κάθε λέξη της αρχικής φράσης με ένα συνώνυμό της. Η σειρά των λέξεων παραμένει ως έχει από την αρχική φράση.

Αν αυτή η λέξη κλειδί δοθεί για ακριβές ευρετήριο αγνοείται και χρησιμοποιείται η **PRECISE FORM OF**. Για διπλό ευρετήριο, η αναζήτηση γίνεται βασισμένη στο γλωσσολογικό μέρος του διπλού ευρετηρίου, αντίθετα με το ακριβές μέρος. Η **SYNONYM FORM OF** δεν μπορεί να χρησιμοποιηθεί όταν χρησιμοποιείται ο τελεστής **NOT** στον παράγοντα αναζήτησης ή όταν η λέξη ή φράση προς αναζήτηση περιέχει χαρακτήρες μάσκες.

BOUND

Λέξη κλειδί που χρησιμοποιείται για κείμενα με CCSID που αντιστοιχεί στην κορεατική γλώσσα. Αν χρησιμοποιηθεί, η αναζήτηση λαμβάνει υπόψη τα περιθώρια των φράσεων. Αν χρησιμοποιείται η παράμετρος της γλώσσας, η λέξη κλειδί αγνοείται, διαφορετικά γίνεται υπόθεση ότι πρόκειται για κορεατικό κείμενο.

SOUNDS LIKE

Χρησιμοποιείται για την ανάκληση παρόμοιων «ηχητικά» λέξεων, ιδιότητα χρήσιμη για έγγραφα που περιέχουν λέξεις που «ακούγονται» το ίδιο, αλλά έχουν διαφορετική ορθογραφία.

Γλώσσα (language)

Μεταβλητή που δηλώνει πιο λεξικό θα χρησιμοποιηθεί για τη γλωσσολογική επεξεργασία των εγγράφων κατά τη διάρκεια ευρετηριοποίησης και ανάκτησης. Αυτό ισχύει όχι μόνο για γλωσσολογικό και διπλό, αλλά και για ακριβές ευρετήριο επειδή χρησιμοποιούν λεξικό για να επεξεργαστούν τις στοπ λέξεις.

Η γλωσσολογική επεξεργασία περιλαμβάνει επίσης και επεξεργασία συνωνύμων και αποκοπή των κοινότητων καταλήξεων των λέξεων.

Η παράμετρος αυτή πρέπει να δίνεται για κείμενα που δεν είναι στην αγγλική Αμερικής, ανεξάρτητα με το ποια γλώσσα είναι η εξ' ορισμού.

«Λέξη ή φράση» (word-or-phrase)

Αυτή η παράμετρος είναι η λέξη ή φράση προς αναζήτηση. Οι χαρακτήρες που μπορούν να χρησιμοποιηθούν είναι ανεξάρτητοι γλώσσας.

Για ακριβές ή γλωσσολογικό ευρετήριο, ο Extender μπορεί να αναζητήσει χρησιμοποιώντας είτε την ακριβή μορφή της λέξης ή φράσης είτε μια παραλλαγή της. Αν δεν καθοριστεί καμία παράμετρος από τον πίνακα 3.4, χρησιμοποιούνται οι εξ' ορισμού επιλογές ανάλογα με το είδος του ευρετηρίου.

Όσον αφορά στους χαρακτήρες μάσκες, μια λέξη μπορεί να χρησιμοποιεί τους εξής χαρακτήρες :

- `_` (underscore) : Αντιπροσωπεύει ένα χαρακτήρα.
- `%` (percent) : Αντιπροσωπεύει ένα οποιοδήποτε αριθμό οποιονδήποτε χαρακτήρων. Αν μια λέξη περιέχει μόνο ένα `%` τότε αντιπροσωπεύει μια λέξη οποιουδήποτε μήκους.

Μια λέξη δεν μπορεί να αποτελείται αποκλειστικά από χαρακτήρες μάσκες, εκτός από την περίπτωση του απλού `%`. Επίσης, αν χρησιμοποιούνται χαρακτήρες μάσκες, δεν μπορεί να χρησιμοποιηθεί καμία από τις λέξεις κλειδιά `SYNONYM FORM OF` και `THESAURUS`, καθώς και το χαρακτηριστικό.

ESCAPE «χαρακτήρας διαφυγής»

Η παράμετρος αυτή ορίζει τον χαρακτήρα διαφυγής. Στη λέξη ή φράση προς αναζήτηση, ο χαρακτήρας που ακολουθεί τον χαρακτήρα διαφυγής αναγνωρίζεται ως απλός χαρακτήρας κι όχι ως χαρακτήρας μάσκα. Για παράδειγμα, αν οριστεί ως χαρακτήρας διαφυγής το `$`, τότε στη φράση το `$%` αντιμετωπίζεται ως απλός χαρακτήρας `'%`.

3.10.2 Περίληψη κανόνων και περιορισμών

Παρακάτω συνοψίζονται οι κανόνες και οι περιορισμοί όσον αφορά στη σύνταξη της ερώτησης.

- Στους λογικούς τελεστές, δεν επιτρέπεται ο τελεστής `NOT` να ακολουθεί το `OR`.
- Για το διπλό ευρετήριο, η εξ' ορισμού επιλογή είναι η `STEMMED FORM OF`. Αν χρησιμοποιούνται χαρακτήρες μάσκες, οι αναζητήσεις δεν βασίζονται σε κεφαλαία-πεζά.
- Η επιλογή `IN SAME PARAGRAPH AS` δεν μπορεί να χρησιμοποιηθεί αν υπάρχει ο τελεστής `NOT` στον παράγοντα αναζήτησης, όπως και η `IN SAME SENTENCE AS` και η `SYNONYM FORM OF`.
- Το γλωσσολογικό ευρετήριο αποτρέπει τη χρησιμοποίηση της επιλογής `PRECISE FORM OF`. Ως εξ' ορισμού δέχεται την `STEMMED FORM OF`. Χαρακτήρες μάσκες είναι επιτρεπτοί και οι αναζητήσεις δεν βασίζονται σε κεφαλαία-πεζά.
- Όταν χρησιμοποιούνται χαρακτήρες μάσκες, δεν μπορούν να χρησιμοποιηθούν οι επιλογές `SYNONYM FORM OF`, `THESAURUS`, καθώς και το χαρακτηριστικό.

- Για ngram ευρετήριο, οι χαρακτήρες μάσκες είναι επιτρεπτοί όταν αντικαθιστούν μόνο αλφαριθμητικούς χαρακτήρες. Οι αναζητήσεις δεν βασίζονται σε κεφαλαία-πεζά, εκτός αν το ευρετήριο έχει οριστεί με την επιλογή CASE_ENABLED και χρησιμοποιείται η επιλογή PRECISE FORM OF.
- Το ακριβές ευρετήριο αποτρέπει την χρησιμοποίηση των επιλογών STEMMED FORM OF και SYNONYM FORM OF. Ως εξ' ορισμού δέχεται την PRECISE FORM OF. Χαρακτήρες μάσκες επιτρέπονται και οι αναζητήσεις βασίζονται σε κεφαλαία-πεζά.

3.10.3 Συναρτήσεις του Text Extender

Οι συναρτήσεις που παρέχει ο Extender είναι οι εξής :

CONTAINS

Η συνάρτηση αυτή κάνει αναζήτηση κειμένου. Καλείται με δύο ορίσματα και επιστρέφει την τιμή τύπου INTEGER 1 αν η αναζήτηση είναι επιτυχής, διαφορετικά 0. Η πρώτη παράμετρος είναι το όνομα του χειριστή της στήλης κειμένου στην οποία γίνεται αναζήτηση και η δεύτερη είναι η ερώτηση, που είναι του τύπου LONG VARCHAR.

NO_OF_DOCUMENTS

Η NO_OF_DOCUMENTS επιστρέφει ένα αριθμό τύπου INTEGER που δείχνει τον αριθμό των εγγράφων που επέστρεψε μια αναζήτηση. Ο αριθμός αυτός είναι ο αριθμός των εγγραφών μιας λίστας χειριστών. Η συνάρτηση δέχεται μια παράμετρο, που είναι του τύπου DB2TEXTHLISTP ή DB2TEXTFHLISTP. Η παράμετρος αυτή λαμβάνεται από τη συνάρτηση HANDLE_LIST και ουσιαστικά είναι ένας δείκτης (*pointer*) σε μια λίστα χειριστών των εγγράφων από το αποτέλεσμα μιας αναζήτησης. Η HANDLE_LIST πρέπει να βρίσκεται εντός της κλήσης της NO_OF_DOCUMENTS, διότι η λίστα είναι στην εμβέλεια (*scope*) της SQL δήλωσης.

HANDLE_LIST

Η συνάρτηση HANDLE_LIST κάνει αναζήτηση και επιστρέφει μια τιμή του τύπου DB2TEXTHLISTP ή DB2TEXTFHLISTP που δείχνει σε μια λίστα χειριστών για τα έγγραφα που βρέθηκαν από την αναζήτηση. Ο δείκτης αυτός υπάρχει μόνο στην εμβέλεια της δήλωσης SQL που καλεί τη συνάρτηση αυτή. Αν φυσικά δεν βρεθούν έγγραφα που να ικανοποιούν την ερώτηση, το ευρετήριο είναι κενό. Η HANDLE_LIST μπορεί να χρησιμοποιηθεί σε συνδυασμό με τη NO_OF_DOCUMENTS για να βρεθεί ο αριθμός των εγγράφων που βρέθηκαν.

Η συνάρτηση δέχεται δύο παραμέτρους, εκ των οποίων η πρώτη είναι το όνομα του χειριστή της στήλης κειμένου και η δεύτερη είναι η ερώτηση.

INIT_TEXT_HANDLE

Η INIT_TEXT_HANDLE χρησιμοποιείται για την επιστροφή ενός μερικώς αρχικοποιημένου (*initialized*) χειριστή που περιέχει εξ' ορισμού τιμές για τη μορφοποίηση του κειμένου και τη γλώσσα που μπορεί να εισαχθεί σε στήλη χειριστή στον πίνακα. Αυτό είναι χρήσιμο για την περίπτωση που γίνεται εισαγωγή εγγραφής στον πίνακα, της οποίας η μορφοποίηση και η γλώσσα του κειμένου είναι διαφορετική από αυτές που έχει η παραμετροποίηση του κειμένου. Η τιμή επιστροφής της συνάρτησης είναι του τύπου

DB2TEXTH. Στην περίπτωση αναζήτησης σε κείμενο που αποθηκεύεται εκτός βάσης, η συνάρτηση χρησιμοποιείται για να επιστρέψει ένα τελείως αρχικοποιημένο χειριστή που περιέχει τιμές για το CCSID, τη μορφοποίηση, τη γλώσσα και την τοποθεσία του αρχείου. Σε αυτή την περίπτωση η τιμή που επιστρέφει η συνάρτηση είναι του τύπου DB2TEXTFH.

NO_OF_MATCHES

Η συνάρτηση NO_OF_MATCHES επιστρέφει ένα αριθμό τύπου INTEGER που δείχνει τον αριθμό των επιτυχιών ανά έγγραφο. Δέχεται δύο παραμέτρους, από τις οποίες η πρώτη είναι το όνομα του χειριστή της στήλης κειμένου και η δεύτερη είναι η ερώτηση αναζήτησης.

RANK

Αυτή η συνάρτηση μπορεί να κάνει αναζήτηση και να επιστρέψει για κάθε έγγραφο που θα βρεθεί μια τιμή διαβάθμισης που δείχνει πόσο καλά περιγράφεται από την ερώτηση αναζήτησης το έγγραφο αυτό. Η τιμή που επιστρέφει είναι του τύπου DOUBLE και είναι μεταξύ του 0 και 1. Η τιμή διαβάθμισης είναι απόλυτη και δείχνει το βαθμό που το έγγραφο ικανοποιεί την ερώτηση σε σχέση με τα άλλα έγγραφα που βρέθηκαν.

Η συνάρτηση δέχεται δύο παραμέτρους, από τις οποίες η πρώτη είναι το όνομα του χειριστή της στήλης κειμένου και η δεύτερη είναι η ερώτηση.

REFINE

Η συνάρτηση REFINE δέχεται δύο ορίσματα, τα οποία είναι δύο διαφορετικές ερωτήσεις αναζήτησης, και επιστρέφει μία συνδυασμένη ερώτηση του τύπου LONG VARCHAR, αποτελούμενη από τις δύο αρχικές ερωτήσεις συνδυασμένες με το λογικό τελεστή AND.

SEARCH_RESULT

Η SEARCH_RESULT επιστρέφει τα αποτελέσματα μιας ερώτησης σε ένα ενδιάμεσο πίνακα. Η συνάρτηση αυτή χρησιμοποιείται στο τμήμα FROM της SQL δήλωσης. Ο ενδιάμεσος πίνακας έχει τη δομή :

Όνομα στήλης	Τύπος
HANDLE	DB2TEXTH ή DB2TEXTFH
NUMBER_OF_MATCHES	INTEGER
RANK	DOUBLE

Η συνάρτηση είναι πιο γρήγορη από τις CONTAIN και RANK στην επεξεργασία μεγάλων πινάκων. Τα ορίσματα που δέχεται είναι δύο, από τα οποία το πρώτο είναι το όνομα του χειριστή της στήλης κειμένου και το δεύτερο είναι η ερώτηση αναζήτησης.

ΚΕΦΑΛΑΙΟ 4

ORACLE CONTEXT CARTRIDGE

4.1 Εισαγωγή στο Oracle ConText cartridge

Το ConText cartridge¹ είναι επιλογή του Σ.Δ.Β.Δ. Oracle που ενεργοποιεί ερωτήσεις αναζήτησης κειμένου μέσω SQL και PL/SQL².

Οι δυνατότητες του ConText cartridge υλοποιούνται μέσω διεργασιών (*processes*) στον διακομιστή της βάσης και ευρετηρίων σε κάθε στήλη πίνακα που περιέχει το κείμενο. Από προγραμματιστικής πλευράς, μπορεί να χρησιμοποιηθεί SQL και PL/SQL για να εμπλουτιστούν οι εφαρμογές με ConText δυνατότητες.

Εργαλεία που μπορούν να καλέσουν αποθηκευμένες διαδικασίες της βάσης μπορούν να εκτελέσουν και ερωτήσεις κειμένου.

Το ConText χειρίζεται τα δεδομένα κειμένου με παρόμοιο τρόπο που χειρίζεται η βάση τους κλασικούς τύπους δεδομένων. Όταν εισάγεται, ενημερώνεται ή διαγράφεται κείμενο, το ConText αναλαμβάνει τη διαχείριση.

4.1.1 Χαρακτηριστικά του ConText cartridge

Το ConText παρέχει εξελιγμένες διαδικασίες ευρετηριοποίησης, ανάλυσης, ανάκτησης και παρουσίασης που μπορούν να ενσωματωθούν σε οποιαδήποτε εφαρμογή χρησιμοποιεί Oracle8. Μια συνοπτική αναφορά των χαρακτηριστικών του είναι :

- Πλήρης ολοκλήρωση δομημένης πληροφορίας και κειμένου
- Εσωτερική και εξωτερική (εκτός βάσης) αποθήκευση του κειμένου
- Υποστήριξη για μια ποικιλία μορφοποιήσεων εγγράφων, περιλαμβανομένων απλού κειμένου, κειμένου HTML και μερικών από τις πιο δημοφιλείς μορφοποιήσεις
- Υποστήριξη για ευρετηριοποίηση και εκτέλεση ερωτήσεων σε στήλες πινάκων κειμένου όλων των τύπων, περιλαμβανομένων των τύπων LONG, LONG RAW και LOBs (*large objects*)
- Ευρετηριοποίηση και εκτέλεση ερωτήσεων σχεσιακών πινάκων/όψεων (*views*) και πινάκων/όψεων αντικειμένων
- Ευρετηριοποίηση και ανάκτηση για όλες τις υποστηριζόμενες NLS γλώσσες, περιλαμβανομένων και των εξής γλωσσών που χρησιμοποιούν πολλά bytes ανά χαρακτήρα : Ιαπωνικά, Κινέζικα και Κορεάτικα
- Εξελιγμένες δυνατότητες ανάκτησης για κείμενο στην αγγλική, γαλλική, γερμανική, ιταλική, ισπανική και ολλανδική γλώσσα
- Εξελιγμένες γλωσσολογικές δυνατότητες, περιλαμβανομένων ερωτήσεων βασισμένες σε θέματα, για κείμενο στην αγγλική γλώσσα
- Τονισμός κειμένου

4.2 Ο διακομιστής ConText

4.2.1 Οι διακομιστές ConText και Oracle

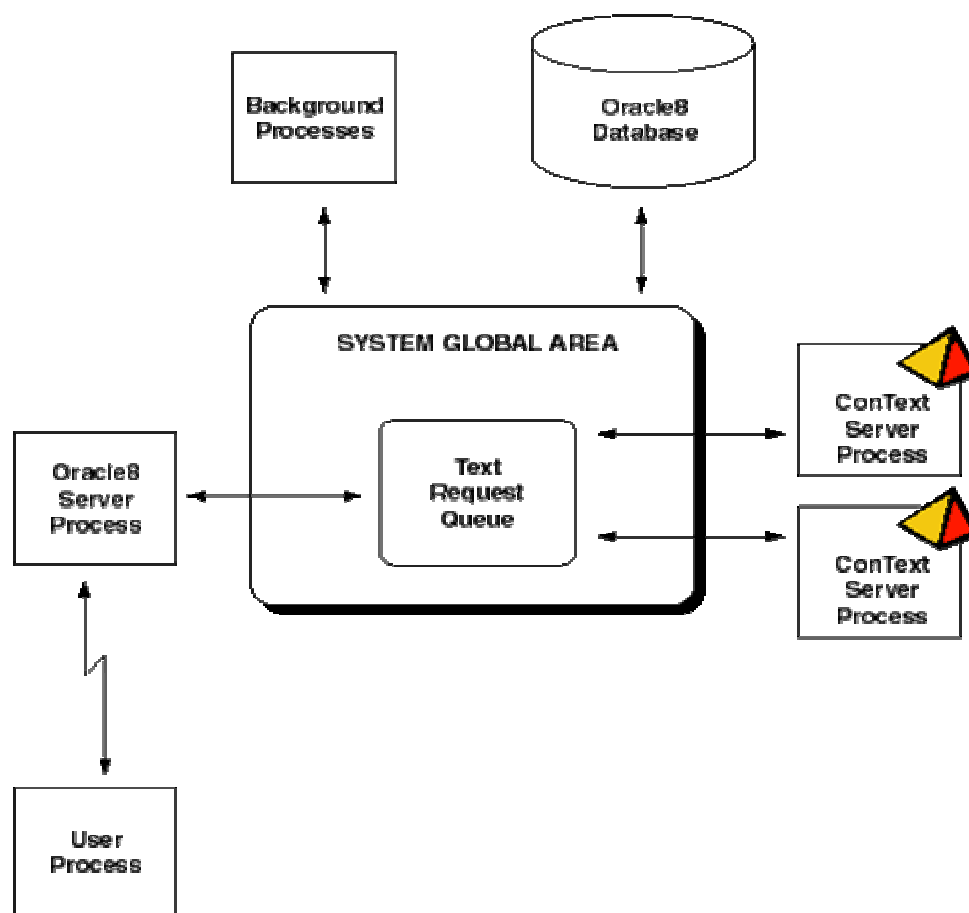
Ο διακομιστής ConText ακολουθεί τη λογική του μοντέλου του διακομιστή Oracle. Παράλληλα με τις διεργασίες του ConText υπάρχει και μια *ουρά ερωτήσεων κειμένου*

¹ Τα cartridges της Oracle μπορούν να θεωρηθούν ως επεκτάσεις (modules) που εγκαθίστανται και προσθέτουν δυνατότητες στη βάση.

² PL/SQL (Programming Language/SQL) είναι η γλώσσα που υποστηρίζει η Oracle για τη δημιουργία αποθηκευμένων διαδικασιών (*stored procedures*) στη βάση. Είναι γλώσσα διαδικασιακή (*procedural*), δομείται σε μπλοκ και παράλληλα ενσωματώνει με διάφανο τρόπο δηλώσεις SQL. Έχει μεγάλη ολοκλήρωση με όλο το φάσμα της Oracle.

(text request queue). Το σχήμα 4.1 δείχνει τους διακομιστές ConText και την ουρά των ερωτήσεων κειμένου σε συνεργασία με τον διακομιστή Oracle σε ένα περιβάλλον που υπάρχει ένας διακομιστής Oracle ανά χρήστη.

Στο κλασικό μοντέλο διακομιστή Oracle, όταν ένας χρήστης ανοίγει σύνδεση με τη βάση, η βάση δημιουργεί μια διεργασία διακομιστή αποκλειστικά αφιερωμένη να εξυπηρετεί τις αιτήσεις του χρήστη. Με το ConText, όταν μια αίτηση του χρήστη έχει ερώτηση κειμένου, η διεργασία του διακομιστή λαμβάνει την αίτηση του χρήστη, αναγνωρίζει την ερώτηση και την προωθεί στην ουρά των ερωτήσεων κειμένου για να εξυπηρετηθεί από μια διεργασία διακομιστή ConText. Αφού ο διακομιστής επεξεργαστεί την ερώτηση, το αποτέλεσμα επιστρέφεται στο χρήστη.



Σχήμα 4.1: Διακομιστές ConText και Oracle

ρετεί τις αιτήσεις του χρήστη. Με το ConText, όταν μια αίτηση του χρήστη έχει ερώτηση κειμένου, η διεργασία του διακομιστή λαμβάνει την αίτηση του χρήστη, αναγνωρίζει την ερώτηση και την προωθεί στην ουρά των ερωτήσεων κειμένου για να εξυπηρετηθεί από μια διεργασία διακομιστή ConText. Αφού ο διακομιστής επεξεργαστεί την ερώτηση, το αποτέλεσμα επιστρέφεται στο χρήστη.

4.2.2 Ρόλοι του διακομιστή ConText

Ένα ιδιαίτερο χαρακτηριστικό του ConText διακομιστή είναι ότι μπορεί να συμπεριφέρεται διαφορετικά. Αυτή η διαφορετική συμπεριφορά ονομάζεται *ρόλος (role)*. Ο ρόλος του διακομιστή καθορίζεται όταν ξεκινά και μπορεί να έχει παραπάνω από έναν. Υπάρχουν πέντε τέτοιοι ρόλοι που καθορίζουν το είδος επεξεργασίας του κειμένου που θα γίνει από τον διακομιστή. Οι ρόλοι αυτοί είναι οι εξής :

- Ρόλος φορτωτή (R)
- Ρόλος DDL (D)
- Ρόλος DML (M)
- Ρόλος ερωτήσεων (Q)

- Ρόλος γλωσσολογικής επεξεργασίας (L)

Παράλληλα με τους παραπάνω ρόλους, κάθε διακομιστής ConText έχει ταυτόχρονα και το ρόλο DBA, που του παρέχει δυνατότητες ανακάλυψης λαθών στον διακομιστή και πελάτη και δυνατότητες ανάκαμψης. Για κάθε διακομιστή δίνεται και ο ρόλος αυτός, γεγονός που αποτρέπει την αποτυχία ενός σε ένα περιβάλλον με πολλούς διακομιστές.

4.2.2.1 Ρόλος φορτωτή (R)

Ο ρόλος αυτός επιτρέπει στον διακομιστή ConText να ελέγχει ανά τακτά χρονικά διαστήματα συγκεκριμένους καταλόγους ενός συστήματος αρχείων για την ύπαρξη νέων αρχείων κειμένου που πρέπει να φορτωθούν σε στήλες πινάκων στη βάση. Όταν ο διακομιστής ανακαλύψει την ύπαρξη ενός νέου αρχείου, χρησιμοποιεί το εργαλείο `ctxload` για να φορτώσει τα περιεχόμενα του αρχείου στη βάση.

4.2.2.2 DDL ρόλος (D)

Ο ρόλος DDL επιτρέπει στον διακομιστή ConText να δημιουργεί, καταργεί και βελτιστοποιεί *ευρετήρια κειμένου (text indexes)*. Το ευρετήριο κειμένου είναι αυτό που επιτρέπει στους χρήστες να δίνουν ερωτήσεις για κείμενο αποθηκευμένο σε μια στήλη πίνακα. Επίσης, ο διακομιστής μπορεί να εξυπηρετήσει και ερωτήσεις που απαιτούν το ρόλο DML στην περίπτωση που οι αιτήσεις αυτές εξυπηρετούνται με μαζικό ρυθμό.

4.2.2.3 DML ρόλος (M)

Έχοντας το ρόλο αυτό, ο διακομιστής ConText μπορεί να ενημερώνει αυτόματα τα ευρετήρια κειμένου ενός πίνακα όταν υπάρχουν αλλαγές στον πίνακα αυτό που απαιτούν την ενημέρωση του ευρετηρίου. Τέτοιες αλλαγές περιλαμβάνουν την εισαγωγή νέων εγγράφων, την αλλαγή ή και τη διαγραφή υπαρχόντων.

4.2.2.4 Ρόλος ερωτήσεων (Q)

Ο ρόλος ερωτήσεων επιτρέπει στον διακομιστή την επεξεργασία ερωτήσεων για κείμενο αποθηκευμένο είτε εντός της βάσης είτε εκτός. Αν δεν υπάρχει τουλάχιστον ένας διακομιστής ConText με αυτό το ρόλο, τότε οποιαδήποτε ερώτηση θα αποτύχει.

4.2.2.5 Ρόλος γλωσσολογικής επεξεργασίας (L)

Αυτός ο ρόλος επιτρέπει στον διακομιστή να επεξεργαστεί ερωτήσεις γλωσσολογικής φύσεως και να δημιουργήσει την κατάλληλη έξοδο. Η έξοδος περιλαμβάνει τα θέματα και τις έννοιες και περιλήψεις θεμάτων. Ο ρόλος αυτός απαιτείται μόνο όταν γίνεται επεξεργασία ερωτήσεων του τύπου που αναφέρεται παραπάνω.

4.2.2.6 DBA ρόλος

Όπως αναφέρθηκε και παραπάνω, ο DBA ρόλος είναι ένας ρόλος που αποδίδεται αυτόματα σε κάθε διακομιστή ConText. Εξυπηρετεί διαχειριστικούς σκοπούς, επιτρέπει στον διακομιστή να αναλάβει σε περίπτωση που αποτύχει κάποιος άλλος και να «καθαρίξει» το σύστημα ανά τακτά χρονικά διαστήματα.

Στην πρώτη περίπτωση, κάνει τα εξής :

- Ο εν λειτουργία διακομιστής καθαρίζει την περιοχή μνήμης του διακομιστή που απέτυχε και περνά σε ημερολόγιο το γεγονός

- Αν ο αποτυχημένος διακομιστής επεξεργαζόταν μια ερώτηση, ο εν λειτουργία διακομιστής απελευθερώνει πόρους (*resources*) που είχε δεσμεύσει
- Αν ο αποτυχημένος διακομιστής επεξεργαζόταν αίτηση DDL, ο εν λειτουργία διακομιστής καταργεί όσους πίνακες δεν χρειάζονται για την ανάκαμψη από τη συγκεκριμένη αίτηση
- Αν ο αποτυχημένος διακομιστής επεξεργαζόταν αίτηση DML, ο εν λειτουργία διακομιστής τοποθετεί οποιεσδήποτε μη ολοκληρωμένες αιτήσεις DML πίσω στην ουρά
- Αν ο αποτυχημένος διακομιστής επεξεργαζόταν αίτηση DDL ή ερώτηση, ο εν λειτουργία αναλαμβάνει να ειδοποιήσει τον εν αναμονή πελάτη

Στη δεύτερη περίπτωση, όταν ένας πίνακας που έχει στήλη κειμένου με ένα υπάρχον ευρετήριο ConText καταργείται χωρίς να καταργηθεί πρώτα το ευρετήριο και η πολιτική της στήλης, τόσο το ευρετήριο όσο και η πολιτική της στήλης παραμένουν στο σύστημα ωσότου γίνει η ανάκαμψη. Αυτή η διαδικασία γίνεται αυτόματα από ένα διακομιστή ConText κάθε 15 λεπτά. Μπορεί φυσικά να γίνει και με ρητό τρόπο, δίνοντας προς εκτέλεση τη δήλωση `CTX_ADM.RECOVER`.

4.3 Διαδικασία ευρετηριοποίησης

Το σχήμα 4.2 δείχνει τη συσχέτιση μεταξύ πινάκων κειμένου, πολιτικών, ευρετηρίων ConText και ερωτήσεων.

Σε ένα τυπικό σύστημα ConText, το κείμενο φορτώνεται σε μια στήλη κειμένου σε πίνακα στη βάση και δημιουργείται μια πολιτική (βλέπε §4.3.1) για τη στήλη αυτή.

Η πολιτική χρησιμοποιείται για να δημιουργήσει το ευρετήριο ConText, που ανήκει σε ξεχωριστούς πίνακες συσχετισμένους με τη στήλη κειμένου μέσω της πολιτικής. Από τη στιγμή που θα δημιουργηθεί το ευρετήριο για μια στήλη, μπορούν να δοθούν ερωτήσεις για το κείμενο της στήλης αυτής.

Όταν δίνεται μια ερώτηση για κείμενο σε στήλη η οποία έχει ευρετήριο ConText, αντί να σαρωθεί το κείμενο του πίνακα για να βρεθούν έγγραφα που ικανοποιούν την ερώτηση, γίνεται αναζήτηση στους πίνακες ευρετηρίων ConText για να βρεθεί αν ένα έγγραφο την ικανοποιεί.

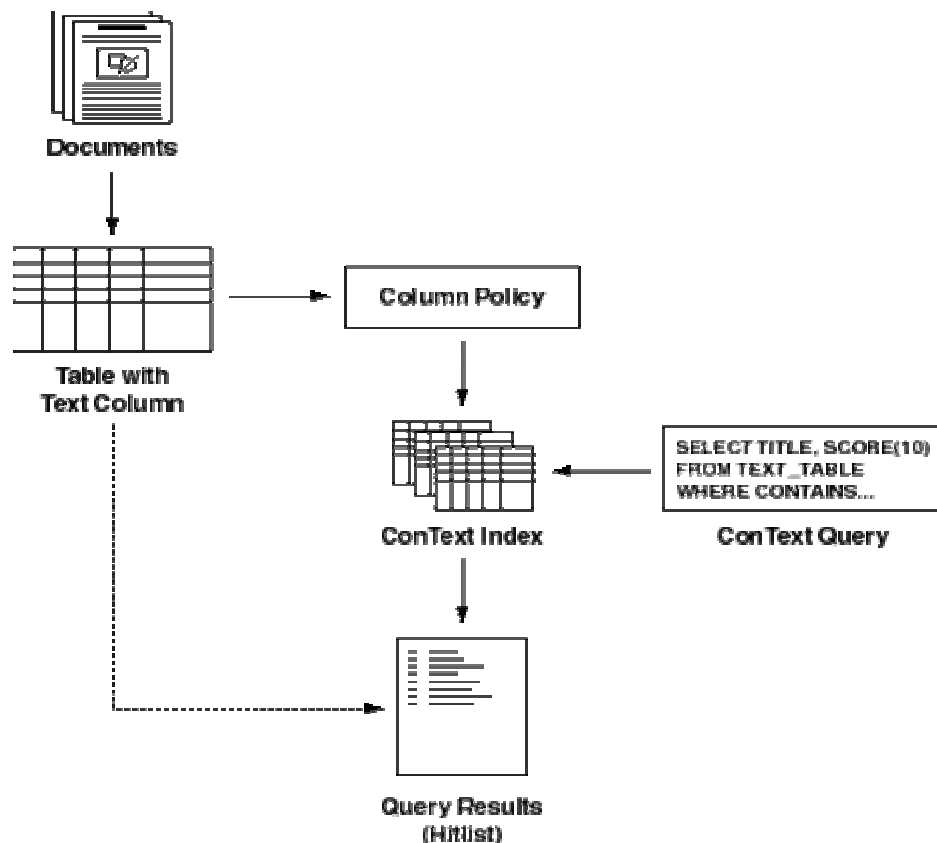
Στη συνέχεια, επιστρέφονται τα αποτελέσματα στο χρήστη σε μορφή *λίστας επιτυχιών* (*hit list*). Το αποτέλεσμα της ερώτησης μπορεί να επιστραφεί απευθείας ή να συνδυαστεί με δεδομένα από τον κανονικό πίνακα για να διαμορφώσει ο χρήστης διαφορετικά την ερώτηση ή να προβληθούν περισσότερες πληροφορίες σχετικά με τα έγγραφα που ικανοποιούν την ερώτησή του.

4.3.1 Πολιτικές (*policies*) ConText

Για να δημιουργήσει ένα ευρετήριο για κείμενο αποθηκευμένο σε στήλη της βάσης, ο διακομιστής ConText χρειάζεται τις εξής πληροφορίες σχετικά με το κείμενο :

- Πώς αποθηκεύεται το κείμενο στη στήλη;
- Ποια είναι η μορφοποίησή του;
- Πώς θα αναγνωρισθούν τα τεκμήρια;
- Πώς θα δημιουργηθεί το ευρετήριο και πού θα αποθηκευθεί;
- Θα χρησιμοποιηθεί κάποια ειδική παράμετρος στις ερωτήσεις;
- Υπάρχουν λέξεις που δεν θα πρέπει να ευρετηριοποιηθούν;

Η πολιτική λοιπόν παρέχει αυτές τις πληροφορίες για την προς ευρετηριοποίηση



Σχήμα 4.2: Πίνακες κειμένου, πολιτικές, ευρετήρια ConText και ερωτήσεις

στήλη με τη μορφή επιλογών του ευρετηρίου (μία επιλογή για κάθε απαίτηση). Οι πολιτικές μπορούν να δημιουργηθούν από οποιοδήποτε χρήστη που του έχει δοθεί ο ρόλος³ CTXAPP και αποθηκεύονται στο λεξικό δεδομένων του ConText.

Από τα παραπάνω γίνεται κατανοητό ότι χωρίς να έχει προσδιορισθεί πολιτική δεν μπορεί να γίνει ευρετηριοποίηση.

Οι πολιτικές που δημιουργεί ο χρήστης πρέπει να είναι μοναδικές για αυτόν. Η ίδια πολιτική δεν μπορεί να εφαρμοστεί σε παραπάνω από μια στήλες.

4.3.1.1 Πολιτική στήλης (*column policy*)

Πολιτική στήλης ονομάζεται η πολιτική που έχει καθοριστεί σε μια στήλη κειμένου. Μόνο πολιτικές στηλών μπορούν να χρησιμοποιηθούν για να δημιουργηθούν ευρετήρια ConText.

4.3.1.2 Πολιτική πρότυπο (*policy template*)

Πολιτική πρότυπο ονομάζεται η πολιτική η οποία δεν έχει καθοριστεί σε στήλη κειμένου. Οι πολιτικές αυτές χρησιμοποιούνται ως βάση για τη δημιουργία άλλων πολιτικών προτύπων ή στηλών. Όταν η πολιτική πρότυπο χρησιμοποιείται για τη δημιουργία άλλων πολιτικών, οι επιλογές της πολιτικής αυτής χρησιμοποιούνται ως οι εξ' ορισμού επιλογές της νέας πολιτικής. Το ConText cartridge έχει μια πολιτική πρότυπο, που λέγεται

³ Ρόλος (role) στα πλαίσια του Σ.Δ.Β.Δ. λέγεται μια ομάδα δικαιωμάτων που δίνεται στους χρήστες για λόγους διευκόλυνσης της διαχείρισης. Παράδειγμα ρόλου είναι ο DBA (DataBase Administrator) στην Oracle.

DEFAULT_POLICY και που αποτελεί τη βάση για όλες τις πολιτικές. Φυσικά, προσφέρεται η δυνατότητα να ξανά οριστούν οι επιλογές κάθε πολιτικής καθορίζοντας ρητά το όνομα της επιλογής. Οι εξ' ορισμού πολιτικές που προσφέρει το ConText περιέχονται στο σχήμα⁴ CTXSYS.

4.3.1.3 Πολλαπλές πολιτικές σε μια στήλη

Μπορούν να καθορισθούν πολλαπλές πολιτικές σε μια στήλη, αρκεί καθεμία να είναι μοναδική για το χρήστη. Ως αποτέλεσμα, μια στήλη μπορεί να έχει παραπάνω από ένα ευρετήρια. Όταν υποβάλλεται μια ερώτηση, μπορεί να ορίζεται το όνομα της πολιτικής, για να καθορίζει έτσι ο χρήστης ποιο ευρετήριο θα χρησιμοποιήσει ο ConText διακομιστής.

Αυτό το χαρακτηριστικό είναι ιδιαίτερα χρήσιμο στην περίπτωση που υπάρχει κείμενο στην αγγλική για το οποίο ο χρήστης θέλει να εκτελεί τόσο απλές ερωτήσεις όσο και ερωτήσεις θεμάτων. Για να γίνει αυτό, ο χρήστης θα πρέπει να δημιουργήσει δυο πολιτικές, μία για την ευρετηριοποίηση για απλό κείμενο και μία για την ευρετηριοποίηση θεμάτων.

4.3.1.4 Χαρακτηριστικά της DEFAULT_POLICY πολιτικής

Η εξ' ορισμού πολιτική πρότυπο DEFAULT_POLICY χρησιμοποιεί τις εξής επιλογές :

DEFAULT_DIRECT_DATASTORE	Δείχνει ότι το κείμενο αποθηκεύεται απευθείας στη βάση.
DEFAULT_NULL_FILTER	Δείχνει ότι το κείμενο είναι απλό κείμενο ASCII, χωρίς να χρειάζεται φιλτράρισμα για την ευρετηριοποίηση και τον τονισμό.
DEFAULT_LEXER	Δείχνει ότι χρησιμοποιούνται οι εξ' ορισμού επιλογές για την αναγνώριση των λέξεων και τα όρια των προτάσεων.
DEFAULT_INDEX	Δείχνει το ποσό μνήμης που θα δεσμευτεί για την ευρετηριοποίηση. Η εξ' ορισμού τιμή είναι 12.582.912 bytes.
NO_SOUNDSEX	Δείχνει ότι δεν θα αποθηκευθούν παράμετροι στο ευρετήριο που χρησιμοποιούνται για την ευρετηριοποίηση και των «ηχητικά» παρόμοιων λέξεων.
DEFAULT_STOPLIST	Δείχνει ότι θα χρησιμοποιηθεί η εξ' ορισμού λίστα στοπ λέξεων.

4.3.2 Επιλογές ευρετηριοποίησης

Οι επιλογές (*preferences*) ευρετηριοποίησης χρησιμοποιούνται από το cartridge για τη δημιουργία των ευρετηρίων ConText. Υπάρχουν έξι κατηγορίες επιλογών, που αντιστοιχούν στις πληροφορίες που χρειάζεται το ConText cartridge για τα ευρετήρια. Οι επιλογές των κατηγοριών αυτών έχουν σχέση με την αποθήκευση των δεδομένων, τα φίλτρα, την αναγνώριση των τεκμηρίων, τη μηχανή ευρετηριοποίησης, την επεξεργασία των λέξεων και τη λίστα με τις στοπ λέξεις.

Οι επιλογές που είναι διαθέσιμες σχετικά με την αποθήκευση έχουν σχέση με το που θα αποθηκεύονται τα έγγραφα. Μπορούν έτσι να αποθηκεύονται απευθείας στη βάση,

⁴ Σχήμα (schema) είναι μια συλλογή αντικειμένων της βάσης (πίνακες, ευρετήρια, διαδικασίες, κτλ). Στην Oracle, οι έννοιες σχήμα και χρήστης είναι ταυτόσημες.

σε αρχείο, να καλούνται από διεύθυνση URL⁵ και να αποθηκεύονται και πάλι στη βάση αλλά σε πίνακες *master-detail*, είτε σε δυαδική μορφή είτε σε μορφή κειμένου.

Οι επιλογές σε σχέση με τα φίλτρα έχουν να κάνουν με το πώς θα το ConText θα διαβάσει τα έγγραφα, δηλαδή καθορίζεται σε ποια μορφοποίηση είναι τα έγγραφα, αν είναι απλό κείμενο HTML ή με εξελεγμένη μορφοποίηση HTML και παράλληλα υπάρχει ειδική παράμετρος για τη μορφοποίηση εγγράφων σε Microsoft Word 6.

Η τρίτη κατηγορία επιλογών έχει σχέση με τα λεκτικά, τον τρόπο δηλαδή που θα σαρώσει το ConText το κείμενο για να αναγνωρίσει τα τεκμήρια, πώς θα αναγνωρίζει τις λέξεις, τα άκρα των προτάσεων, κτλ. Υπάρχουν επιλογές για αναγνώριση HTML εγγράφων, για απλό κείμενο, για κείμενο στην κορεατική και κινεζική γλώσσα και επιλογή για δημιουργία ευρετηρίων θεμάτων.

Μία είναι η διαθέσιμη επιλογή που αφορά στη μηχανή ευρετηριοποίησης και έχει σχέση με το ποσό μνήμης που θα χρησιμοποιήσει το ConText για την ευρετηριοποίηση.

Η επόμενη κατηγορία επιλογών έχει σχέση με τις λέξεις, αν δηλαδή θα δεικτοδοτηθούν λέξεις που μοιάζουν «χηητικά» και αν θα χρησιμοποιηθεί *fuzzy* ταύτιση για κείμενο στην κορεατική, κινεζική και ιαπωνική γλώσσα.

Η τελευταία κατηγορία επιλογών έχει σχέση με τη λίστα των στοπ λέξεων, αν θα χρησιμοποιηθεί ή όχι.

Όταν δημιουργείται μια πολιτική, καθορίζονται έξι παράμετροι, μια για κάθε κατηγορία που αναφέρθηκε παραπάνω. Αν δεν καθοριστεί κάποια από αυτές τις παραμέτρους χρησιμοποιείται η αντίστοιχη επιλογή από την `DEFAULT_POLICY` πολιτική. Δύο επιλογές της ίδιας κατηγορίας δεν μπορούν να χρησιμοποιηθούν για την ίδια πολιτική. Ακόμη κι όταν αυτό είναι θεμιτό, πρέπει να δημιουργηθούν δύο πολιτικές που θα έχουν τα ίδια χαρακτηριστικά, αλλά διαφορετικό όνομα.

Οι διαφορετικές τιμές που μπορεί να πάρει μια παράμετρος δίνονται από κάποιες σταθερές του λεξικού δεδομένων του ConText και ονομάζονται *tiles*.

Υπάρχει ένας αριθμός ορισμένων από πριν επιλογών (ανήκουν στο σχήμα `CTXSYS`) έτσι ώστε ο χρήστης, που του έχει δοθεί ο ρόλος `CTXAPP`, να μπορεί να δημιουργήσει κατευθείαν πολιτικές χωρίς να δημιουργήσει πρώτα επιλογές.

4.4 Ενεργοποίηση του ConText cartridge, τερματισμός και εκκίνηση ενός διακομιστή ConText

4.4.1 Ενεργοποίηση του ConText

Οι παρακάτω οδηγίες έχουν δοκιμαστεί με επιτυχία σε λειτουργικό σύστημα Windows NT Server 4 (service pack 6a), Oracle Enterprise Edition 8.0.5 (με patchset 80522) και ConText Cartridge 2.4.5 (αναβαθμισμένο σε 2.4.5.4). Κατά τη διάρκεια της εγκατάστασης της βάσης, ο χρήστης ρωτάται αν θέλει να εγκαταστήσει κάποιο από τα cartridges τα οποία περιλαμβάνει η βάση. Αφού ο χρήστης επιλέξει το ConText, στη συνέχεια υπάρχει μια σειρά εργασιών προκειμένου να ενεργοποιηθεί το cartridge.

1. Η παράμετρος `SHARED_POOL_SIZE` στο αρχείο `init.ora` που διαβάσει η Oracle κατά την εκκίνηση, πρέπει να έχει τιμή μεγαλύτερη του 11.534.336 bytes. Αφού γίνει η αλλαγή, πρέπει να γίνει επανεκκίνηση της βάσης προκειμένου να διαβαστεί η αλλαγή.

⁵ Uniform Resource Locator. Ονομάζεται η ηλεκτρονική διεύθυνση (πχ. <http://www.oracle.com>) που χρησιμοποιείται στο διαδίκτυο και αποτελείται από τέσσερα μέρη, το όνομα του πρωτοκόλλου (πχ. `http`), το όνομα της τοποθεσίας που δείχνει η διεύθυνση (`www`), το όνομα του οργανισμού που διατηρεί τη σελίδα (πχ. `oracle`), καθώς και μια κατάληξη που δείχνει το είδος του οργανισμού (πχ. `com`).

2. Αν το λειτουργικό σύστημα είναι τα Windows NT, τότε από το Control Panel, πρέπει να ξεκινήσει η υπηρεσία (*service*) με όνομα OracleConTextService80.
3. Σε αυτό το βήμα δημιουργείται ο χρήστης CTXSYS. Ανοίγουμε το πρόγραμμα Oracle Database Assistant, επιλέγουμε Modify a database, πατάμε Next, επιλέγουμε το στιγμιότυπο της βάσης στο οποίο θα ενεργοποιηθεί το ConText, εισάγουμε τον κωδικό (*password*) για το χρήστη internal και πατάμε Next. Στην επόμενη οθόνη επιλέγουμε Configure Oracle cartridges και Πατάμε Next. Επιλέγουμε ConText cartridge. Αν επιλέξουμε Enable Demo, τότε θα δημιουργηθεί ένας χρήστης CTXDEMO με κάποιους πίνακες επίδειξης. Πατώντας Finish, θα αρχίσει η διαδικασία ενεργοποίησης του ConText cartridge για το συγκεκριμένο στιγμιότυπο βάσης που επιλέχθηκε στην αρχή. Αν επιλέχθηκε η επιλογή Enable Demo, τότε θα πρέπει να έχει ήδη ξεκινήσει ένας διακομιστής ConText από τη γραμμή εντολών των Windows NT δίνοντας την εντολή :

```
START CTXSRV80 -USER CTXSYS/PASSWORD -PERSONALITY DQL
```

Αυτή η εντολή ξεκινά ένα διακομιστή ConText από το χρήστη CTXSYS με κωδικό CTXSYS (που είναι ο εξ' ορισμού).

Στο σημείο αυτό θα πρέπει να έχει τελειώσει το στάδιο ενεργοποίησης του ConText cartridge.

4.4.2 Εκκίνηση και τερματισμός διακομιστή

Για την εκκίνηση ενός διακομιστή ConText χρησιμοποιείται το εργαλείο γραμμής εντολών ctxsrv80 (για περιβάλλον Windows NT). Η σύνταξή του είναι :

```
ctxsrv80 -user ctxsys /passwd[@sqlnet_address]
        [-personality RQDML]
        [-logfile log_name]
        [-sqltrace]
```

όπου -user είναι η παράμετρος που δίνει το όνομα του χρήστη (πάντα ctxsys, είναι ο μόνος χρήστης που μπορεί να εκκινήσει ένα ConText διακομιστή)

passwd είναι ο κωδικός για το χρήστη ctxsys

@sqlnet_address είναι *όνομα είδωλο (alias)* μιας απομακρυσμένης βάσης

-personality είναι το όνομα του ρόλου με τον οποίο θα ξεκινήσει ο διακομιστής και μπορεί να έχει τις τιμές R, Q, D, M ή L

-logfile *όνομα αρχείου* είναι το όνομα του αρχείου ημερολογίου στο οποίο θα γράφονται όλες οι πληροφορίες από το διακομιστή

-sqltrace είναι μια παράμετρος που καθορίζει την εγγραφή πληροφοριών για αποσφαλμάτωση στον κατάλογο που καθορίζεται από την μεταβλητή USER_DUMP_DEST του αρχείου initorcl.ora

Ένα άλλο εργαλείο είναι το ctxctl80 (για περιβάλλον Windows NT), με το οποίο μπορεί να δει κανείς την κατάσταση των διακομιστών ConText, ποιοι έχουν ξεκινήσει και με ποιο ρόλο, καθώς και να εκκινήσει ή σταματήσει ένα διακομιστή. Και αυτό το εργαλείο, όπως και το ctxsrv80, τρέχει από τη γραμμή εντολών των Windows. Με *start αριθμός διακομιστών ρόλος* ξεκινάνε τόσοι διακομιστές όσοι ορίζονται με τον αντίστοιχο ρόλο. Με *status* ο χρήστης μπορεί να δει ποιες διεργασίες ConText διακομιστών έχουν ξεκινήσει και με *stop pid*, όπου *pid* είναι ο αριθμός της διεργασίας, να σταματήσει ένα διακομιστή.

4.5 Ρόλοι και χρήστες του ConText στη βάση

Με την εγκατάσταση και την ενεργοποίηση του ConText για το στιγμιότυπο της βάσης, δημιουργούνται τρεις νέοι ρόλοι της βάσης. Οι ρόλοι αυτοί είναι οι εξής :

1. CTXADMIN ρόλος που παρέχει δικαίωμα στο χρήστη που θα του δοθεί να εκτελέσει όλες τις ConText λειτουργίες. Θα πρέπει να δίνεται σε χρήστες που εκτελούν διαχειριστικές εργασίες για το ConText cartridge.
2. CTXAPP ρόλος που παρέχει τα δικαιώματα στο χρήστη να χειρίζεται το λεξικό δεδομένων του ConText, να χειρίζεται τα γλωσσολογικά και να δίνει ερωτήσεις προς εκτέλεση. Θα πρέπει να δίνεται στους χρήστες που αναπτύσσουν εφαρμογές ConText.
3. CTXUSER ρόλος που παρέχει το δικαίωμα εκτέλεσης ερωτήσεων (κειμένου και θεμάτων). Πρέπει να δίνεται σε χρήστες που χρησιμοποιούν τις εφαρμογές ConText.

Δεν είναι απαραίτητο να δίνονται παραπάνω από ένας ρόλοι σε ένα χρήστη, εφόσον αυτοί είναι σε μια ιεραρχία με την εξής σειρά : CTXADMIN, CTXAPP και CTXUSER.

Παράλληλα με τους ρόλους, δημιουργείται και ο χρήστης CTXSYS που χρησιμοποιείται για διαχειριστικούς λόγους, όπως να εκκινεί και να σταματά διακομιστές ConText και να διαχειρίζεται τις ουρές. Έχει το ρόλο του διαχειριστή βάσης (DBA), είναι ιδιοκτήτης του λεξικού δεδομένων του ConText και είναι κάτοχος του ρόλου CTXADMIN.

Εάν κατά την ενεργοποίηση του ConText στη βάση επιλέχθηκε η επιλογή *Enable Demo*, τότε δημιουργήθηκε και ένας άλλος χρήστης με όνομα CTXDEMO για σκοπούς επίδειξης. Ο χρήστης αυτός είναι κάτοχος του ρόλου CTXAPP και παράλληλα έχει κάποιους πίνακες, όψεις, κτλ. μαζί με αντικείμενα του λεξικού δεδομένων του ConText (πολιτικές, επιλογές, κτλ.).

4.6 Είδη ερωτήσεων

Υπάρχουν τρεις τρόποι για την εκτέλεση ερωτήσεων από το ConText cartridge. Τα είδη των ερωτήσεων είναι τα εξής :

1. Ερωτήσεις δύο βημάτων
2. Ερωτήσεις ενός βήματος και
3. Ερωτήσεις από μνήμη

4.6.1 Ερωτήσεις δύο βημάτων

Για την εκτέλεση τέτοιου είδους ερωτήσεων, απαιτούνται δύο εργασίες. Η πρώτη εργασία είναι η δημιουργία του πίνακα αποτελεσμάτων. Ο πίνακας μπορεί να ονομάζεται με οποιοδήποτε όνομα, αλλά πρέπει να έχει την εξής δομή :

```
create table ctx_temp (textkey varchar2(64), score number, conid number);
```

Από τη δήλωση αυτή (ονομάζουμε τον πίνακα `ctx_temp`) φαίνεται και η δομή του. Η στήλη `textkey` περιέχει το κύριο κλειδί των εγγράφων, ενώ η στήλη `score` περιέχει ένα αριθμό που αποτελεί το σκορ κάθε εγγραφής που δημιουργεί το ConText. Η τρίτη στήλη χρησιμοποιείται για την αποθήκευση ενός αριθμού για κάθε ερώτηση σε περιβάλλον πολλών χρηστών οι οποίοι δίνουν πολλές ερωτήσεις ο καθένας.

Οι ερωτήσεις δύο βημάτων χρησιμοποιούνται συνήθως σε εφαρμογές βασισμένες σε PL/SQL που απαιτούν όλα τα αποτελέσματα των ερωτήσεων. Τα πλεονεκτήματα αυτής της μεθόδου είναι ότι οι πίνακες των αποτελεσμάτων (αυτοί που δημιουργούνται στο

πρώτο βήμα) μπορούν να χειριστούν από το χρήστη, οι ερωτήσεις σε γενικές γραμμές είναι γρηγορότερες από εκείνες του ενός βήματος, μπορεί να γίνει ανάκληση όλων των αποτελεσμάτων με μία προσπάθεια καθώς επίσης ότι μπορεί να υπάρχει μια δομημένη συνθήκη. Τα μειονεκτήματα της μεθόδου είναι ότι απαιτείται προ-εκχώρηση των πινάκων αποτελεσμάτων, οι ερωτήσεις χρησιμοποιούν ονόματα πολιτικών, απαιτείται *σύνδεση* (*join*) για να ανακληθούν λεπτομερείς πληροφορίες για τα έγγραφα που επέστρεψε η εκτέλεση των ερωτήσεων, καθώς βέβαια και το γεγονός ότι είναι μια διαδικασία δύο βημάτων για την περάτωσή της.

4.6.2 Ερωτήσεις ενός βήματος

Οι ερωτήσεις αυτού του τύπου είναι συνήθως απλές ερωτήσεις SQL που περιέχουν τη συνάρτηση `CONTAINS` στο τμήμα `WHERE` της SQL δήλωσης. Στην περίπτωση αυτή, δεν απαιτούνται ενδιάμεσοι πίνακες αποτελεσμάτων.

Ερωτήσεις αυτού του τύπου χρησιμοποιούνται περισσότερο για διαδραστικές (*interactive*) ερωτήσεις που δίνει ο χρήστης. Πλεονεκτήματά τους είναι ότι δεν υπάρχει απαίτηση για ύπαρξη ενδιάμεσων πινάκων αποτελεσμάτων, η χρησιμοποίηση στάνταρ ερωτήσεων SQL, η χρησιμοποίηση ονομάτων πινάκων και στηλών και η επιστροφή των αποτελεσμάτων σε ένα βήμα. Στα μειονεκτήματα συγκαταλέγονται η σε γενικές γραμμές πιο αργή εκτέλεση σε σχέση με τους άλλους δύο τύπους ερωτήσεων και η αδυναμία χρησιμοποίησής τους σε εφαρμογές PL/SQL λόγω και του γεγονότος ότι δεν υπάρχει ενδιάμεσος πίνακας αποτελεσμάτων που θα βοηθούσε τον προγραμματιστή σε ορισμένες λειτουργίες.

4.6.3 Ερωτήσεις από μνήμη

Αυτή η κατηγορία ερωτήσεων χρησιμοποιεί ένα χώρο στη μνήμη (*buffer*) και ένα δρομέα (*cursor*) για να επιστρέψει τα αποτελέσματα. Επιστρέφοντας τα αποτελέσματα των ερωτήσεων σε ένα χώρο στη μνήμη βελτιώνει τη συνολική επίδοση σε σχέση με το γράψιμο και το διάβασμα από πίνακες στη βάση, κάτι που γίνεται εκτενώς στις ερωτήσεις των άλλων τύπων.

Τέτοιου είδους ερωτήσεις χρησιμοποιούνται σε εφαρμογές PL/SQL στις οποίες υπάρχει μια μεγάλη λίστα επιτυχιών από την οποία όμως υπάρχει ενδιαφέρον για ένα υποσύνολό της. Πλεονεκτήματα αυτού του είδους ερωτήσεων είναι ότι δεν χρησιμοποιούνται ενδιάμεσοι πίνακες, υπάρχει γρηγορότερη απόκριση σε σχέση με τις ερωτήσεις δύο βημάτων λόγω του γεγονότος ότι θα ληφθεί μόνο ένα υποσύνολο του συνόλου των αποτελεσμάτων, μπορεί να καθοριστεί ο αριθμός των αποτελεσμάτων που θα ληφθούν και η ερώτηση μπορεί να περιέχει μια δομημένη συνθήκη. Μειονεκτήματα της μεθόδου είναι η χρησιμοποίηση ονομάτων πολιτικών, το γεγονός ότι η ανάκληση όλων των αποτελεσμάτων δεν είναι δυνατή με μία προσπάθεια, η διαφορά στην απόδοση για μικρές λίστες αποτελεσμάτων είναι αμελητέα σε σχέση με τις ερωτήσεις δύο βημάτων, απαιτεί τρία βήματα, περιλαμβανομένου και ενός βρόχου για την περάτωση και δεν υποστηρίζονται οι τελεστές για το μέγιστο, προηγούμενο και επόμενο.

4.6.4 Παράδειγμα ερώτησης δύο βημάτων

Η παρακάτω ερώτηση χρησιμοποιεί την πολιτική με όνομα `ARTICLES_POL` για την αναζήτηση σε στήλη κειμένου ενός πίνακα με όνομα `TEXTTAB` για οποιαδήποτε άρθρα περιέχουν τη λέξη 'petroleum' (πετρέλαιο). Η διαδικασία `CONTAINS` γεμίζει τον πίνακα

CTX_TEMP με τα κύρια κλειδιά των άρθρων που ικανοποιούν την ερώτηση. Η δήλωση SELECT στη συνέχεια προκαλεί σύζευξη των αποτελεσμάτων από τον πίνακα CTX_TEMP με τον TEXTTAB για να δημιουργηθεί μια λίστα με τα άρθρα ταξινομημένα με το σκορ.

Πρώτα απ' όλα δημιουργείται ο πίνακας CTX_TEMP :

```
create table CTX_TEMP(
    textkey varchar2(64),
    score number,
    conid number);
```

Η H CONTAINS συμπληρώνει τον CTX_TEMP με τα έγγραφα που ικανοποιούν την ερώτηση :

```
execute ctx_query.contains('ARTICLE_POLICY','petroleum','CTX_TEMP');
```

Τέλος, γίνεται ανάκτηση των εγγράφων που ικανοποιούν την ερώτηση ταξινομημένα με φθίνουσα σειρά βάσει του σκορ :

```
SELECT SCORE, title
FROM CTX_TEMP, TEXTTAB
WHERE texttab.PK=ctx_temp.textkey
ORDER BY SCORE DESC;
```

4.6.5 Παράδειγμα ερώτησης ενός βήματος

Η παρακάτω ερώτηση είναι μια ερώτηση ενός βήματος. Γίνεται αναζήτηση στον πίνακα TEXTTAB για άρθρα που περιέχουν τη λέξη 'petroleum'.

```
SELECT *
FROM texttab
WHERE CONTAINS (text, 'petroleum') > 0;
```

4.6.6 Παράδειγμα ερώτησης από μνήμη

Το επόμενο παράδειγμα είναι ένα μπλοκ σε PL/SQL. Η χρησιμοποίηση τέτοιου είδους ερωτήσεων γίνεται μόνο από PL/SQL. Το παράδειγμα χρησιμοποιεί την πολιτική με όνομα ARTICLES_POL για την αναζήτηση στον πίνακα TEXTTAB για άρθρα που περιέχουν τη λέξη 'petroleum'.

```
declare
    score    char(5);
    pk       char(5);
    curid    number;
    title    char(256);

begin
    dbms_output.enable(100000);
    curid := ctx_query.open_con(
        policy_name => 'ARTICLES_POL',
        text_query  => 'petroleum',
        score_sorted => true,
        other_cols  => 'title');
    while (ctx_query.fetch_hit(curid, pk, score, title)>0)
    loop
        dbms_output.put_line(score||pk||substr(title,1,50));
    end loop;
    ctx_query.close_con(curid);
end;
```

Ο παραπάνω PL/SQL κώδικας εκτελεί τα εξής :

1. Καλείται η συνάρτηση `CTX_QUERY.OPEN_CON` η οποία ανοίγει ένα δρομέα στο χώρο μνήμης, προσπελαύνει τη στήλη κειμένου χρησιμοποιώντας την πολιτική που δίνεται καθώς και τη συνθήκη της ερώτησης και αποθηκεύει στο χώρο της μνήμης τα κλειδιά και τα σκορ των άρθρων που ικανοποιούν την ερώτηση.
2. Καλείται η συνάρτηση `CTX_QUERY.FETCH_HIT` για κάθε κλειδί στο χώρο μνήμης για να ανακτήσει τα έγγραφα, ένα κάθε φορά, μέχρι ο επιθυμητός αριθμός εγγράφων επιστραφεί ή να μην απομένουν άλλα στο χώρο μνήμης.
3. Καλείται η `CTX_QUERY.CLOSE_CON` διαδικασία για να απελευθερωθεί ο χώρος μνήμης που δεσμεύτηκε.

4.7 Ερωτήσεις

Μία ερώτηση καθορίζει τα κριτήρια αναζήτησης για την ανάκτηση εγγράφων χρησιμοποιώντας το ConText. Η ερώτηση αποτελείται από τους όρους (λέξεις ή φράσεις) και άλλα στοιχεία όπως τελεστές και ειδικούς χαρακτήρες που καθορίζει ο χρήστης για να ορίσουν ποια έγγραφα θα ανακτηθούν από το ConText.

Μία ερώτηση μπορεί επίσης να καλεί αποθηκευμένες ερωτήσεις για την επιστροφή αποθηκευμένων αποτελεσμάτων ή συναρτήσεις PL/SQL για να επιστρέψουν τιμές που χρησιμοποιούνται μέσα σε αυτήν.

Οι όροι της ερώτησης μπορεί να είναι λέξεις ή φράσεις. Οι λέξεις αποτελούν τα τεκμήρια πάνω στα οποία ενεργούν οι τελεστές. Εάν πολλές λέξεις χωρισμένες με κενό (όχι τελεστές) περιέχονται σε μια ερώτηση, τότε όλη αυτή η σειρά των λέξεων θεωρείται μια πρόταση.

Οι όροι μπορεί να περιέχουν και στοπ λέξεις. Όταν καθορίζεται μια πολιτική, το ConText επιτρέπει τον ορισμό στοπ λέξεων, οι οποίες δεν ευρετηριοποιούνται. Οι στοπ λέξεις δεν έχουν καμιά επίδραση στα ευρετήρια θεμάτων που θα περιγραφούν παρακάτω.

Εκτός των παραπάνω (λέξεις, φράσεις και στοπ λέξεις), μια ερώτηση μπορεί να περιέχει τα εξής :

- Τελεστές, που ορίζουν τις σχέσεις μεταξύ των όρων της ερώτησης και καθορίζουν την έξοδό της. Υπάρχουν λογικοί τελεστές, τελεστές για τη διαβάθμιση, για την εγγύτητα, την επαύξηση και το θησαυρό.
- Χαρακτήρες μάσκες, που επαυξάνουν τους όρους της ερώτησης χρησιμοποιώντας ταύτιση σχεδίων.
- Χαρακτήρες για γκρουπ όρων και τελεστών.
- *Εκφράσεις αποθηκευμένων ερωτήσεων (stored query expressions)*, που επιστρέφουν αποτελέσματα ερωτήσεων που έχουν εκτελεστεί και τα οποία είναι αποθηκευμένα σε ειδικούς πίνακες.
- Συναρτήσεις PL/SQL που επιστρέφουν κάποια αποτελέσματα στην ερώτηση.

Το ConText υποστηρίζει ερωτήσεις με διάκριση ή μη κεφαλαίων–πεζών. Για τις απλές ερωτήσεις κειμένου, η διάκριση βασίζεται στις επιλογές του λεκτικού που χρησιμοποιείται για την ευρετηριοποίηση των κειμένων. Εξ' ορισμού δεν γίνεται διάκριση. Η διάκριση κεφαλαίων–πεζών θα πρέπει να χρησιμοποιείται για λέξεις που έχουν διαφορετικό νόημα όταν έχουν διαφορετική ορθογραφία όσον αφορά στα γράμματα των λέξεων αυτών. Όταν χρησιμοποιείται η διάκριση κεφαλαίων–πεζών ισχύει και για τις στοπ λέξεις.

Οι ερωτήσεις θεμάτων γίνονται με διάκριση κεφαλαίων–πεζών.

Η γερμανική και ολλανδική γλώσσα υποστηρίζουν σύνθετες λέξεις. Με το ConText μπορεί να υπάρξει ένα ευρετήριο που να υποστηρίζει αυτές τις γλώσσες και να γίνεται αναζήτηση με βάση απλές λέξεις.

Για γλώσσες που χρησιμοποιούν 8-bit σελ χαρακτηριστών, όπως τα γαλλικά ή ισπανικά, το ConText έχει την επιλογή για μετατροπή των χαρακτηριστών στη βασική τους αναπαράσταση. Αυτό σημαίνει ότι χαρακτήρες με τόνους και με τα διάφορα σημάδια τονισμού μετατρέπονται στους αντίστοιχους μη τονισμένους. Όταν καθοριστεί αυτή η μετατροπή κατά την ευρετηριοποίηση, το ConText μετατρέπει τους χαρακτήρες του όρου της ερώτησης στους αντίστοιχους μη τονισμένους. Για παράδειγμα, αναζήτηση για την ισπανική λέξη *‘manaña’* θα επιστρέψει έγγραφα που περιέχουν τόσο τη λέξη *‘manaña’* όσο και την *‘manana’*. Αν ωστόσο δεν καθοριστεί η μετατροπή, τότε θα επιστραφούν μόνο τα έγγραφα που περιέχουν τη λέξη *‘manaña’*.

Για ερωτήσεις θησαυρού, δεν γίνεται μετατροπή των χαρακτηριστών στη βασική τους αναπαράσταση, διότι η μετατροπή αυτή λαμβάνει χώρα μετά την αναζήτηση στο θησαυρό.

4.7.1 Λογικοί τελεστές

Οι λογικοί τελεστές συνδυάζουν τους όρους μιας ερώτησης. Αυτό ισχύει για όλες τις λέξεις ή φράσεις. Όταν συμβαίνει κάτι τέτοιο, ο αριθμός των κενών χαρακτηριστών (διαστημάτων) γύρω από τους τελεστές δεν παίζει ρόλο.

Οι λογικοί τελεστές συνδέουν τους όρους με τρόπο ώστε να παράγουν σκορ που βασίζονται στη σχέση των όρων. Οι τελεστές συνδυάζουν τα σκορ των τελεστών τους σε μέγιστη τιμή 100.

Στον πίνακα 4.1 περιγράφονται οι λογικοί τελεστές. Τελεστές λέγονται οι όροι στους οποίους ενεργούν οι τελεστές.

Τελεστής	Σύνταξη	Περιγραφή
AND	term1&term2 term1 and term2	Επιστρέφει έγγραφα που περιέχουν τους όρους term1 και term2. Επιστρέφει το ελάχιστο σκορ των τελεστών του.
OR	term1 term2 term1 or term2	Επιστρέφει τα έγγραφα που περιέχουν έναν τουλάχιστον από τους όρους term1 ή term2. Επιστρέφει το μέγιστο σκορ των τελεστών του.
NOT	term1~term2 term1 not term2	Επιστρέφει τα έγγραφα που περιέχουν τον όρο term1 και όχι τον term2.
EQUIVALENCE	term1=term2 term1 equiv term2	Καθορίζει ότι ο όρος term2 είναι μια αποδεκτή αντικατάσταση του όρου term1.

Πίνακας 4.1: Λογικοί τελεστές

Φυσικά μπορούν να γίνουν συνδυασμοί των τελεστών, για παράδειγμα η πρόταση :
windows and (linux or unix)

ζητά τα έγγραφα που περιέχουν τις λέξεις windows και linux ή windows και unix.

Ο τελεστής NOT δεν επηρεάζει το σκορ που παράγουν οι άλλοι τελεστές.

4.7.2 Τελεστής WITHIN

Ο τελεστής WITHIN χρησιμοποιείται για να περιορίσει την ερώτηση σε μέρη ενός εγγράφου. Μέρη ενός εγγράφου μπορεί να είναι προτάσεις ή παράγραφοι και προκαθορισμένα μέρη. Ο τελεστής μπορεί να συνδυαστεί με τρεις τρόπους :

- *έκφραση* WITHIN SENTENCE : Με αυτό τον τρόπο γίνεται αναζήτηση για έγγραφα που περιέχουν την *έκφραση* μέσα στην ίδια πρόταση. Η *έκφραση* μπορεί να περιέχει τους τελεστές AND ή NOT.

- *έκφραση* WITHIN PARAGRAPH : Η σύνταξη αυτή αναζητεί για την *έκφραση* μέσα στην ίδια παράγραφο. Η *έκφραση* μπορεί να περιέχει τους τελεστές AND ή NOT.
- *όρος* WITHIN τμήμα : Γίνεται αναζήτηση του *όρου* σε ένα προκαθορισμένο *τμήμα*.

Ο τελεστής WITHIN δεν επηρεάζει το σκορ.

4.7.2.1 Παραδείγματα

Για να βρεθούν τα έγγραφα που περιέχουν τη λέξη dog και cat στην ίδια πρόταση :
(dog and cat) WITHIN SENTENCE

Για να βρεθούν τα έγγραφα που περιέχουν τη λέξη dog και cat στην ίδια παράγραφο :
(dog and cat) WITHIN PARAGRAPH

Για να βρεθούν τα έγγραφα που περιέχουν τη λέξη dog και όχι cat στην ίδια πρόταση :
(dog not cat) WITHIN SENTENCE

4.7.2.2 Ο τελεστής WITHIN σε προκαθορισμένα τμήματα

Ο τελεστής μπορεί να χρησιμοποιηθεί για να γίνει αναζήτηση σε προκαθορισμένο τμήμα, όπως αναφέρθηκε παραπάνω. Για παράδειγμα, σε ένα έγγραφο HTML ο διαχειριστής ConText μπορεί να καθορίσει ένα τμήμα βασισμένο στην παραπομπή <HEAD> και </HEAD> και να γίνει έτσι η αναζήτηση σε όλα τα έγγραφα μόνο στο συγκεκριμένο σημείο.

Για να χρησιμοποιήσει ο χρήστης αυτή τη δυνατότητα που προσφέρει ο τελεστής WITHIN, πρέπει να ξέρει το όνομα του τμήματος στο οποίο επιθυμεί να κάνει την αναζήτηση. Η λίστα των προκαθορισμένων τμημάτων μπορεί να βρεθεί από τις όψεις CTX_ALL_SECTIONS και CTX_USER_SECTIONS.

4.7.2.3 Παραδείγματα

Για να βρεθούν τα έγγραφα που περιέχουν τη φράση San Francisco μέσα στο προκαθορισμένο τμήμα με όνομα Headings :
San Francisco WITHIN Headings

Για να βρεθούν τα έγγραφα που περιέχουν τη λέξη sailing και τη φράση San Francisco μέσα στο προκαθορισμένο τμήμα με όνομα Headings :
(San Francisco WITHIN Headings) and sailing ή
sailing and San Francisco WITHIN Headings

Για να βρεθούν τα έγγραφα που περιέχουν τις λέξεις dog και cat μέσα στο ίδιο προκαθορισμένο τμήμα με όνομα Headings :
(dog and cat) WITHIN Headings

το οποίο νοηματικά είναι διαφορετικό από το :
dog WITHIN Headings and cat WITHIN Headings

που βρίσκει τα έγγραφα που περιέχουν τις λέξεις dog και cat, οι οποίες βρίσκονται στο τμήμα Headings, άσχετα με το γεγονός ότι μπορεί να βρίσκονται στο ίδιο τμήμα ή σε διαφορετικά.

4.7.2.4 Περιορισμοί του τελεστή WITHIN

Ο τελεστής έχει τους ακόλουθους περιορισμούς :

- Δεν υποστηρίζεται από το λεκτικό για τις ερωτήσεις θεμάτων
- Δεν μπορεί να περιέχεται σε μια φράση. Για παράδειγμα, απαγορεύεται η φράση : *όρος₁ WITHIN τμήμα όρος₂*
- Δεν μπορεί να συνδυαστεί με τελεστές επαύξησης
- Υποερωτήσεις που περνιούνται στον τελεστή δεν μπορούν να χρησιμοποιούν το μέγιστο ή πρώτο/επόμενο τελεστή
- Δεν μπορεί να φωλιάζεται με άλλους WITHIN τελεστές. Για παράδειγμα, η φράση *'dog WITHIN body WITHIN heading'*, απαγορεύεται
- Από τη στιγμή που η λέξη WITHIN είναι δεσμευμένη, πρέπει να συμπεριληφθεί σε εισαγωγικά προκειμένου να γίνει αναζήτηση για τη λέξη 'within'.

4.7.3 Τελεστής NEAR

Ο τελεστής NEAR επιστρέφει σκορ βασισμένο στην εγγύτητα δύο ή περισσότερων όρων. Το ConText επιστρέφει υψηλότερα σκορ για όρους που είναι κοντά και χαμηλότερα για όρους που είναι μακριά σε ένα έγγραφο.

Ο τελεστής χρησιμοποιείται μόνο για ερωτήσεις κειμένου κι όχι για ερωτήσεις θεμάτων. Η σύνταξή του είναι :

```
NEAR((word1, word2, ..., wordn)[, MAX_SPAN [, ORDER]])
```

όπου word_n είναι οι όροι χωρισμένοι με κόμμα. Οι όροι μπορεί να είναι λέξεις ή φράσεις. MAX_SPAN είναι προαιρετική παράμετρος που καθορίζει το μέγεθος της συστάδας. Συστάδα είναι το μεγαλύτερο γκρουπ λέξεων στο οποίο εμφανίζονται όλοι οι όροι. Εξ' ορισμού είναι 100, που είναι και το μεγαλύτερο μέγεθος. Για δύο όρους, η συστάδα είναι η μεγαλύτερη απόσταση μεταξύ των όρων αυτών. Η συστάδα αρχίζει και τελειώνει με ένα όρο. ORDER είναι μια προαιρετική παράμετρος του τύπου αλήθειας (TRUE/FALSE). Καθορίζει τη σειρά με την οποία θα γίνει η αναζήτηση. Αν είναι TRUE, τότε η αναζήτηση γίνεται με τη σειρά με την οποία καθορίστηκαν οι όροι στην ερώτηση. Εξ' ορισμού είναι FALSE. Η παράμετρος αυτή καθορίζεται μαζί με την MAX_SPAN.

Για παράδειγμα, αναζήτηση για τις λέξεις Monday, Tuesday, Wednesday για μέγιστο μέγεθος συστάδας 20 και για τη σειρά με την οποία δίνονται οι όροι, πρέπει να δοθεί : `near(Monday, Tuesday, Wednesday), 20, TRUE)`

Το ConText μπορεί να επιστρέφει διαφορετικά αποτελέσματα για το ίδιο έγγραφο και για τις ίδιες ερωτήσεις με την παράμετρο ORDER διαφορετική.

Το σκορ για τον τελεστή NEAR συνδυάζει τη συχνότητα εμφάνισης των όρων με την εγγύτητά τους. Για κάθε έγγραφο που ικανοποιεί την ερώτηση, το ConText επιστρέφει ένα σκορ μεταξύ 1 και 100 που είναι ανάλογο με τον αριθμό των συστάδων του εγγράφου και αντιστρόφως ανάλογο με το μέσο μέγεθος των συστάδων. Μικρός αριθμός συστάδων σε ένα έγγραφο σημαίνει υψηλότερο σκορ, διότι μικρός αριθμός σημαίνει εγγύτητα των όρων. Ο αριθμός των όρων επίσης επηρεάζει το σκορ. Ερωτήσεις με πολλούς όρους, όπως για παράδειγμα επτά, σε γενικές γραμμές χρειάζονται λιγότερες συστάδες σε ένα έγγραφο για σκορ 100, από ότι ερωτήσεις με λίγους όρους, όπως για παράδειγμα δύο.

Ο τελεστής μπορεί να χρησιμοποιηθεί με άλλους τελεστές, όπως ο AND ή OR. Με τον τελεστή WITHIN μπορεί να χρησιμοποιηθεί ως εξής :

```
Near((dog, cat), 10) WITHIN Headings
```

Όταν υπάρχουν τέτοιες εκφράσεις, το ConText κοιτάει για συστάδες που υπάρχουν μέσα στα δοθέντα τμήματα. Στο παραπάνω παράδειγμα, λαμβάνονται μόνο οι συστάδες που περιέχουν τις λέξεις dog και cat στο τμήμα Headings. Αν για παράδειγμα, η λέξη dog

υπάρχει μέσα στο τμήμα Headings και η λέξη cat είναι πέντε λέξεις μακριά από την dog, αλλά εκτός του τμήματος Headings, τότε το έγγραφο δεν ικανοποιεί την έκφραση αυτή.

4.7.4 Τελεστές αλλαγής του σκορ

Οι τελεστές αυτού του είδους συμπεριφέρονται με τρόπο ανάλογο των λογικών τελεστών. Ωστόσο, επηρεάζουν το σκορ διαφορετικά κι έτσι μπορούν να αλλάξουν τη θέση ενός εγγράφου στη λίστα επιτυχιών. Οι τελεστές αυτοί είναι οι ACCUMULATE, MINUS και WEIGHT.

4.7.4.1 Τελεστής ACCUMULATE

Συντάσσεται ως : *όρος₁,όρος₂* ή *όρος₁ accum όρος₂*

Επιστρέφει τα έγγραφα που περιέχουν έναν τουλάχιστον από τους όρους *όρος₁* ή *όρος₂*. Το σκορ υπολογίζεται προσθέτοντας το σκορ κάθε τελεστή. Παρόμοιο με το λογικό OR, με τη διαφορά ότι το επιστρεφόμενο σκορ είναι το *άθροισμα* όλων. Ο τελεστής OR βασίζει το σκορ του μόνο στον όρο που εμφανίζεται πιο συχνά στο έγγραφο, ενώ ο τελεστής συσσώρευσης, προσθέτει όλα τα σκορ, όλων των όρων, με πάνω όριο το 100, όταν το σκορ υπερβεί το νούμερο αυτό.

Ο τελεστής αυτός μπορεί να χρησιμοποιηθεί για να γίνει αναζήτηση οποιουδήποτε όρου της ερώτησης σε έγγραφα, από τα οποία αυτά που περιέχουν τις περισσότερες εμφανίσεις έχουν και το υψηλότερο σκορ. Αυτά δηλαδή τα έγγραφα εμφανίζονται και πρώτα στη λίστα επιτυχιών.

4.7.4.2 Τελεστής MINUS

Συντάσσεται ως : *όρος₁- όρος₂* ή *όρος₁ minus όρος₂*

Επιστρέφει τα έγγραφα που περιέχουν τον *όρο₁*, υπολογίζοντας το σκορ με αφαίρεση του αριθμού εμφανίσεων του *όρου₂* από τον αριθμό εμφανίσεων του *όρου₁*.

Ο τελεστής αυτός θα χρησιμοποιηθεί όταν είναι επιθυμητό ένα έγγραφο που περιέχει ένα όρο να τοποθετηθεί χαμηλότερα στη λίστα επιτυχιών, λόγω της ύπαρξης ενός άλλου όρου. Ο τελεστής αυτός είναι χρήσιμος για την ελάττωση του σκορ ενός εγγράφου που περιέχει «θόρυβο».

Για παράδειγμα, έστω ότι αναζήτηση για τον όρο cars (αυτοκίνητα) επιστρέφει πάντα έγγραφα με υψηλό σκορ που είναι σχετιζόμενα με αυτοκίνητα της Ford (Ford cars). Ο τελεστής θα μπορούσε να χρησιμοποιηθεί έτσι ώστε να ελαττώσει το σκορ αυτών των εγγράφων. Αυτό θα γινόταν ως εξής :

`'cars - Ford'`

Στην ουσία, επιστρέφονται τα έγγραφα που περιέχουν τη λέξη cars. Ωστόσο, το σκορ κάθε εγγράφου είναι ο αριθμός εμφανίσεων της λέξης cars μείον τον αριθμό εμφανίσεων της λέξης Ford. Όταν ένα έγγραφο δεν περιέχει τη λέξη Ford, τότε ο αριθμός εμφανίσεων της είναι μηδέν.

4.7.4.3 Τελεστής WEIGHT

Συντάσσεται ως : *όρος*v*

Επιστρέφει τα έγγραφα που περιέχουν τον *όρο*, υπολογίζοντας το σκορ με πολλαπλασιασμό του σκορ του *όρου* με το συντελεστή *v*, όπου *v* αριθμός μεταξύ 0.1 και 10. Το μέγιστο τελικό σκορ είναι 100.

Σε ερωτήσεις που περιέχουν παραπάνω από έναν όρους, ο τελεστής μπορεί να χρησιμοποιηθεί για να μεταβάλλει το σχετικό σκορ των όρων της ερώτησης. Το σκορ του

όρου μπορεί να μειωθεί αν ο συντελεστής είναι κάτω της μονάδας. Αντίθετα, μπορεί να αυξηθεί αν ο συντελεστής είναι μεταξύ μονάδας και του 10.

Ο τελεστής είναι χρήσιμος για ερωτήσεις που περιέχουν τους τελεστές συσσώρευσης και τους λογικούς OR ή AND όταν η ερώτηση περιέχει παραπάνω από έναν όρους. Χωρίς μεταβολή του βάρους κάθε όρου, το σκορ δεν μπορεί να πληροφορήσει σχετικά με το ποιος όρος εμφανίζεται πιο συχνά, γεγονός που ενδιαφέρει σε περιπτώσεις που ο χρήστης ενδιαφέρεται για ένα συγκεκριμένο όρο περισσότερο από ότι για έναν άλλο, η σειρά διαβάθμισης δεν λέει πολλά για το ποια έγγραφα περιέχουν τον όρο για τον οποίο ενδιαφέρεται ο χρήστης.

Για παράδειγμα, έστω ότι υπάρχει μια συλλογή από άρθρα σχετικά με τον αθλητισμό. Ο χρήστης ενδιαφέρεται για εκείνα που είναι για το ποδόσφαιρο, συγκεκριμένα το βραζιλιάνικο ποδόσφαιρο. Μετά από σχετική αναζήτηση, επιστρέφονται αρκετά άρθρα για το αμερικανικό όμως ποδόσφαιρο. Για να αυξηθεί η βαθμολογία των άρθρων σχετικά με το βραζιλιάνικο ποδόσφαιρο, η ερώτηση μπορεί να δοθεί ως εξής :

`'soccer,Brazil*3'`

Ο πίνακας 4.2 δείχνει πώς λειτουργεί ο τελεστής βαρύτητας για τρία υποθετικά έγγραφα A, B, Γ, που περιέχουν όλα πληροφορίες για ποδόσφαιρο. Οι στήλες του πίνακα δείχνουν το σκορ για τέσσερις διαφορετικές ερωτήσεις.

	soccer	Brazil	soccer,Brazil	soccer,Brazil*3
A	20	10	30	50
B	10	30	40	100
Γ	50	10	60	80

Πίνακας 4.2: Παράδειγμα λειτουργίας του τελεστή βαρύτητας

Το σκορ στην τρίτη στήλη που χρησιμοποιείται ο τελεστής συσσώρευσης είναι το άθροισμα των σκορ των δύο πρώτων στηλών. Το σκορ της τέταρτης στήλης είναι το άθροισμα του σκορ της πρώτης στήλης και της δεύτερης επί τρία. Με την αρχική ερώτηση (soccer,Brazil) η σειρά διαβάθμισης των εγγράφων είναι Γ Β Α, ενώ με την ερώτηση με το συντελεστή βαρύτητας (soccer,Brazil*3) η σειρά είναι η Β Γ Α, που είναι και η σωστή.

4.7.5 Τελεστές ελέγχου του συνόλου αποτελεσμάτων

Οι τελεστές αυτοί χρησιμοποιούνται για να ελέγχουν ποια έγγραφα θα επιστραφούν από το σύνολο των αποτελεσμάτων. Οι τελεστές για αυτούς τους τελεστές μπορεί να είναι εκφράσεις, που μπορεί να είναι από μόνες τους απλές ερωτήσεις ή συνδυασμός ερωτήσεων με τελεστές.

Λόγω του γεγονότος ότι οι τελεστές αυτοί χειρίζονται ένα σύνολο αποτελεσμάτων, πρέπει να εφαρμόζονται στο πιο εξωτερικό επίπεδο της έκφρασης της ερώτησης.

Οι τελεστές αυτοί χρησιμοποιούνται για να αποκλείουν «θόρυβο» από τη λίστα επιτυχιών, που είναι τα μη σχετιζόμενα έγγραφα, και για την ανάκτηση εγγράφων με περισσότερη ευελιξία. Υπάρχουν τρεις τελεστές :

- **THRESHOLD**, που συντάσσεται ως εξής : *έκφραση* > n, που επιστρέφει τα έγγραφα των οποίων το σκορ είναι μεγαλύτερο ενός αριθμού n, που ονομάζεται κατώφλι (*threshold*) και *όρος* > n, όπου μέσα σε μια έκφραση, επιστρέφει τα έγγραφα που περιέχουν τον *όρο* με σκορ τουλάχιστον n.

- MAX, που συντάσσεται ως εξής : *έκφραση:n*, για το οποίο επιστρέφονται τα n πρώτα έγγραφα από τη λίστα επιτυχιών (αρχίζοντας από αυτά με το μεγαλύτερο σκορ). Το n πρέπει να έχει τιμή μεταξύ 1 και 65.536.
- FIRST/NEXT, που συντάσσεται *έκφραση#m-n*, που επιστρέφει τα έγγραφα που αριθμούνται m έως n στη λίστα επιτυχιών. Αυτή η αρίθμηση δεν γίνεται βάσει του σκορ, αλλά της σειράς με την οποία το ConText βρίσκει τα έγγραφα στη στήλη κειμένου.

4.7.5.1 Παραδείγματα

Για να επιστραφούν τα έγγραφα με σκορ μεγαλύτερο του 30 για τον όρο 'lion' :

```
lion > 30
```

Για να επιστραφούν τα πρώτα 25 έγγραφα από τη λίστα επιτυχιών για τον όρο 'dance' :

```
dance:25
```

Για να επιστραφούν τα πρώτα 10 έγγραφα από τη λίστα επιτυχιών για τον όρο 'dog' :

```
dog#1-10
```

Για να επιστραφούν τα επόμενα 20 έγγραφα :

```
dog#11-30
```

Φυσικά, μπορεί να υπάρξει συνδυασμός των παραπάνω. Για να ανακληθούν τα πρώτα 10 έγγραφα από τα πρώτα 50 που περιέχουν τον όρο 'cat' :

```
cat:50#1-10
```

4.7.6 Τελεστές επαύξεσης

Οι τελεστές αυτού του είδους χρησιμοποιούνται για την επαύξηση του όρου της ερώτησης ώστε να περιλαμβάνουν παραλλαγές του όρου που δίνει ο χρήστης. Υπάρχουν τρεις τέτοιοι τελεστές, ο τελεστής STEM, SOUNDEX και FUZZY.

Οι τελεστές επαύξεσης είναι μοναδιαίοι τελεστές. Αυτό σημαίνει ότι εφαρμόζονται μόνο σε ένα όρο. Μπορούν να συνδυαστούν μεταξύ τους και με άλλους τελεστές που περιγράφονται παραπάνω. Μια αναζήτηση μπορεί να γίνει πιο ευρεία, αν εφαρμοστεί επαύξηση πάνω στην επαύξηση.

4.7.6.1 Τελεστής STEM

Συντάσσεται ως : \$*όρος*

Ο τελεστής αυτός επαυξάνει τον όρο της ερώτησης ώστε να περιλαμβάνει όλους τους όρους που έχουν την ίδια ρίζα stem με τον αρχικό όρο. Ο λεκτικός αναλυτής του ConText υποστηρίζει αγγλικά, γαλλικά, ισπανικά, ιταλικά, γερμανικά και ολλανδικά. Για παράδειγμα, το \$scream θα επαυξηθεί στο scream, screaming, screamed, κτλ.

4.7.6.2 Τελεστής SOUNDEX

Συντάσσεται ως : !*όρος*

Ο τελεστής SOUNDEX επιτρέπει αναζητήσεις για λέξεις που έχουν παρόμοιο ήχο με τον όρο που δίνει ο χρήστης. Αυτό επιτρέπει τη σύγκριση λέξεων που έχουν διαφορετική ορθογραφία, αλλά «ακούγονται» το ίδιο στα Αγγλικά.

Ο τελεστής χρησιμοποιεί την ίδια λογική με τη συνάρτηση SOUNDEX της SQL για αναζήτηση παρόμοιων «ηχητικά» λέξεων.

Για παράδειγμα, η επόμενη SQL δήλωση είναι μια ερώτηση ενός βήματος :


```

SELECT ID, COMMENT FROM EMP_RESUME
      WHERE CONTAINS(COMMENT, '!SMYTHE') > 0
(Αποτέλεσμα)
ID COMMENT
--
23 Smith is a hard worker who...

```

Ο τελεστής δουλεύει πιο αποτελεσματικά για γλώσσες που χρησιμοποιούν 7-bit σετ χαρακτήρων, όπως τα αγγλικά. Μπορεί να χρησιμοποιηθεί με λιγότερη αποτελεσματικότητα για γλώσσες που χρησιμοποιούν 8-bit σετ χαρακτήρων, όπως πολλές δυτικο-ευρωπαϊκές γλώσσες.

4.7.6.3 Τελεστής FUZZY

Συντάσσεται ως : ?*όρος*

Οι επαυξήσεις fuzzy έχουν σκοπό την επαύξηση λέξεων με παρόμοια ορθογραφία. Αυτός ο τελεστής είναι χρήσιμος για περιπτώσεις που υπάρχουν συχνά ορθογραφικά λάθη στα έγγραφα της βάσης. Ανάλογα με την επαύξηση του τελεστή STEM, ο αριθμός των λέξεων που παράγει η επαύξηση του τελεστή FUZZY εξαρτώνται από το τι υπάρχει στο ευρετήριο του κειμένου.

Παράδειγμα, το ?cat επαυξάνεται σε cat, cats, calc, case, κτλ.

Όπως και ο τελεστής SOUNDEX, η fuzzy επαύξηση δουλεύει καλύτερα για γλώσσες με 7-bit σετ χαρακτήρων. Για γλώσσες με 8-bit σετ, όπως οι δυτικο-ευρωπαϊκές, δουλεύει με μικρότερη αποτελεσματικότητα. Για τα ιαπωνικά, υπάρχουν περιορισμένες δυνατότητες fuzzy επαύξησης.

Για τους SOUNDEX και FUZZY τελεστές και εάν έχει καθοριστεί να γίνεται μετατροπή των χαρακτήρων στη βασική τους αναπαράσταση, οι τελεστές εφαρμόζονται στους όρους μετά την μετατροπή. Ο STEM τελεστής δεν υποστηρίζει τη μετατροπή στους βασικούς χαρακτήρες.

4.7.7 Τελεστές θησαυρού

Οι τελεστές θησαυρού επαυξάνουν έναν όρο χρησιμοποιώντας θησαυρό που καθορίζει σχέσεις μεταξύ του όρου που εισάγει ο χρήστης και άλλων σχετιζόμενων νοηματικών όρων.

Υπάρχουν δέκα τελεστές, που αντιστοιχούν στα δέκα είδη σχέσεων που ορίζονται από το πρότυπο ISO2788 για τους θησαυρούς. Στον πίνακα 4.3 περιγράφεται η σύνταξη και η σημασία των τελεστών αυτών. Ο *όρος* μπορεί να είναι είτε απλή λέξη είτε φράση.

Οι τελεστές υλοποιούνται στο ConText ως συναρτήσεις PL/SQL και για αυτό το λόγο έχουν ορίσματα που πρέπει να δοθούν μαζί με τον τελεστή. Οι κανόνες που ισχύουν για την PL/SQL φυσικά ισχύουν και για τους τελεστές αυτούς. Οι παράμετροι που δέχονται είναι οι εξής :

- *Όρος*. Καθορίζει τον τελεστέο, τον όρο δηλαδή στον οποίο θα ενεργήσει ο τελεστής. Για τον τελεστή NT, πρέπει να οριστεί όρος. Στους τελεστές PT και TT, ο όρος αντικαθίσταται από τους αντίστοιχους όρους που ορίζουν οι τελεστές, αν ωστόσο δεν βρεθεί τέτοιος όρος, χρησιμοποιείται ο αρχικός όρος.
- *Επίπεδο*. Καθορίζει τον αριθμό επιπέδων στην ιεραρχία του θησαυρού που θα προσπελαστούν για να επιστραφεί ένας ευρύτερος (τελεστές BT, BTG, BTP) ή στενότερος (τελεστές NT, NTG, NTP) όρος για τον όρο που δίνεται. Για παράδειγμα, επίπεδο 1 για μια BT ερώτηση επιστρέφει μόνο τον ευρύτερο όρο, αν υπάρχει, για τον όρο που δίνεται. Επίπεδο 2 θα επέστρεφε τον ευρύτερο όρο για τον ευρύτερο

όρο, αν υπάρχει, του όρου που δίνεται. Αυτή η παράμετρος είναι προαιρετική και έχει εξ' ορισμού τιμή 1. Για τιμές κάτω του μηδενός, επιστρέφεται ο αρχικός όρος.

- **Θησαυρός.** Καθορίζει το όνομα του θησαυρού που χρησιμοποιείται για την επαύξηση του όρου που δίνεται. Είναι προαιρετική παράμετρος και έχει εξ' ορισμού τιμή το 'DEFAULT'. Ως αποτέλεσμα, ένας θησαυρός με όνομα DEFAULT πρέπει να υπάρχει στους πίνακες θησαυρού προκειμένου να χρησιμοποιηθούν οι τελεστές θησαυρού.

Τελεστής	Σύνταξη	Περιγραφή
SYNONYM	SYN(όρος[,θησ.])	Επαυξάνει τον όρο ώστε να συμπεριλαμβάνει και όρους που θεωρούνται συνώνυμά του.
PREFERRED	PT(όρος[,θησ.])	Αντικαθιστά τον όρο της ερώτησης με τον προτεινόμενο του όρο. Δεν επιστρέφονται έγγραφα που περιέχουν τον όρο.
RELATED	RT(όρος[,θησ.])	Επαυξάνει τον όρο ώστε να συμπεριλαμβάνει και όρους που θεωρούνται σχετιζόμενοι με τον όρο.
TOP	TT(όρος[,θησ.])	Αντικαθιστά τον όρο με τον όρο στην κορυφή της ιεραρχίας (BT, NT). Δεν επιστρέφονται έγγραφα που περιέχουν τον όρο.
NARROWER	NT(όρος[,επίπεδο[,θησ.]])	Επαυξάνει τον όρο ώστε να συμπεριλαμβάνει όλους τους όρους μέχρι και σε επίπεδο που θεωρούνται να είναι στενότεροι του όρου.
NARROWER GENERIC	NTG(όρος[,επίπεδο[,θησ.]])	
NARROWER PARTITIVE	NTP(όρος[,επίπεδο[,θησ.]])	
NARROWER INSTANCE	NTI(όρος[,επίπεδο[,θησ.]])	
BROADER	BT(όρος[,επίπεδο[,θησ.]])	Επαυξάνει τον όρο ώστε να συμπεριλαμβάνει όλους τους όρους μέχρι και σε επίπεδο που θεωρούνται να είναι ευρύτεροι του όρου.
BROADER GENERIC	BTG(όρος[,επίπεδο[,θησ.]])	
BROADER PARTITIVE	BTP(όρος[,επίπεδο[,θησ.]])	
BROADER INSTANCE	BTI(όρος[,επίπεδο[,θησ.]])	

Πίνακας 4.3: Τελεστές θησαυρού

Εσωτερικά, το ConText επεξεργάζεται την επαύξηση βάζοντας σε αγκύλη κάθε ξεχωριστό όρο που επιστρέφεται από αυτήν και στο τέλος όλοι οι όροι συσσωρεύονται χρησιμοποιώντας τον τελεστή ACCUMULATE. Για παράδειγμα, αν οι λέξεις bird, birdy και avian είναι συνώνυμες, το SYN(bird) επαυξάνεται στο {bird}, {avian}, {birdy}.

Αν ένας όρος δεν έχει εγγραφή στο θησαυρό, δεν γίνεται επαύξηση και ο ίδιος όρος χρησιμοποιείται στην ερώτηση.

4.7.7.1 Περιορισμοί

Οι τελεστές θησαυρού μπορούν να χρησιμοποιηθούν και με άλλους τελεστές, με την εξαίρεση του τελεστή NEAR.

Το μέγιστο μήκος μιας επαυξημένης ερώτησης είναι 32.000 χαρακτήρες.

Οι τελεστές δεν μπορούν να φωλιαστούν. Για παράδειγμα, η έκφραση :

SYN(BT(bird))

δεν επιτρέπεται.

Οι επαυξήσεις θησαυρού στις ερωτήσεις κειμένου μπορούν να κάνουν διάκριση μεταξύ κεφαλαίων-πεζών. Για παράδειγμα, δημιουργείται ένας θησαυρός με όνομα *thes1* και η λέξη Mercury ορίζεται ως στενότερος όρος για τον όρο planets, ενώ ο όρος mercury ορίζεται ως στενότερος όρος για την metals. Σε μια ερώτηση, συμβαίνουν οι ακόλουθες επαυξήσεις :

BT (mercury,1,thes1) επαυξάνεται στο {MERCURY},{METALS}

BT (Mercury,1,thes1) επαυξάνεται στο {MERCURY},{PLANETS}

Δεν υπάρχει δυνατότητα να ενεργοποιηθεί ή απενεργοποιηθεί η διάκριση κεφαλαίων-πεζών. Το ConText διατηρεί τους όρους όπως είναι όταν εισάγονται στο θησαυρό βασιζόμενο στο αν ο θησαυρός είχε καθοριστεί να κάνει διάκριση κεφαλαίων-πεζών κατά τη δημιουργία του. Με παρόμοιο τρόπο, οι ερωτήσεις χρησιμοποιούν τους όρους όπως δίνονται από το χρήστη για να γίνει η αναζήτηση στο θησαυρό.

Οι θησαυροί που κάνουν διάκριση κεφαλαίων-πεζών επηρεάζουν την επαύξηση των όρων και όχι τους όρους που χρησιμοποιούνται στην ερώτηση. Το αν θα υπάρξει διάκριση κεφαλαίων-πεζών στους επαυξημένους όρους εξαρτάται από το αν το ευρετήριο έχει καθοριστεί να κάνει κάτι τέτοιο. Για παράδειγμα, για θησαυρό *thes1* με διάκριση κεφαλαίων-πεζών χρησιμοποιείται με ευρετήριο χωρίς διάκριση, θα γινόταν η ακόλουθη επαύξηση :

BT (Mercury,1,thes1) επαυξάνεται στο {MERCURY},{PLANETS}

Η ερώτηση θα επέστρεφε όλα τα έγγραφα στα οποία εμφανίζονται οι δύο όροι, άσχετα με το αν είναι στα κεφαλαία ή πεζά. Αν χρησιμοποιούνταν ευρετήριο με διάκριση, τότε η παραπάνω επαύξηση θα γινόταν ως εξής :

BT (Mercury,1,thes1) επαυξάνεται στο {Mercury},{planets}

το οποίο θα επιστρέφει τα έγγραφα που περιέχουν τα Mercury και planets.

4.7.8 Χαρακτήρες μάσκες (wildcard characters)

Οι χαρακτήρες μάσκες χρησιμοποιούνται για την επαύξηση των αναζητήσεων λέξεων σε αναζητήσεις *σχεδίων (patterns)*. Για το σκοπό αυτό, χρησιμοποιούνται δύο χαρακτήρες :

- %, που αντικαθιστά μηδέν ή περισσότερους οποιουδήποτε χαρακτήρες, και
- _, αντικαθιστά έναν οποιοδήποτε χαρακτήρα

4.7.9 Χαρακτήρες καθορισμού γκρουπ

Οι χαρακτήρες αυτοί ελέγχουν την προτεραιότητα των εκφράσεων μιας ερώτησης με το να ομαδοποιούν τους όρους και τους τελεστές. Οι χαρακτήρες που χρησιμοποιούνται είναι οι παρενθέσεις '()' και οι αγκύλες '[']'.

4.7.10 Αποθηκευμένες ερωτήσεις

Τα αποτελέσματα μιας ερώτησης μπορούν να αποθηκευτούν και να ανακληθούν σε μεταγενέστερο χρόνο. Για το σκοπό αυτό υπάρχει ο τελεστής SQE, που δέχεται ως παράμετρο το όνομα μιας *αποθηκευμένης ερώτησης (stored query)*.

Το πλεονέκτημα των αποθηκευμένων ερωτήσεων είναι ότι σε γενικές γραμμές είναι πιο γρήγορες από μια ερώτηση στην οποία καθορίζονται οι όροι αναζήτησης, διότι το

ConText δεν έχει να προσπελάσει τους πίνακες κειμένου απευθείας. Επίσης, οι αποθηκευμένες ερωτήσεις μπορούν να χρησιμοποιηθούν για επαναληπτικές ερωτήσεις, όπου μια ερώτηση τροποποιείται (*refine*) χρησιμοποιώντας μια ή περισσότερες πρόσθετες ερωτήσεις.

Η διαδικασία για τη χρησιμοποίηση των ερωτήσεων αυτών αποτελείται από δύο στάδια. Στο πρώτο, καλείται η διαδικασία `CTX_QUERY.STORE_SQE` που αποθηκεύει τα αποτελέσματα μιας ερώτησης. Οι παράμετροι που δέχεται η διαδικασία αυτή είναι το όνομα της αποθηκευμένης ερώτησης, το όνομα μιας πολιτικής που αναγνωρίζει τη στήλη κειμένου για την αποθηκευμένη ερώτηση, η ερώτηση για τη στήλη κειμένου και τέλος μια παράμετρος που υποδεικνύει αν η αποθηκευμένη ερώτηση θα είναι του συστήματος ή αν θα ισχύει μόνο για την τρέχουσα σύννοδο του χρήστη. Στο δεύτερο στάδιο καλείται η αποθηκευμένη ερώτηση μέσω του τελεστή `SQE` σε μια απλή ερώτηση κειμένου ή θέματος. Το ConText επιστρέφει τα αποτελέσματα της αποθηκευμένης ερώτησης με τον ίδιο τρόπο όπως και με τα απλές ερωτήσεις. Αν τα αποτελέσματα της αποθηκευμένης ερώτησης δεν ισχύουν πλέον, το ConText αυτόματα αναθεωρεί την ερώτηση πριν επιστρέψει τα αποτελέσματα.

Για παράδειγμα, η δημιουργία μιας αποθηκευμένης ερώτησης με όνομα `PROG_LANG` για την τρέχουσα σύννοδο του χρήστη γίνεται ως εξής :

```
exec ctx_query.store_sqe('emp_resumes','prog_lang','cobol','session');
```

Η παραπάνω δήλωση προσπελαύνει τη στήλη που έχει καθοριστεί με την πολιτική `EMP_RESUMES` και επιστρέφει όλα τα έγγραφα που περιέχουν τη λέξη `cobol`. Τα αποτελέσματα αποθηκεύονται σε ένα ειδικό πίνακα που προσπελάσσεται μέσω του τελεστή `SQE`. Στη συνέχεια, η ερώτηση `PROG_LANG` μπορεί να κληθεί σε μια απλή ερώτηση ως εξής :

```
select score, docid from emp
where contains(resume, 'sqe(prog_lang)')>0
order by score;
```

4.7.11 Χαρακτήρες διαφυγής

Όπως έχει διαπιστώσει κανείς υπάρχουν αρκετές δεσμευμένες λέξεις, οι οποίες αν χρησιμοποιηθούν όπως είναι στις ερωτήσεις, το ConText κατά την συντακτική ανάλυση (*parsing*) της ερώτησης, θα λάβει αυτές τις λέξεις με την έννοια της δεσμευμένης λέξης κλειδί. Αν όμως ο χρήστης θέλει να χρησιμοποιήσει αυτού του είδους τις λέξεις ως απλές λέξεις κι όχι ως τελεστές, πρέπει με κάποιο τρόπο να άρει τον μετα-χαρακτηρικό χαρακτήρα αυτών των λέξεων. Αυτό γίνεται με τους χαρακτήρες διαφυγής, οι οποίοι είναι οι αγκύλες '{ }' και η ανάποδη κάθετος '\'. Οι αγκύλες περικλείουν την έκφραση της οποίας ο χρήστης θέλει να άρει το μετα-χαρακτηρικό χαρακτήρα, ενώ η ανάποδη κάθετος χρησιμοποιείται για τη διαφυγή του αμέσως επόμενου χαρακτήρα.

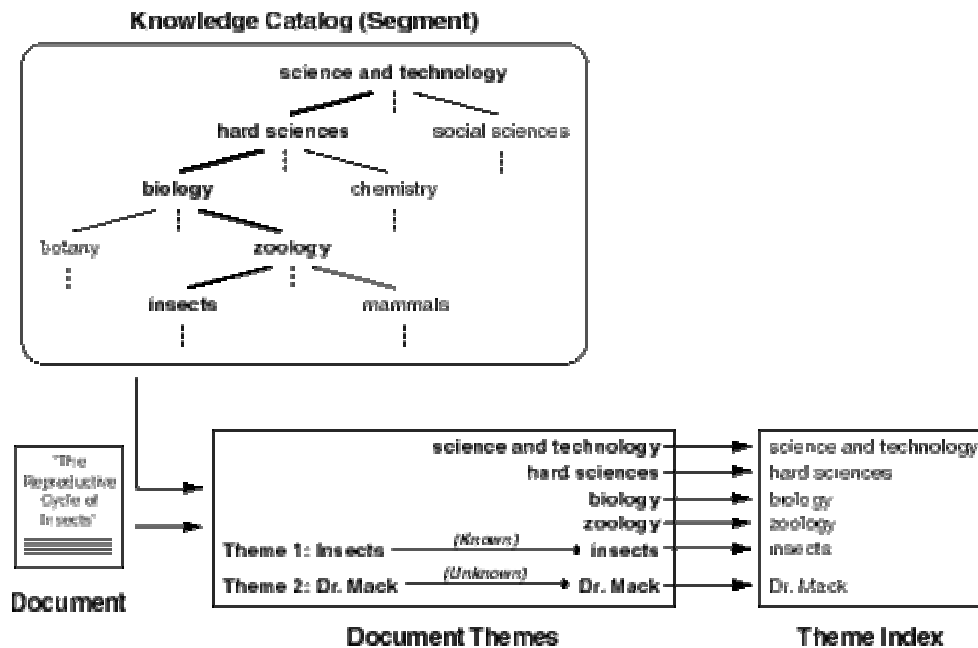
4.8 Ερωτήσεις θεμάτων

Οι ερωτήσεις κειμένου επιτρέπουν την αναζήτηση εγγράφων με βάση την κύρια ιδέα τους.

4.8.1 Αρχές ευρετηριοποίησης

Πριν ο χρήστης δώσει μια τέτοια ερώτηση προς εκτέλεση, πρέπει πρώτα να ευρετηριοποιηθεί ένα σύνολο εγγράφων. Κατά τη διάρκεια της ευρετηριοποίησης, το ConText εξάγει μέχρι και πενήντα κεντρικές έννοιες ή αλλιώς *θέματα* (*themes*) τα οποία θα αποθη-

κευθούν στο ευρετήριο θεμάτων. Ένα βάρος επίσης συσχετίζεται με κάθε θέμα που ευρετηριοποιείται. Ένα θέμα μπορεί να είναι μια συγκεκριμένη έννοια, όπως για παράδειγμα *έντομα*, ή μια αφηρημένη, για παράδειγμα *επιτυχία*, έννοια η οποία αναπτύσσεται εκτενώς στο έγγραφο. Το σχήμα 4.3 δείχνει πως το ConText χρησιμοποιεί τη γνωστική του βάση



Σχήμα 4.3: Παράδειγμα ερώτησης θέματος

για να εξάγει τα θέματα του εγγράφου «The reproductive cycle of insects», το οποίο έχει πληροφορίες σχετικά με τα έντομα. Αυτό το παράδειγμα δείχνει ότι το ConText αναγνωρίζει δύο είδη θεμάτων, τα γνωστά και τα άγνωστα θέματα.

4.8.2 Γνωστά θέματα

Γνωστά είναι τα θέματα τα οποία μπορούν να τοποθετηθούν σε ένα κλαδί της γνωστικής βάσης του ConText. Στο σχήμα 4.3, το έγγραφο με τίτλο «The reproductive cycle of insects» έχει τέσσερα γονικά θέματα που αντιστοιχούν στα εξής κλαδιά της γνωστικής βάσης : *science and technology*, *hard sciences*, *biology*, *zoology* και *insects*. Κάθε θέμα εισάγεται ως μια γραμμή προς αναζήτηση στο ευρετήριο παράλληλα με ένα βάρος. Όταν δοθεί μια ερώτηση σχετική με έντομα ή οποιοδήποτε από τα γονικά θέματα που έχουν ευρετηριοποιηθεί, τότε επιστρέφεται το έγγραφο «The reproductive cycle of insects».

4.8.3 Άγνωστα θέματα

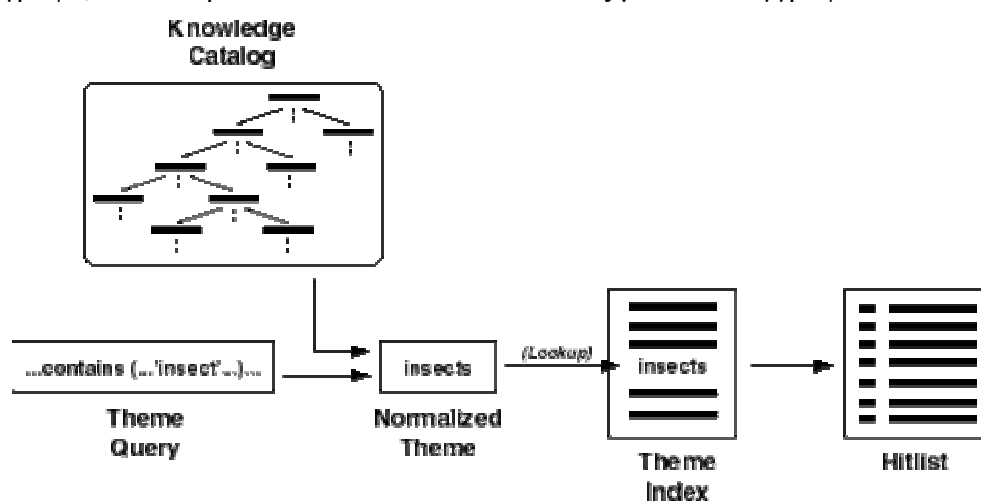
Άγνωστα είναι τα θέματα τα οποία δεν μπορούν να βρεθούν στη γνωστική βάση, διότι είτε επειδή είναι άγνωστα σε αυτή ή από τη φύση τους είναι ασαφή. Το σχήμα 4.3 δείχνει ότι έχει εξαχθεί ένα άγνωστο θέμα (Dr. Mack) χωρίς αντιστοίχιση με κάποιο κλαδί της γνωστικής βάσης. Τέτοια άγνωστα θέματα εισάγονται στο ευρετήριο ως μία εγγραφή. Προκειμένου να γίνει αναζήτηση σε τέτοια ασαφή θέματα, ο χρήστης πρέπει να βασίζεται σε γνωστά θέματα που έχουν ευρετηριοποιηθεί μαζί με τα ασαφή.

4.8.4 Βάρος θέματος

Το βάρος θέματος είναι το μέτρο σύγκρισης της «ισχύος» του θέματος σε σχέση με τα άλλα θέματα του εγγράφου. Τα βάρη ευρετηριοποιούνται μαζί με κάθε θέμα και τα γονικά του. Το ConText χρησιμοποιεί τα βάρη για να βαθμολογήσει τις ερωτήσεις θεμάτων.

4.8.5 Διαδικασία ερώτησης

Για να εκτελεστεί μια ερώτηση θεμάτων, ο χρήστης καθορίζει μια συμβολοσειρά που μπορεί να είναι πρόταση με ή χωρίς τελεστές. Το ConText χρησιμοποιεί τη γνωστική βάση για να κανονικοποιήσει αυτή την πρόταση σε μια στάνταρ μορφή. Στη συνέχεια, ψάχνει στο ευρετήριο για το κανονικοποιημένο θέμα και επιστρέφει τα έγγραφα που είχαν ευρετηριοποιηθεί με το δοθέν θέμα. Το σχήμα 4.4 παρουσιάζει αυτή τη διαδικασία. Τα σκορ για τις ερωτήσεις θεμάτων υπολογίζονται βάσει των βαρών που είναι συσχετισμένα με κάθε θέμα στο ευρετήριο. Προκειμένου όμως να αντιστοιχίσει το ConText ένα θέμα με ένα έγγραφο, εκείνο πρέπει να αναπτύσσεται εκτενώς μέσα στο έγγραφο.



Σχήμα 4.4: Διαδικασία ερώτησης θέματος

Για παράδειγμα, βάσει του παραδείγματος που αναφέρθηκε παραπάνω, ερώτηση για τα έντομα (*insects*) θα επιστρέψει το έγγραφο «The reproductive cycle of insects». Το ίδιο έγγραφο θα επιστραφεί αν γίνει ερώτηση που να αναφέρεται στα γονικά θέματα *science and technology*, *hard sciences*, *biology* και *zoology*.

4.8.6 Υπολογισμός σκορ

Μετά από ερώτηση, το ConText επιστρέφει με κάθε έγγραφο και ένα σκορ που φανερώνει το βαθμό σχετικότητας του εγγράφου με την ερώτηση. Όσο πιο μεγάλο είναι το σκορ, τόσο πιο σχετιζόμενο είναι το έγγραφο. Μέγιστο σκορ είναι το 100 και βασίζεται στο βάρος του ευρετηριοποιημένου θέματος.

Σε γενικές γραμμές, όταν καθορίζεται ευρύ θέμα σε μια ερώτηση, επιστρέφονται και υψηλότερα σκορ. Όταν χρησιμοποιούνται τελεστές σε ερώτηση θέματος, ο υπολογισμός του σκορ είναι παρόμοιος με τις κανονικές ερωτήσεις κειμένου. Για παράδειγμα, ερώτηση με τον λογικό τελεστή OR θα επιστρέψει το υψηλότερο σκορ των τελεστέων του, ενώ το AND θα επιστρέψει το μικρότερο.

4.8.7 Διάκριση κεφαλαίων-πεζών

Οι ερωτήσεις θεμάτων κάνουν διάκριση μεταξύ κεφαλαίων-πεζών. Για παράδειγμα, ερώτηση για το ουσιαστικό turkey (γαλοπούλα) θα επιστρέφει έγγραφο σχετικά με το πουλί. Τέτοια ερώτηση δεν θα επιστρέφει έγγραφο σχετικά με το κύριο όνομα Τουρκία.

Το ConText είναι ανεκτικό με είσοδο που έχει φτωχή μορφοποίηση για γνωστά θέματα. Για παράδειγμα, για την ερώτηση microSoft ή microsoft, το ConText επιστρέφει έγγραφα που περιέχουν τον όρο Microsoft.

Το ConText πάντα προσπαθεί να προσαρμόσει το θέμα που εισάγει ο χρήστης με θέματα του ευρετηρίου.

4.8.8 Δημιουργία ερωτήσεων θεμάτων

Για τις ερωτήσεις θεμάτων, οι ακόλουθοι τελεστές έχουν την ίδια σημασία όπως και στις κανονικές ερωτήσεις κειμένου. Οι τελεστές αυτοί είναι : ο τελεστής συσώρευσης (''), το λογικό ΕΙΤΕ ('|'), το λογικό ΚΑΙ ('&'), η διαφορά ('-'), η άρνηση ('~'), ο τελεστής βάρους ('*'), το κατώφλι ('>') και ο τελεστής μέγιστου (':').

Επίσης, με τον ίδιο τρόπο λειτουργούν και οι τελεστές θησαυρού, οι χαρακτήρες καθορισμού γκρουπ και οι χαρακτήρες μάσκες.

Οι τελεστές που δεν υποστηρίζονται είναι οι τελεστές NEAR, FUZZY, SOUNDEX και STEM.

Όταν δίνεται μια ερώτηση θέματος, το ConText κανονικοποιεί τη λέξη ή φράση με τέτοιο τρόπο ώστε η ερώτηση να έρθει σε μια μορφή που να μπορεί να συγκριθεί με άλλα θέματα εγγράφων στο ευρετήριο. Αυτή η κανονικοποιημένη μορφή αποτελείται από ουσιαστικά και φράσεις ουσιαστικών, όπως *chemistry* και *personal computer*. Είναι λοιπόν προτιμότερο ο χρήστης να χρησιμοποιεί ουσιαστικά και φράσεις ουσιαστικών για την υποβολή ερωτήσεων θεμάτων. Δεν θα πρέπει να χρησιμοποιούνται προτάσεις και μεγάλες φράσεις. Δεν πρέπει επίσης να χωρίζονται οι φράσεις που περιγράφουν την ιδέα ως σύνολο, για παράδειγμα, αντί της φράσης *physical and chemistry*, είναι καλύτερο να χρησιμοποιηθεί η φράση *physical chemistry*.

4.8.9 Τροποποίηση των ερωτήσεων (refining queries)

Ανάλογα με το πώς έχει δοθεί η ερώτηση, το ConText συνήθως επιστρέφει τόσο έγγραφα που είναι σχετιζόμενα με αυτήν όσο και μη σχετιζόμενα έγγραφα. Όπως είναι φυσικό, ο χρήστης δεν είναι δυνατό να γνωρίζει από πριν τα αποτελέσματα και ο καλύτερος τρόπος να το μάθει είναι να υποβάλλει την ερώτηση. Στη συνέχεια, μπορεί να εξετάσει το σύνολο των εγγράφων που λαμβάνει ως αποτέλεσμα και να ξαναυποβάλλει την ερώτηση χρησιμοποιώντας αυτή τη φορά τελεστές προκειμένου να εξαλείψει τα ανεπιθύμητα έγγραφα. Αυτή η μέθοδος μπορεί να επιτευχθεί με δύο τρόπους :

- Περιορισμός της ερώτησης. Ο χρήστης επιλέγει μια ευρύτερη κατηγορία ή θέμα, εξετάζει τα αποτελέσματα και υποβάλλει ξανά την ερώτηση χρησιμοποιώντας τους τελεστές AND ή NOT για να αποκλείσει από τη λίστα επιτυχίων τα ανεπιθύμητα έγγραφα.
- Επαύξηση της ερώτησης. Ο χρήστης επιλέγει μια συγκεκριμένη κατηγορία, εξετάζει το αποτέλεσμα και στη συνέχεια επαυξάνει την ερώτηση ώστε να συμπεριληφθούν περισσότερα έγγραφα στη λίστα επιτυχίων.

4.8.9.1 Περιορισμός της ερώτησης

Αρχίζοντας με ευρύ θέματα, οι ερωτήσεις είναι δυνατό να επιστρέψουν «θόρυβο» ή ανεπιθύμητα έγγραφα. Αυτό συμβαίνει εξαιτίας των γεγονότων ότι η λέξη ή φράση που χρησιμοποιείται στην ερώτηση μπορεί να αντιπροσωπεύει περισσότερες από μία ιδέες και ότι ένα έγγραφο είναι πολύ πιθανό να του έχουν αντιστοιχηθεί περισσότερα του ενός θέματα. Μπορούν να χρησιμοποιηθούν οι τελεστές AND ή NOT για τον αποκλεισμό των μη σχετιζόμενων εγγράφων, ωστόσο το γεγονός αυτό απαιτεί προσοχή, προκειμένου να μην αποκλειστούν σχετιζόμενα έγγραφα. Είναι καλύτερα να υπάρχει πάντα ένα ποσοστό «θόρυβου» παρά να μην υπάρχει κανένα αποτέλεσμα.

Για παράδειγμα, αν ερώτηση με τη λέξη *cricket* επιστρέψει έγγραφα που αναφέρονται τόσο στο άθλημα, όσο και στο έντομο (τριζόνι), ο χρήστης θα μπορούσε να χρησιμοποιήσει τον τελεστή AND προκειμένου να περιορίσει τα έγγραφα σε αυτά που αναφέρονται μόνο στο άθλημα, με την εξής ερώτηση :

```
cricket and sports
```

Το μειονέκτημα χρησιμοποίησης του τελεστή είναι ότι για να επιστρέψει σωστά αποτελέσματα η ερώτηση, πρέπει τα έγγραφα να αναπτύσσουν εκτενώς και τις δύο έννοιες (*cricket*, *sports*). Μια άλλη προσέγγιση θα ήταν να χρησιμοποιηθεί ο τελεστής NOT. Σε αυτή την περίπτωση, ο τελεστής θα μπορούσε να αποκλείσει την ανεπιθύμητη έννοια, για το παραπάνω παράδειγμα, ο χρήστης θα μπορούσε να υποβάλει την εξής ερώτηση

```
cricket not insects
```

προκειμένου να πάρει έγγραφα που σχετίζονται με το άθλημα *cricket* και όχι το έντομο. Ένα μειονέκτημα του τελεστή NOT είναι ότι δε θα επιστραφούν έγγραφα που συμβαίνει να αναφέρονται και στο θέμα που ενδιαφέρει το χρήστη, αλλά και στο θέμα που απέκλεισε με τον τελεστή. Ένα άλλο εγγενές μειονέκτημα, είναι ότι σχεδόν πάντα, ο χρήστης έχει καλύτερη γνώση των θεμάτων για τα οποία ενδιαφέρεται και όχι αυτών τα οποία δεν επιθυμεί. Για αυτό το λόγο, ο τελεστής NOT ταιριάζει καλύτερα για τον αποκλεισμό μη σχετιζόμενων εγγράφων, τα οποία είναι όμως ψηλά στη σειρά διαβάθμισης της λίστας επιτυχιών και των οποίων το θέμα είναι γνωστό ότι είναι άσχετο.

4.8.9.2 Επαύξηση της ερώτησης

Μερικές φορές είναι καλύτερα ο χρήστης να αρχίσει με συγκεκριμένες κατηγορίες και μετά να επαυξάνει τις ερωτήσεις σε μερικές πιο γενικές, ειδικά όταν η ερώτηση καλύπτει ένα θέμα που έχει καλή κατηγοριοποίηση. Για παράδειγμα, έστω ότι ο χρήστης ψάχνει για έγγραφα σχετικά με ‘bees’ (μέλισσες) που είναι μια συγκεκριμένη κατηγορία εντόμων. Αν η αναζήτηση δεν επιστρέψει επιθυμητά αποτελέσματα, τότε ο χρήστης μπορεί να ψάξει για ‘insects’ (έντομα) που είναι μια ελαφρώς ευρύτερη κατηγορία. Στη συνέχεια, ο χρήστης θα μπορούσε να χρησιμοποιήσει τους τελεστές AND ή NOT, όπως περιγράφεται και στην προηγούμενη παράγραφο.

4.9 Διαδικασίες και συναρτήσεις για την εκτέλεση ερωτήσεων

Το ConText με την ενεργοποίησή του εγκαθιστά στη βάση μια σειρά πακέτων⁶. Τα πακέτα αυτά χρησιμοποιούνται για την εκτέλεση ερωτήσεων και για τη δημιουργία γλωσσολογικής εξόδου, για τη διαχείριση της ουράς εξυπηρέτησης των αιτήσεων, για τη διαχείριση των διακομιστών ConText και για τη λήψη πληροφοριών σχετικά με την εγκατε-

⁶ Πακέτο (package) στην Oracle είναι μια συλλογή από διαδικασίες και συναρτήσεις μαζί με τους δρομείς (*cursors*) και μεταβλητές που χρησιμοποιούν, αποθηκευμένη στη βάση για συνεχή χρήση. Τα πακέτα γράφονται σε PL/SQL, μεταγλωττίζονται και αποθηκεύονται στη βάση για την κλήση τους.

στημένη έκδοση του ConText cartridge, για τη δημιουργία πολιτικών ή ευρετηρίων, κτλ.. Προκειμένου να γίνει ανάπτυξη εφαρμογών σε PL/SQL, ο διαχειριστής του ConText πρέπει να έχει δώσει το δικαίωμα `EXECUTE` για τα πακέτα αυτά στους ανάλογους χρήστες.

Παρακάτω, αναπτύσσονται σε λίγες γραμμές και με αλφαβητική σειρά οι συναρτήσεις και οι διαδικασίες που περιέχουν τα πακέτα `CTX_QUERY`, `CTX_LING`, `CTX_DDL` και `CTX_DML`.

4.9.1 Πακέτο CTX_QUERY

Το πακέτο αυτό περιέχει συναρτήσεις και διαδικασίες για την επεξεργασία ερωτήσεων από μνήμη και ερωτήσεων δύο βημάτων, καθώς και για τον τονισμό των εγγράφων που επιστρέφονται. Παρέχονται επίσης συναρτήσεις για τον χειρισμό των αποθηκευμένων ερωτήσεων.

4.9.1.1 CLOSE_CON

Η διαδικασία αυτή κλείνει ένα δρομέα που έχει ανοίξει η `OPEN_CON`. Χρησιμοποιείται σε συνδυασμό με την `FETCH_HIT` και `OPEN_CON` για την εκτέλεση ερωτήσεων από μνήμη. Η μοναδική παράμετρος της διαδικασίας είναι ο αριθμός του δρομέα που ανοίχτηκε με την `OPEN_CON`.

4.9.1.2 CONTAINS

Η διαδικασία `CONTAINS` επιλέγει τα έγγραφα που ικανοποιούν τη συνθήκη που καθορίζεται ως παράμετρος της και αποθηκεύει τα αποτελέσματα στον πίνακα που καθορίζεται. Οι παράμετροι που δέχεται το όνομα της πολιτικής, η ερώτηση, ο πίνακας αποτελεσμάτων, μια παράμετρος που καθορίζει αν τα αποτελέσματα θα είναι προσπελάσιμα και από άλλες δηλώσεις `CONTAINS` (όταν συμβαίνει αυτό πρέπει επίσης να δίνεται και ένας αριθμός ταυτότητα για κάθε δήλωση `CONTAINS`), μια παράμετρος που καθορίζει τον αριθμό των ConText διακομιστών που θα εκτελέσουν την ερώτηση και τέλος μια συνθήκη `WHERE` που σχετίζεται με την ερώτηση.

4.9.1.3 COUNT_HITS

Η `COUNT_HITS` είναι συνάρτηση που επιστρέφει τον αριθμό των αποτελεσμάτων μιας ερώτησης. Δεν χρησιμοποιεί πίνακα για την αποθήκευση αποτελεσμάτων. Οι παράμετροι της συνάρτησης είναι το όνομα της πολιτικής, η ερώτηση, μία συνθήκη `WHERE` για την ερώτηση, καθώς και μια παράμετρος του τύπου αλήθειας, η οποία καθορίζει τον τρόπο λειτουργίας της συνάρτησης, αν θα επιστρέψει τα αποτελέσματα με προσέγγιση (αλλά με μεγαλύτερη ταχύτητα) ή με ακρίβεια.

4.9.1.4 COUNT_LAST

Επιστρέφει τον αριθμό των αποτελεσμάτων μετά από μια εκτέλεση της `CONTAINS` σε ερώτηση δύο βημάτων ή της `OPEN_CON` για ερώτηση από μνήμη. Μια εναλλακτική μέθοδος για την ανάκληση του αριθμού των αποτελεσμάτων είναι η εκτέλεση της ερώτησης με την `COUNT_HITS` για την ανάκληση του αριθμού και στη συνέχεια η εκτέλεση της ίδιας ερώτησης για την ανάκληση των αποτελεσμάτων.

4.9.1.5 FEEDBACK

Η διαδικασία `FEEDBACK` χρησιμοποιείται για τη δημιουργία πληροφορίας για τον τρόπο με τον οποίο το ConText αναλύει την ερώτηση πριν την εκτελέσει. Τα αποτελέσματα αποθηκεύονται στον πίνακα που ορίζεται. Η `FEEDBACK` χρησιμοποιείται για λόγους τροποποίησης και αποσφαλμάτωσης των ερωτήσεων. Οι παράμετροι που δέχεται είναι το όνομα της πολιτικής, η ερώτηση, το όνομα του πίνακα για την αποθήκευση των αποτελεσμάτων, παράμετρος που θα καθορίζει αν τα αποτελέσματα είναι προσπελάσιμα και από άλλες δηλώσεις `FEEDBACK`, καθώς και τον αντίστοιχο αριθμό ταυτότητα για την περίπτωση αυτή.

4.9.1.6 `FETCH_HIT`

Συνάρτηση για την επιστροφή ενός αποτελέσματος από τον δρομέα που έχει ανοίξει προηγουμένως με την `OPEN_CON`. Επιστρέφει 1 για την επιτυχή επιστροφή αποτελέσματος και 0 για αποτυχία. Δέχεται παραμέτρους τον αριθμό του δρομέα και επιστρέφει στις υπόλοιπες παραμέτρους το κύριο κλειδί του εγγράφου (που είναι μια στήλη κειμένου), το σκορ του εγγράφου και στις υπόλοιπες πέντε (το μέγιστο) παραμέτρους, διάφορες πληροφορίες για το έγγραφο.

4.9.1.7 `GETTAB`

Η διαδικασία αυτή κατανέμει πίνακα αποτέλεσμα για χρησιμοποίηση από την `HIGHLIGHT` και `CONTAINS`. Δέχεται τρεις παραμέτρους, από τις οποίες η πρώτη είναι αλφαριθμητικού τύπου σταθερά που φανερώνει το είδος του πίνακα που πρέπει να δημιουργηθεί για χρησιμοποίηση από τις αντίστοιχες διαδικασίες και συναρτήσεις, η δεύτερη που είναι παράμετρος εξόδου και είναι το όνομα του πίνακα που δημιουργήθηκε και η τρίτη είναι ένας αριθμός για τον αριθμό των κλειδιών κειμένου του πίνακα, εξ' ορισμού ένα.

4.9.1.8 `HIGHLIGHT`

Διαδικασία για τον τονισμό του κειμένου. Χρησιμοποιείται συνήθως μετά την εκτέλεση μιας ερώτησης για την αναγνώριση του εγγράφου που θα επεξεργαστεί. Δέχεται μια πληθώρα παραμέτρων, από τις οποίες είναι το όνομα της πολιτικής, το κλειδί κειμένου για το έγγραφο, καθώς και η ερώτηση που χρησιμοποιήθηκε αρχικά. Οι άλλες παράμετροι έχουν σχέση με το πώς θα γίνει ο τονισμός του κειμένου.

4.9.1.9 `OPEN_CON`

Χρησιμοποιείται για το άνοιγμα δρομέα για ερώτηση από μνήμη. Οι παράμετροι που δέχεται είναι το όνομα της πολιτικής, η ερώτηση, παράμετρος τύπου αλήθειας για την ταξινόμηση ή όχι των αποτελεσμάτων με βάση το σκορ, μια αλφαριθμητική παράμετρος που είναι ονόματα στηλών που θα εμφανιστούν μαζί με τον αριθμό του εγγράφου και το σκορ στη λίστα επιτυχιών, καθώς και τη συνθήκη `WHERE` για την ερώτηση. Η συνάρτηση επιστρέφει έναν αναγνωριστικό αριθμό ταυτότητα για το δρομέα που άνοιξε, που θα χρησιμοποιηθεί από τις `FETCH_HIT` και `CLOSE_CON`.

4.9.1.10 `PKDECODE`

Χρησιμοποιείται για την εξαγωγή ενός στοιχείου από ένα σύνθετο κλειδί κειμένου. Χρησιμοποιείται μετά από την `FETCH_HIT` και δέχεται ως παραμέτρους το σύνθετο κλειδί, καθώς και ένα αριθμό δείκτη που φανερώνει ποιο στοιχείο του σύνθετου κλειδιού θα εξαχθεί. Επιστρέφει το εξαγόμενο στοιχείο.

4.9.1.11 PKENCODE

Εκτελεί την αντίστροφη διαδικασία από την PKDECODE, δεχόμενη ως και δεκαέξι στοιχεία και επιστρέφει το σύνθετο κλειδί.

4.9.1.12 PURGE_SQE

Αυτή η διαδικασία καταργεί όλες τις αποθηκευμένες ερωτήσεις για την τρέχουσα σύνοδο. Η μοναδική παράμετρος που δέχεται είναι το όνομα της πολιτικής για την οποία οι αποθηκευμένες ερωτήσεις θα καταργηθούν.

4.9.1.13 REFRESH_SQE

Ξαναεκτελεί την αποθηκευμένη ερώτηση και αποθηκεύει εκ νέου τα αποτελέσματα στον πίνακα αποτελεσμάτων. Δέχεται ως παραμέτρους το όνομα της πολιτικής και το όνομα της αποθηκευμένης ερώτησης.

4.9.1.14 RELTAB

Καταργεί ένα πίνακα που δημιουργήθηκε προηγουμένως με την GETTAB. Η παράμετρος που δέχεται είναι το όνομα του πίνακα.

4.9.1.15 REMOVE_SQE

Διαγράφει μια αποθηκευμένη ερώτηση από τον πίνακα αποθηκευμένων ερωτήσεων του συστήματος, καθώς και τον αντίστοιχο πίνακα αποτελεσμάτων. Οι παράμετροι που δέχεται είναι το όνομα της πολιτικής και το όνομα της αποθηκευμένης ερώτησης.

4.9.1.16 STORE_SQE

Διαδικασία που δημιουργεί μια αποθηκευμένη ερώτηση, με την εκτέλεση της ερώτησης που δίνεται ως παράμετρος για την πολιτική που επίσης δίνεται ως παράμετρος και αποθηκεύει την ερώτηση. Το όνομα της αποθηκευμένης ερώτησης δίνεται ως παράμετρος, καθώς επίσης και μια μεταβλητή που καθορίζει αν η ερώτηση θα ισχύει για την τρέχουσα σύνοδο ή αν θα είναι καθολική του συστήματος.

4.9.2 Πακέτο CTX_LING

Το πακέτο αυτό περιέχει τις διαδικασίες και συναρτήσεις για γλωσσολογική έξοδο και για τον έλεγχο του πώς θα γίνονται αιτήσεις στους διακομιστές ConText με το ρόλο γλωσσολογικής επεξεργασίας (L).

4.9.2.1 CANCEL

Ακυρώνει όλες τις εκκρεμείς αιτήσεις για γλωσσολογική επεξεργασία. Δεν δέχεται παραμέτρους.

4.9.2.2 GET_COMPLETION_CALLBACK

Επιστρέφει το όνομα της PL/SQL διαδικασίας που καλείται μετά την περάτωση από το διακομιστή ConText μιας αίτησης στην ουρά εξυπηρέτησης και που έχει οριστεί με την SET_COMPLETION_CALLBACK για την τρέχουσα σύνοδο.

4.9.2.3 GET_ERROR_CALLBACK

Επιστρέφει το όνομα της διαδικασίας που καλείται μετά την εμφάνιση λάθους κατά τη διάρκεια εξυπηρέτησης μιας αίτησης από το διακομιστή ConText για την τρέχουσα σύνοδο και έχει οριστεί από την `SET_ERROR_CALLBACK`.

4.9.2.4 GET_FULL_THEMES

Επιστρέφει τιμή τύπου αλήθειας για το αν δημιουργείται πληροφορία για την ιεραρχία θεμάτων για την τρέχουσα σύνοδο.

4.9.2.5 GET_LOG_PARSE

Επιστρέφει τιμή τύπου αλήθειας για το αν έχει ενεργοποιηθεί για την τρέχουσα σύνοδο το ημερολόγιο (*log*) για την συντακτική ανάλυση.

4.9.2.6 GET_SETTINGS_LABEL

Επιστρέφει το όνομα που χρησιμοποιείται την παραμετροποίηση της τρέχουσας συνόδου.

4.9.2.7 REQUEST_GIST

Χρησιμοποιείται για τη δημιουργία περιλήψεων θεμάτων για ένα έγγραφο. Μπορούν να παραχθούν περιλήψεις σε επίπεδο παραγράφων ή προτάσεων. Εξ' ορισμού η διαδικασία δημιουργεί περιλήψεις για όλα τα θέματα ενός εγγράφου, ωστόσο μπορεί να καθοριστεί το θέμα για το οποίο θα δημιουργηθεί η περίληψη. Οι παράμετροι που δέχεται είναι το όνομα της πολιτικής, το όνομα του κλειδιού για το έγγραφο, το όνομα του πίνακα στον οποίο θα αποθηκευθεί η περίληψη, μια παράμετρος που καθορίζει αν θα προκύψει περίληψη σε επίπεδο παραγράφου ή πρότασης και τέλος μια προαιρετική παράμετρος που καθορίζει το όνομα του θέματος για το οποίο θα δημιουργηθεί η περίληψη.

4.9.2.8 REQUEST_THEMES

Παράγει μια λίστα με το πολύ πενήντα θέματα για ένα έγγραφο. Δέχεται παραμέτρους το όνομα της πολιτικής, το όνομα του κλειδιού για το έγγραφο και το όνομα του πίνακα για την αποθήκευση της λίστας των θεμάτων.

4.9.2.9 SET_COMPLETION_CALLBACK

Καθορίζει το όνομα της διαδικασίας ή συνάρτησης που θα κληθεί όταν ο διακομιστής ConText τελειώνει την εξυπηρέτηση μιας αίτησης από την ουρά. Η παράμετρος που δέχεται είναι το όνομα της διαδικασίας ή συνάρτησης.

4.9.2.10 SET_ERROR_CALLBACK

Καθορίζει το όνομα της διαδικασίας ή συνάρτησης που θα κληθεί όταν κατά τη διάρκεια εξυπηρέτησης μιας αίτησης από τον διακομιστή ConText, προκύψει λάθος. Δέχεται ως παράμετρο το όνομα της διαδικασίας ή συνάρτησης.

4.9.2.11 SET_FULL_THEMES

Αυτή η διαδικασία ενεργοποιεί την εγγραφή πληροφορίας σχετικά με την ιεραρχία θεμάτων στον πίνακα θεμάτων. Το ConText καταγράφει την πληροφορία αυτή όταν καλείται η `REQUEST_THEMES`. Η παράμετρος που δέχεται είναι του τύπου αλήθειας.

4.9.2.12 SET_LOG_PARSE

Ενεργοποιεί το ημερολόγιο για την συντακτική ανάλυση. Δέχεται παράμετρο τύπου αλήθειας για την ενεργοποίηση ή όχι.

4.9.2.13 SET_SETTINGS_LABEL

Η διαδικασία αυτή ονοματίζει τις ρυθμίσεις για τη γλωσσολογική επεξεργασία. Δέχεται ως παράμετρο το όνομα που δίνεται για τις ρυθμίσεις.

4.9.2.14 SUBMIT

Αυτή η συνάρτηση δημιουργεί μια αίτηση (μια γραμμή) στην ουρά εξυπηρέτησης για όλες τις αιτήσεις για γλωσσολογική επεξεργασία που είναι στη μνήμη και επιστρέφει ένα χειριστή για την αίτηση. Οι παράμετροι που δέχεται είναι ένας αριθμός σε δευτερόλεπτα που καθορίζει το μέγιστο διάστημα για το οποίο θα μπλοκάρονται ταυτόχρονες αιτήσεις, καθώς ο διακομιστής ConText επεξεργάζεται την αίτηση, μια δεύτερη παράμετρος του τύπου αλήθειας για το αν θα οριστικοποιηθούν οι αλλαγές στη βάση και μια τρίτη παράμετρος που καθορίζει την προτεραιότητα της αίτησης.

4.9.3 Πακέτο CTX_DDL

Το πακέτο CTX_DDL περιέχει διαδικασίες και συναρτήσεις για τη δημιουργία επιλογών και πολιτικών, καθώς επίσης και την εκτέλεση εργασιών τύπου *DDL (Data Definition Language)*, όπως δημιουργία και βελτιστοποίηση ευρετηρίων.

4.9.3.1 ADD_SECTION

Δημιουργεί ένα προκαθορισμένο τμήμα για την αναζήτηση με τη βοήθεια του τελεστή *WITHIN* και το προσθέτει σε μια υπάρχουσα ομάδα τμημάτων. Δέχεται ως παραμέτρους το όνομα της ομάδας τμημάτων που το ConText θα προσθέσει το νέο τμήμα, το όνομα του νέου τμήματος οι παραπομπές αρχής και τέλους του τμήματος.

4.9.3.2 CLEAR_ATTRIBUTES

Καθαρίζει τη μνήμη από οποιαδήποτε *χαρακτηριστικά (attributes)* έχουν οριστεί με την *SET_ATTRIBUTE*.

4.9.3.3 CREATE_INDEX

Δημιουργεί ένα ευρετήριο για τη στήλη κειμένου που ορίζεται μέσω του ονόματος πολιτικής που δίνεται ως παράμετρος.

4.9.3.4 CREATE_POLICY

Δημιουργεί μια πολιτική. Παράμετροι είναι το όνομα της νέας πολιτικής, το όνομα της στήλης κειμένου του πίνακα με την οποία θα συσχετιστεί η νέα πολιτική, προαιρετικά το όνομα μιας πολιτικής πρότυπο και στη συνέχεια οι επιλογές της πολιτικής (όπως περιγράφονται στην §4.3.2).

4.9.3.5 CREATE_PREFERENCE

Δημιουργεί μια επιλογή στο λεξικό δεδομένων του ConText. Οι ρυθμίσεις για μια επιλογή έχουν από πριν οριστεί με τη διαδικασία `SET_ATTRIBUTE`. Παράμετροι είναι το όνομα της νέας επιλογής, μια περιγραφή και το όνομα της ρύθμισης `tile`.

4.9.3.6 CREATE_SECTION_GROUP

Δημιουργεί μια ομάδα τμημάτων. Παράμετρος είναι το όνομα της ομάδας.

4.9.3.7 CREATE_SOURCE

Ορίζει μια πηγή για τη φόρτωση κειμένου σε μια στήλη κειμένου. Παράμετροι είναι το όνομα της πηγής, το όνομα της στήλης που αντιστοιχίζεται η πηγή, μια περιγραφή και τον αριθμό σε λεπτά που ο διακομιστής ConText θα ελέγχει για ύπαρξη νέων αρχείων στον κατάλογο της πηγής προς φόρτωση στη βάση.

4.9.3.8 CREATE_TEMPLATE_POLICY

Δημιουργεί μια πολιτική πρότυπο που δεν αντιστοιχίζεται σε στήλη κειμένου. Παράμετροι είναι το όνομα της νέας πολιτικής, προαιρετικά μια άλλη πολιτική πρότυπο, μια περιγραφή και οι επιλογές της.

4.9.3.9 DROP_INDEX

Κατάργηση ενός ευρετηρίου. Μοναδική παράμετρος το όνομα του ευρετηρίου.

4.9.3.10 DROP_INTTRIG

Καταργεί τη σκανδάλη ενός πίνακα, η οποία δημιουργείται/ενημερώνεται αυτόματα όταν δημιουργεί ένα ευρετήριο ConText για μια στήλη κειμένου στον πίνακα. Παράμετρος είναι το όνομα του πίνακα.

4.9.3.11 DROP_POLICY

Καταργεί την πολιτική της οποίας το όνομα δίνεται ως παράμετρος.

4.9.3.12 DROP_PREFERENCE

Καταργεί την επιλογή το όνομα της οποίας δίνεται ως παράμετρος.

4.9.3.13 DROP_SECTION_GROUP

Καταργεί την ομάδα τμημάτων το όνομα της οποίας δίνεται ως παράμετρος.

4.9.3.14 DROP_SOURCE

Καταργεί την πηγή το όνομα της οποίας δίνεται ως παράμετρος.

4.9.3.15 OPTIMIZE_INDEX

Βελτιστοποιεί ένα ευρετήριο. Παράμετροι είναι το όνομα της πολιτικής, μια παράμετρος που καθορίζει το είδος της βελτιστοποίησης που θα γίνει καθώς και παράμετροι που έχουν σχέση με τη συμπίεση των όρων, τον αριθμό των διακομιστών που θα χρησιμοποιηθούν στη διαδικασία, κτλ.

4.9.3.16 REMOVE_SECTION

Αφαιρεί ένα τμήμα από μια ομάδα τμημάτων. Το όνομα του τμήματος και της αντίστοιχής του ομάδας δίνονται ως παράμετροι.

4.9.3.17 RESUME_FAILED_INDEX

Συνεχίζει από εκεί που σταμάτησε μια ανεπιτυχής διαδικασία δημιουργίας ή βελτιστοποίησης ευρετηρίου. Παράμετροι είναι το όνομα της πολιτικής και μια παράμετρος που δηλώνει το είδος της διαδικασίας που είχε αποτύχει (δημιουργία/βελτιστοποίηση).

4.9.3.18 SET_ATTRIBUTE

Ορίζει την τιμή μιας ρύθμισης tile. Χρησιμοποιείται σε συνδυασμό με την CREATE_PREFERENCE. Παράμετροι είναι το όνομα του tile, οι τιμές της ρύθμισης (μπορούν να είναι το πολύ δύο) και το όνομα μιας ακολουθίας που χρησιμοποιείται για tiles που υποστηρίζουν πολλαπλές τιμές.

4.9.3.19 UPGRADE_INDEX

Αναβαθμίζει ένα ευρετήριο ConText παλιότερης έκδοσης (2.0 και παλιότερη) στην τρέχουσα.

4.9.3.20 UPDATE_POLICY

Αλλάζει την περιγραφή ή τις επιλογές μιας πολιτικής. Παράμετροι είναι το όνομα της πολιτικής, η περιγραφή της και οι επιλογές της.

4.9.3.21 UPDATE_SOURCE

Αλλάζει τις ρυθμίσεις μιας πηγής. Παράμετροι είναι το όνομα της πηγής και οι νέες ρυθμίσεις.

4.9.4 Πακέτο CTX_DML

Το πακέτο CTX_DML περιλαμβάνει διαδικασίες για εργασίες τύπου *DML (Data Manipulation Language)*.

4.9.4.1 REINDEX

Η διαδικασία αυτή δημιουργεί/ενημερώνει το ευρετήριο ενός εγγράφου. Στην ουσία, αυτό που κάνει είναι να προσθέτει μια αίτηση στην ουρά DML. Δέχεται ως παραμέτρους το όνομα της πολιτικής ή το κλειδί για το έγγραφο του οποίου το ευρετήριο πρέπει να ενημερωθεί.

4.9.4.2 SYNC

Η SYNC προωθεί όλες τις εκκρεμείς αιτήσεις στην DML ουρά προς εκτέλεση από τους διαθέσιμους ConText διακομιστές με το DDL ρόλο. Οι παράμετροι που δέχεται είναι πότε θα αρχίσει η εξυπηρέτηση των αιτήσεων, το όνομα της πολιτικής και ο αριθμός των διακομιστών ConText που θα αρχίσουν την εκτέλεση.

4.9.4.3 SYNC_QUERY

Συνάρτηση που επιστρέφει την ημερομηνία του χαμηλότερου ορίου για το οποίο αιτήσεις στην DML ουρά έχουν εξυπηρετηθεί.

4.9.5 Συναρτήσεις CONTAINS και SCORE

Εκτός από τα πακέτα που προσφέρει το ConText για την εκτέλεση ερωτήσεων, υπάρχουν και δύο συναρτήσεις για την εκτέλεση διαδραστικών ερωτήσεων (όπως για παράδειγμα, αυτά που εκτελούνται από SQL Plus⁷). Προκειμένου να χρησιμοποιηθούν οι δύο αυτές συναρτήσεις, πρέπει στο αρχείο εκκίνησης της βάσης (init_{orcl}.ora) να έχει δηλωθεί η παράμετρος TEXT_ENABLE = TRUE.

4.9.5.1 Συνάρτηση CONTAINS

Η CONTAINS χρησιμοποιείται στο WHERE ενός SELECT ερώτησης για ερωτήσεις ενός βήματος. Οι παράμετροι που δέχεται είναι η στήλη κειμένου που θα γίνει η αναζήτηση, η ερώτηση, μια αλφαριθμητική μεταβλητή που είναι ετικέτα για το σκορ που βγάζει η συνάρτηση και τέλος μια παράμετρος που είναι το όνομα πολιτικής στην περίπτωση στηλών με πολλαπλές πολιτικές.

4.9.5.2 Συνάρτηση SCORE

Χρησιμοποιείται σε συνδυασμό με την CONTAINS για να επιστρέψει τις τιμές των σκορ που επιστρέφει η CONTAINS σε ερώτηση ενός βήματος.

4.9.6 Ερώτηση SELECT

Παρακάτω φαίνεται η σύνταξη της SELECT ερώτησης :

```
SELECT SCORE(label1), SCORE(label2), ...SCORE(labeln),
       column1, column2, ... columnn
FROM table
   WHERE CONTAINS (columnid, 'text_query', label1, polhint1) > 0
CONTAINS (columnid, 'text_query', label2, polhint2) > 0,
.
.
.
CONTAINS (columnid, 'text_query', labeln, polhintn) > 0
ORDER BY SCORE(labeln)
```

label_x είναι ένας αριθμός που καθορίζει ποια δήλωση CONTAINS δημιούργησε το σκορ (απαραίτητος μόνο αν υπάρχουν πολλές δηλώσεις CONTAINS στην ερώτηση)

column_n είναι η στήλη που επιστρέφεται από την ερώτηση. Κάθε δήλωση CONTAINS δημιουργεί μια εικονική στήλη SCORE που μπορεί να κληθεί με την αριθμητική της ετικέτα (label_x) και να συμπεριληφθεί στην έξοδο της ερώτησης

table είναι το όνομα του πίνακα που έχει τη στήλη κειμένου στην οποία γίνεται αναζήτηση

column_{id} είναι το όνομα της στήλης κειμένου

text_query είναι η ερώτηση για την αναζήτηση και

pol_hint είναι το όνομα μιας πολιτικής όταν η στήλη column_{id} έχει πολλαπλές πολιτικές

⁷ SQL Plus είναι το εργαλείο της Oracle για την εκτέλεση διαδραστικών ερωτήσεων.

Η συνάρτηση `CONTAINS` πρέπει πάντα να ακολουθείται από την έκφραση `>0` που καθορίζει ότι το σκορ που δημιουργεί πρέπει να είναι μεγαλύτερο του μηδενός προκειμένου να ανακτηθεί η εγγραφή.

ΚΕΦΑΛΑΙΟ 5

ΠΡΟΤΥΠΟ ΠΕΡΙΒΑΛΛΟΝ ΙΣΤΟΣΕΛΙΔΩΝ

5.1 Εισαγωγή

Το τρέχον κεφάλαιο περιγράφει το περιβάλλον που δημιουργήθηκε για την επίδειξη των δυνατοτήτων του Text Extender και του ConText cartridge των βάσεων DB2 και Oracle αντίστοιχα.

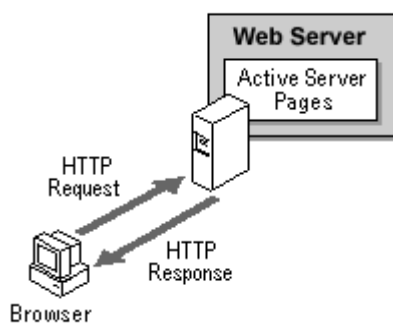
Συγκεκριμένα, χρησιμοποιήθηκε η τεχνολογία των *Active Server Pages* (asp χάριν συντομίας) που παρέχει ο διακομιστής ιστοσελίδων *Internet Information Server* (IIS) της Microsoft σε περιβάλλον Windows NT. Η σύνδεση με τη βάση επιτεύχθηκε μέσω ODBC¹ και χρησιμοποιήθηκε το προγραμματιστικό μοντέλο *ADO* (*ActiveX Data Objects*) για την επικοινωνία με τους διακομιστές των βάσεων.

Παρακάτω, θα περιγραφεί σύντομα το μοντέλο των asp σελίδων, καθώς και το προγραμματιστικό μοντέλο ADO. Παράλληλα, θα υπάρξει και παρουσίαση των σελίδων του περιβάλλοντος.

5.2 Σύντομη περιγραφή των asp σελίδων

Οι *Active Server Pages* σελίδες είναι περιβάλλον εκτέλεσης σεναρίων (*scripts*) στο διακομιστή που χρησιμοποιείται για τη δημιουργία σελίδων HTML με δυναμικό περιεχόμενο. Δημιουργούνται έτσι *διαδραστικές* (*interactive*) web εφαρμογές. Προς το παρόν, υποστηρίζονται μόνο από τον διακομιστή *Internet Information Server* της Microsoft.

Ένα σενάριο asp αρχίζει να εκτελείται όταν ο πελάτης ζητήσει μια σελίδα asp από το πρόγραμμα πλοήγησής του. Ο διακομιστής των ιστοσελίδων IIS καλεί την asp σελίδα και αρχίζει να την διαβάζει από την αρχή, εκτελώντας οποιαδήποτε σενάρια περιέχονται σε αυτή. Στο τέλος προκύπτει η σελίδα HTML που επιστρέφεται στον πελάτη. Το σχήμα 5.1 δείχνει αυτή τη διαδικασία.



Σχήμα 5.1: Διαδικασία αίτησης asp σελίδας

Λόγω του γεγονότος ότι το σενάριο τρέχει στον διακομιστή IIS και όχι στον πελάτη, ο χρήστης δεν πρέπει να ανησυχεί για το αν το πρόγραμμα πλοήγησής του υποστηρίζει τα σενάρια, εφόσον αυτό που τελικά λαμβάνει είναι μια σελίδα με στάνταρ HTML. Επίσης, τα σενάρια που εκτελούνται στον IIS και που δημιουργούν τις σελίδες που στέλνονται στον πελάτη, δεν μπορεί να τα δει κανείς.

Τα σενάρια που τρέχουν στον IIS διακομιστή γράφονται σε *γλώσσες σεναρίων* (*scripting languages*) που υποστηρίζει ο IIS. Προκειμένου να χρησιμοποιηθεί μια γλώσσα σεναρίων πρέπει να υπάρχει εγκατεστημένη η αντίστοιχη *μηχανή σεναρίων* (*scripting engine*) που μεταφράζει και εκτελεί τις εντολές της γλώσσας σεναρίων που υποστηρίζει. Ο IIS έρχεται εξ' ορισμού με δύο μηχανές, που υποστηρίζουν αντίστοιχα τις γλώσσες σεναρίων VBScript (Visual Basic Scripting Edition) και JScript.

Μια χρήσιμη δυνατότητα είναι η χρησιμοποίηση των *ActiveX συστατικών* (*components*). Ένα ActiveX συστατικό είναι ένα αρχείο που περιέχει κώδικα που εκτελεί ορισμένες εργασίες. Τα συστατικά χρησιμοποιούνται ως προσθήκες για τη χρησιμοποίηση κοινότυπων εργασιών, για παράδειγμα μια κυλιόμενη λωρίδα που εμφανίζει διάφορα μη-

¹ ODBC (Open Database Connectivity) : Πρωτόκολλο που προσφέρει μια κοινή γλώσσα που επιτρέπει την επικοινωνία εφαρμογών με συστήματα διαχείρισης βάσεων δεδομένων. Αναπτύχθηκε από τη Microsoft για την πλατφόρμα των Windows και χρησιμοποιείται ευρέως από Windows εφαρμογές.

νύματα θα μπορούσε να αποτελεί ένα ActiveX συστατικό. Ο κώδικας ενός συστατικού μπορεί να περιέχεται είτε σε ένα αρχείο εκτελέσιμου (.exe αρχείο), είτε σε μια δυναμική βιβλιοθήκη των Windows (.dll αρχείο). Ο IIS έρχεται με μερικά προεγκατεστημένα ActiveX συστατικά.

Οι asp σελίδες έχουν ενσωματωμένα αντικείμενα που χρησιμοποιούνται για την επικοινωνία με τον πελάτη, για διάφορες πληροφορίες του IIS, κτλ. Τα αντικείμενα αυτά είναι έξι στον αριθμό και είναι τα εξής : το *αντικείμενο εφαρμογής (Application Object)*, που χρησιμοποιείται για πληροφορίες που μοιράζονται οι χρήστες μιας εφαρμογής web, το *αντικείμενο αίτησης (Request Object)*, που χρησιμοποιείται για την πρόσβαση στις πληροφορίες μιας αίτησης HTTP, το *αντικείμενο απόκρισης (Response Object)*, που χρησιμοποιείται για τον έλεγχο της πληροφορίας που στέλνεται στον πελάτη, το *αντικείμενο διακομιστή (Server Object)*, που χρησιμοποιείται για την προσπέλαση μεθόδων και ιδιοτήτων του διακομιστή, το *αντικείμενο συνόδου (Session Object)*, για την αποθήκευση πληροφοριών για μια συγκεκριμένη σύνοδο ενός χρήστη και το *αντικείμενοObjectContext* που χρησιμοποιείται για την οριστικοποίηση ή ακύρωση μιας *συναλλαγής (transaction)* που ξεκίνησε ένα asp σενάριο.

5.3 Περιγραφή του μοντέλου ADO

5.3.1 Γενικά

ADO (ActiveX Data Objects) είναι μια τεχνολογία για την προσπέλαση βάσεων δεδομένων μέσω ιστοσελίδων. Το ADO μπορεί να χρησιμοποιηθεί για τη συγγραφή σεναρίων που προσπελαίνουν βάσεις μέσω του πρωτοκόλλου ODBC χρησιμοποιώντας αντικείμενα. Μερικά κύρια χαρακτηριστικά του ADO είναι τα εξής :

- Δημιουργία ανεξάρτητων αντικειμένων. Τα ADO αντικείμενα μπορούν να δημιουργηθούν ανεξάρτητα, χωρίς να απαιτούν τη δημιουργία υψηλότερων στην ιεραρχία αντικειμένων, με αποτέλεσμα ο προγραμματιστής να δημιουργεί μόνο τα αντικείμενα που χρειάζεται και να δουλεύει με ένα μικρό σετ αντικειμένων.
- Το ADO υποστηρίζει *μαζική ενημέρωση (batch update)*, κρατώντας τοπικά τις αλλαγές σε μια *λανθάνουσα μνήμη (cache)* και ενημερώνει τη βάση με μια ενέργεια.
- Υποστήριξη αποθηκευμένων διαδικασιών με πέρασμα και επιστροφή παραμέτρων.
- Υποστήριξη διαφόρων τύπων δρομέων (*cursors*).
- Υποστήριξη για μέγιστο επιστρεφόμενο αριθμό εγγραφών και άλλων βοηθημάτων για βελτιστοποίηση της απόδοσης.
- Υποστήριξη για πολλαπλά *σετ εγγραφών (record sets)* από αποθηκευμένες διαδικασίες ή δηλώσεις.
- Υποστήριξη αντικειμένων που χρησιμοποιούν *νήματα (threads)* του λειτουργικού για καλύτερη απόδοση.

5.3.2 ADO Αντικείμενα, συλλογές και γεγονότα

Το ADO είναι *αντικειμενοστραφές (object-oriented)* μοντέλο που σημαίνει ότι χρησιμοποιεί αντικείμενα για την πρόσβαση στις βάσεις δεδομένων και *μεθόδους (methods)* για τις διάφορες λειτουργίες. Το ADO βρίσκεται πλέον στην έκδοση 2.0 και τα αντικείμενα που έχει είναι τα εξής :

- *Connection (σύνδεση)* αντικείμενο, που χρησιμοποιείται για τη σύνδεση με τη βάση δεδομένων και για τη δημιουργία του κατάλληλου περιβάλλοντος για την α-

νταλλαγή δεδομένων με αυτή. Μέσω αυτού του αντικείμενου μπορούν επίσης να οριστούν συναλλαγές μέσω μεθόδων.

- *Command* (εντολή) αντικείμενο, που ενσαρκώνει την ιδέα της εντολής, η οποία χειρίζεται τα δεδομένα με κάποιο τρόπο, δηλαδή προσθέτει νέες εγγραφές, διαγράφει, ανακτά ή ενημερώνει υπάρχουσες.
- *Parameter* (παράμετρος) αντικείμενο, που χρησιμοποιείται για τη δημιουργία παραμέτρων για ένα *Command* αντικείμενο. Χρησιμοποιείται συνήθως για την κλήση διαδικασιών ή συναρτήσεων που δέχονται ή επιστρέφουν παραμέτρους.
- *Recordset* (σετ εγγραφών) αντικείμενο, που έχει την έννοια του δρομέα (*cursor*) που χρησιμοποιείται για την αποθήκευση των εγγραφών που επιστρέφει η εκτέλεση μιας εντολής. Υποστηρίζει φυσικά την ενημέρωση των εγγραφών. Το *Recordset* αντικείμενο χρησιμοποιείται ακόμη για μια εγγραφή (*row*), διότι δεν υπάρχει ADO αντικείμενο που να αντιπροσωπεύει μια εγγραφή.
- *Field* (πεδίο) αντικείμενο, που αντιπροσωπεύει ένα πεδίο μιας εγγραφής.
- *Error* (σφάλμα) αντικείμενο, που αντιπροσωπεύει τα λάθη που μπορούν να προκύψουν από μια διαδικασία, για παράδειγμα μια προσπάθεια ενημέρωσης μιας εγγραφής. Προγραμματιστικά, για κάθε σελίδα asp υπάρχει το ενσωματωμένο αντικείμενο `Err`, στο οποίο συσσωρεύονται όλα τα λάθη της asp σελίδας. Όταν συμβεί ένα νέο λάθος, τα υπόλοιπα λάθη διαγράφονται από τη συλλογή.
- *Property* (ιδιότητα) αντικείμενο, που αντιπροσωπεύει το πώς ένα αντικείμενο συμπεριφέρεται. Υπάρχουν δύο είδη ιδιοτήτων, οι ενσωματωμένες και οι δυναμικές. Οι ενσωματωμένες ιδιότητες είναι μέρος των ADO αντικειμένων, ενώ οι δυναμικές προστίθενται στη συλλογή *Properties*, ανάλογα με τον *προμηθευτή δεδομένων* (*data provider*) που χρησιμοποιείται, ο οποίος εξαρτάται από τον οδηγό ODBC.

Μια κύρια έννοια για το ADO είναι οι *συλλογές* (*collections*). Οι συλλογές περιέχουν αντικείμενα του ίδιου τύπου. Το ADO παρέχει τέσσερις τύπους συλλογών :

- Το *Connection* αντικείμενο παρέχει τη συλλογή *Errors* που περιέχει όλα τα *Error* αντικείμενα.
- Το *Command* αντικείμενο παρέχει τη συλλογή *Parameters* που περιέχει όλα τα αντικείμενα τύπου *Parameter* που αντιστοιχούν στο συγκεκριμένο *Command* αντικείμενο.
- Το *Recordset* αντικείμενο παρέχει τη συλλογή *Fields* που περιέχει όλα τα αντικείμενα τύπου *Field* για το συγκεκριμένο αντικείμενο *Recordset*.
- Επίσης, τα αντικείμενα *Connection*, *Command*, *Recordset* και *Field* περιέχουν τη συλλογή *Properties*, που περιέχουν όλα τα αντικείμενα τύπου *Property* που αντιστοιχούν σε κάθε αντικείμενο αντίστοιχα.

Η έκδοση 2.0 του ADO εισάγει μια νέα έννοια. Την έννοια του *γεγονότος* (*event*). Τα γεγονότα είναι ένα είδος ειδοποιήσεων ότι ορισμένες ενέργειες πρόκειται ή έχουν ήδη συμβεί. Τα γεγονότα μπορούν να χρησιμοποιηθούν για το συγχρονισμό πολλών ασύγχρονων διαδικασιών. Δεν υπάρχει κάποιο αντικείμενο που να αντιπροσωπεύει τα γεγονότα, αλλά υπάρχουν οι *ρουτίνες χειριστές των γεγονότων* (*event handler routines*). Οι ρουτίνες αυτές χρησιμοποιούνται για να εξετάσουν την κατάσταση μια διαδικασίας και να την επιτρέψουν ή όχι να εκτελεστεί.

Υπάρχουν δύο οικογένειες γεγονότων, τα *γεγονότα συνδέσεων* (*connection events*) και τα *γεγονότα σετ εγγραφών* (*recordset events*). Τα πρώτα συμβαίνουν όταν αρχίζει μια συναλλαγή, όταν οι αλλαγές οριστικοποιούνται ή ακυρώνονται, όταν εκτελούνται εντολές και φυσικά όταν ανοίγει και κλείνει μια σύνδεση με τη βάση. Τα δεύτερα συμβαίνουν για

να δείξουν την κατάσταση των δεδομένων σε ένα σενε εγγραφών, όταν γίνεται προσπέ-
λαση ή αλλαγή μεμονωμένων εγγραφών ή ολόκληρου του σενε εγγραφών.

5.4 Περιγραφή του πρότυπου περιβάλλοντος

Το περιβάλλον δημιουργήθηκε σε λειτουργικό σύστημα Windows NT 4 Server (service pack 6a), με Option Pack για NT Server, το οποίο εγκαθιστά τον Internet Information Server 4. Η συγγραφή των asp σελίδων έγινε με τη βοήθεια του Visual InterDev από το Visual Studio 6 (το οποίο ωστόσο δεν είναι απαραίτητο). Επίσης, το Option Pack περιέχει και τις *επεκτάσεις διακομιστή (server extensions)* FrontPage τα οποία πρέπει να εγκατασταθούν στον IIS για να χρησιμοποιηθεί το InterDev με τον IIS.

Η γλώσσα σεναρίων που χρησιμοποιήθηκε ήταν η VBScript.

5.4.1 Σύνδεση με τη βάση

Για σύνδεση με τις βάσεις χρησιμοποιήθηκε το αντικείμενο Connection του ADO. Ορίζεται μια μεταβλητή η οποία είναι αντικείμενο Connection και ορίζει σύνδεση με ένα *όνομα πηγής δεδομένων (data source name)*. Το όνομα πηγής δεδομένων ορίζεται από το εικονίδιο ODBC Data Sources του Control Panel των Windows και είναι ένα όνομα που ορίζει μια σύνδεση με τη βάση (DB2 και Oracle στην συγκεκριμένη περίπτωση) μέσω ODBC. Ο κώδικας που δημιουργεί ένα αντικείμενο, ανοίγει και κλείνει τη σύνδεση φαίνεται στη λίστα κώδικα 5.1.

```
Dim dbConnection
Set dbConnection=Server.CreateObject("ADODB.Connection")
dbConnection.ConnectionString="DSN=Oracle;UID=ctxdemo;PWD=demopass"
dbConnection.Open
.
.
.
dbConnection.Close
```

Λίστα κώδικα 5.1: Δημιουργία σύνδεσης με βάση

Από τη λίστα φαίνεται η δημιουργία του αντικειμένου dbConnection μέσω της μεθόδου CreateObject του ενσωματωμένου αντικειμένου Server. Στη συνέχεια ορίζεται η ιδιότητα ConnectionString που στην ουσία είναι ο ορισμός του ονόματος πηγής δεδομένων (Oracle όπως ορίστηκε στο ODBC Data Sources). Η μέθοδος Open ανοίγει τη σύνδεση με τη βάση, ενώ παρακάτω η Close την τερματίζει.

5.4.2 Ανάκτηση δεδομένων από τη βάση

Για την ανάκτηση δεδομένων από τις βάσεις χρησιμοποιήθηκε το αντικείμενο Recordset και Command. Το Recordset μπορεί να χρησιμοποιηθεί χωρίς το αντίστοιχο Command αντικείμενο, ωστόσο, όταν η SQL ερώτηση έχει παραμέτρους, πρέπει να χρησιμοποιηθεί το αντικείμενο Command σε συνδυασμό με το Recordset. Παράδειγμα φαίνεται στη λίστα κώδικα 5.2.

```
Dim sqlCommand, recordSet
Set sqlCommand=Server.CreateObject("ADODB.Command")
sqlCommand.CommandText="select doc_id, author from articles " & _
    "order by pub_date desc"
Set recordSet=sqlCommand.Execute
Do Until recordSet.EOF
```

```

Response.Write("<p>Document ID : " & recordSet("doc_id") & _
    " Author : " & recordSet("author") & "</p>")
recordSet.MoveNext
Loop
recordSet.Close

```

Λίστα κώδικα 5.2: Εκτέλεση SQL ερώτησης και ανάκτηση δεδομένων

Όπως φαίνεται από τη λίστα, ορίζονται δύο μεταβλητές `sqlCommand` και `recordSet` που θα αποτελέσουν τα αντικείμενα τύπου `Command` και `Recordset` αντίστοιχα. Αρχικά, ορίζεται το αντικείμενο του τύπου `Command` και η SQL ερώτηση μέσω της ιδιότητας `CommandText`. Στη συνέχεια ορίζεται το `recordSet` μέσω της μεθόδου `Execute` του `sqlCommand` που επιστρέφει αντικείμενο του τύπου `Recordset`. Χρησιμοποιώντας ένα βρόχο τύπου `Until`, προσπελάζονται κάθε εγγραφή του `recordSet` και μέσω της μεθόδου `Write` του ενσωματωμένου αντικειμένου `Response` επιστρέφεται η έξοδος στον πελάτη ως σελίδα HTML (ένα μεμονωμένο μέρος). Η μέθοδος `MoveNext` του `recordSet` προχωρά το δείκτη του σετ εγγραφών κατά μία εγγραφή. Τέλος η μέθοδος `Close` αποδεσμεύει τη μνήμη που είχε αρχικά δεσμευθεί για το `recordSet`.

5.4.3 Περιγραφή των ιστοσελίδων για τον Text Extender

Για τη δημιουργία των ιστοσελίδων για τον Text Extender χρησιμοποιήθηκε το στιγμιότυπο βάσης που δημιουργείται και περιέχει τους πίνακες επίδειξης όπως περιγράφεται στο κεφάλαιο 3 (§3.9). Το σενάριο επίδειξης δημιουργεί τους πίνακες `sample_1` (για το γλωσσολογικό ευρετήριο), `sample_d` (για το διπλό ευρετήριο), `sample_lfe` (για το γλωσσολογικό ευρετήριο με εξαγωγή χαρακτηριστικών), `sample_n` (για το ngram ευρετήριο), `sample_nce` (για το ngram ευρετήριο με την επιλογή `CASE_ENABLED`) και `sample_p` (για το ακριβές ευρετήριο). Όλοι οι πίνακες έχουν την ίδια δομή (κύριο κλειδί που είναι ο αριθμός του εγγράφου, συγγραφέας, θέμα, ημερομηνία και η στήλη κειμένου), με τη διαφορά ότι ευρετηριοποιούνται με διαφορετικό είδος ευρετηρίου κάθε φορά. Παράλληλα, το σενάριο επίδειξης εισάγει στους πίνακες την ίδια συλλογή εγγράφων. Με το πέρας της εκτέλεσης του σεναρίου, ο χρήστης έχει έτοιμους τους πίνακες για αναζήτηση.

Η αρχική σελίδα είναι προσπελάσιμη μέσω του URL `http://localhost/db2/`. Επιλέγοντας τη σελίδα αυτή ο χρήστης έχει πέντε επιλογές που έχουν σχέση με τα είδη ευρετηρίων που υποστηρίζει ο Text Extender. Αφού ο χρήστης επιλέξει κάτι θα οδηγηθεί στη δεύτερη σελίδα, που είναι η σελίδα στην οποία θα δώσει τους όρους αναζήτησης. Στη σελίδα αυτή μπορεί να ορίσει μέγιστο αριθμό εγγράφων στο αποτέλεσμα, τσεκάροντας το αντίστοιχο πεδίο και εισάγοντας ένα ακέραιο μεταξύ 1 και 32.767 (Εικόνα 5.1). Ο μέγι-

Εικόνα 5.1: Επιλογές για μέγιστο αριθμό αποτελεσμάτων και χαρακτήρα διαφυγής

στος αριθμός εγγράφων αντιστοιχεί στην παράμετρο `result limit` της ερώτησης. Επίσης μπορεί να ορίσει χαρακτήρα διαφυγής, αν στην ερώτησή του χρησιμοποιεί τους χαρακτήρες `%` ή `_` και θέλει να ερμηνευθούν ως κυριολεκτικοί χαρακτήρες κι όχι ως μεταχαρακτήρες. Μπορεί να τσεκάρει την κατάλληλη επιλογή και να εισάγει ένα χαρακτήρα ο οποίος θα είναι ο χαρακτήρας διαφυγής. Αντιστοιχεί στην παράμετρο `escape`.

Στη συνέχεια της σελίδας (Εικόνα 5.2) ακολουθούν τα πεδία κειμένου στα οποία ο χρήστης μπορεί να εισάγει τους όρους αναζήτησης. Ο χρήστης και ανάλογα με το είδος ευρετηρίου που επέλεξε αρχικά έχει διαθέσιμες επιλογές, η καθεμία από τις οποίες αντι-

Εισάγετε τους όρους χωρισμένους με κόμμα (πχ. "database","DB2","IBM"). Οι λογικοί τελεστές είναι το & (AND), το | (OR) και το NOT (πχ. NOT "database"). Τα διπλά εισαγωγικά (") είναι απαραίτητα για κάθε όρο.

Αναζήτηση

STEMMED FORM OF

Εισάγετε τον όρο.

Αναζήτηση

IN SAME PARAGRAPH AS

Εισάγετε τον πρώτο όρο, επιλέξτε αν θέλετε να κάνετε αναζήτηση στην ίδια πρόταση ή παράγραφο και εισάγετε και τον δεύτερο όρο.

Αναζήτηση

IS ABOUT

Στο δεύτερο πεδίο κειμένου εισάγετε τη φράση με την οποία θέλετε να κάνετε αναζήτηση. Προαιρετικά, μπορείτε να εισάγετε ένα όρο στο πρώτο πεδίο και τη συνθήκη (AND).

Αναζήτηση

PERSON

Επιλέξτε το χαρακτηριστικό που θέλετε να κάνετε εξαγωγή και εισάγετε τον όρο.

Αναζήτηση

Εικόνα 5.2: Διαθέσιμες επιλογές

στοιχεί σε διαφορετικό τρόπο αναζήτησης. Ο χρήστης πρέπει να συμπληρώσει **μία** από όλες τις επιλογές και να πατήσει το κουμπί «Αναζήτηση» ώστε η ερώτησή του να δοθεί προς εκτέλεση. Η πρώτη επιλογή αντιστοιχεί στην πιο γενική μορφή αναζήτησης. Μπορεί να δεχθεί πολλούς όρους συνδεδεμένους με τους λογικούς τελεστές ή έναν. Αυτό το πεδίο εισαγωγής κειμένου είναι και το μοναδικό στο οποίο οι όροι πρέπει να περικλείονται σε διπλά εισαγωγικά ("). Η δεύτερη επιλογή αφορά στις παραμέτρους *STEMMED FORM OF*, *SYNONYM FORM OF* και *SOUNDS LIKE*. Στην περίπτωση του *ngram* ευρετηρίου η επιλογή *SOUNDS LIKE* αντικαθίσταται από την *PRECISE FORM OF*. Σε αυτό το πεδίο κειμένου ο χρήστης μπορεί να εισάγει ένα μόνο όρο. Η τρίτη επιλογή έχει σχέση με την εύρεση όρων που βρίσκονται στην ίδια πρόταση ή παράγραφο (παραμέτροι *IN SAME SENTENCE AS* και *IN SAME PARAGRAPH AS*). Εδώ ο χρήστης πρέπει να εισάγει και τους δύο όρους. Η τέταρτη επιλογή έχει σχέση με την ερώτηση σε φυσική γλώσσα (παραμέτρος *IS ABOUT*). Προαιρετικά μπορεί να δοθεί και μια συνθήκη *AND* με ένα όρο. Η επιλογή δεν είναι διαθέσιμη στην περίπτωση του *ngram* ευρετηρίου. Στην περίπτωση του γλωσσολογικού ευρετηρίου με εξαγωγή χαρακτηριστικών, υπάρχει μια επιλογή όπου ο χρήστης εισάγει έναν όρο και επιλέγει το είδος της σχέσης (*PERSON*, *PLACE*, *ORGANISATION*, *DOMAIN*, *ANY*). Τέλος, για την περίπτωση του *ngram* ευρετηρίου (Εικόνα 5.3), υπάρχει μια επιλογή η οποία έχει σχέση με εύρεση εγγράφων που περιέχουν όρους που ακούγονται «ηχητικά» το ίδιο περίπου με τον όρο που εισάγει ο χρήστης. Παράλληλα, ο χρήστης μπορεί να επιλέξει και βαθμό ομοιότητας (μεταξύ ένα και πέντε, με ένα το χαλαρότερο και πέντε το αυστηρότερο).

Πρέπει να σημειωθεί ότι στη σελίδα αυτή υπάρχει και έλεγχος των πεδίων κειμένου και όλων των φορμών της σελίδας για τυχόν παραλήψεις του χρήστη. Ο έλεγχος γίνε-

Εισάγετε τον όρο και επιλέξτε συντελεστή ομοιότητας με το 5 να είναι το πιο αυστηρό και 1 το πιο χαλαρό.

Συντελεστής ομοιότητας

5 4 3 2 1

Αναζήτηση

Εικόνα 5.3: Επιπλέον επιλογή για το ngram ευρετήριο

ται με σενάρια γραμμένα σε JavaScript στο πρόγραμμα πλοήγησης του χρήστη, δηλαδή στον πελάτη.

Αφού ο χρήστης εισάγει τον όρο ή τους όρους και πατήσει το κουμπί «Αναζήτηση», οδηγείται στην τρίτη σελίδα. Η σελίδα αυτή είναι η σελίδα που παρουσιάζει το αποτέλεσμα. Αρχικά, εμφανίζεται ο αριθμός των αποτελεσμάτων, ο οποίος λαμβάνεται από μια ερώτηση SQL στη DB2 :

```
select count(*) from db2tx.πίνακας where
db2tx.contains(db2tx.πίνακας.commenthandle, συνθήκη) = 1
```

όπου *πίνακας* είναι το όνομα του πίνακα στη βάση και το οποίο λαμβάνεται ανάλογα με το είδος του ευρετηρίου που επέλεξε αρχικά ο χρήστης και *συνθήκη* είναι η ερώτηση που

Αριθμός αποτελεσμάτων : 3					
Έγγραφο	Διαβάθμιση	Εμφανίσεις	Συγγραφέας	Τίτλος	Ημερομηνία
doc48	0,154039	1	NEWSGATE at TOROLAB2 for SREES at TOROLAB	Re: cobol	26/7/1995
doc47	0,147483	1	NEWSGATE at TOROLAB2 for JHORNIBR at TOROLAB	Re: systable is locked	26/7/1995
doc25	0,144532	1	NEWSGATE at TOROLAB2 for SREES at TOROLAB	Re: Cobol Microfocus 32 bit and DB2/2 V2	26/7/1995

Εικόνα 5.4: Λίστα εγγράφων αποτελέσματος

εισήγαγε στο πεδίο κειμένου. Στη συνέχεια και με μορφή λίστας (Εικόνα 5.4), εμφανίζονται τα έγγραφα που ικανοποιούν τη συνθήκη του χρήστη, συγκεκριμένα εμφανίζεται ο αριθμός ταυτότητα του εγγράφου, η διαβάθμιση του εγγράφου κατά την οποία γίνεται και η φθίνουσα ταξινόμηση των εγγράφων, ο αριθμός εμφάνισης του όρου που εισήγαγε ο χρήστης μέσα στο έγγραφο, το όνομα του συγγραφέα, ο τίτλος, καθώς και η ημερομηνία του εγγράφου. Όλες αυτές οι πληροφορίες λαμβάνονται με την εκτέλεση της SQL ερώτησης :

```
select docid, author, subject, sysibm.date(date) date,
db2tx.rank (db2tx.πίνακας.commenthandle, συνθήκη),
db2tx.no_of_matches (db2tx.πίνακας.commenthandle, συνθήκη) from
db2tx.πίνακας where db2tx.contains (db2tx.πίνακας.commenthandle, συνθήκη)
= 1 order by 5 desc
```

όπου χρησιμοποιείται η συνάρτηση *rank* για την έξοδο του αριθμού διαβάθμισης του εγγράφου, *no_of_matches* για τον αριθμό εμφανίσεων της *συνθήκης* μέσα στο έγγραφο και φυσικά η *contains* η οποία επιστρέφει 1 στην περίπτωση που το έγγραφο ικανοποιεί τη *συνθήκη* του χρήστη. Η συνάρτηση *sysibm.date* απλά επιστρέφει το κομμάτι της ημερομηνίας της στήλης *date* του πίνακα μιας και η στήλη αυτή είναι του τύπου *timestamp* που περιλαμβάνει εκτός της ημερομηνίας, ώρα και χιλιοστά δευτερολέπτου.

Ο αριθμός ταυτότητα κάθε εγγράφου αποτελεί σύνδεσμο που οδηγεί σε μια σελίδα που εμφανίζει όλες τις πληροφορίες του εγγράφου, μαζί με το κείμενό του. Αυτή η σελίδα εκτελεί την ερώτηση :

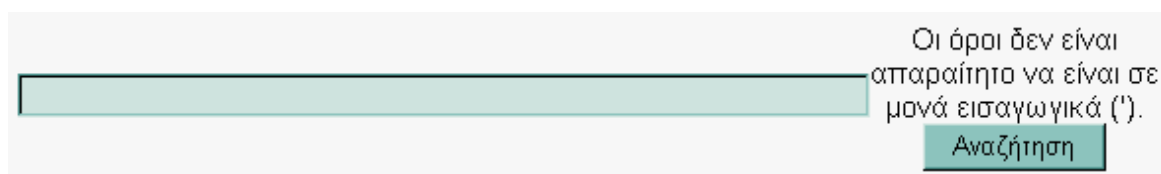
```
select author, subject, date, comment
from db2tx.πίνακας
where docid=αριθμός
```

όπου αριθμός είναι ο αριθμός ταυτότητα του εγγράφου.

5.4.4 Περιγραφή των ιστοσελίδων για το ConText cartridge

Για τις ιστοσελίδες του ConText cartridge χρησιμοποιήθηκαν οι πίνακες που δημιουργήθηκαν κατά την ενεργοποίηση του ConText cartridge όπως περιγράφεται στο κεφάλαιο 4 (§4.4.1). Οι πίνακες που χρησιμοποιούνται σε αυτή την περίπτωση είναι οι articles, article_themes, article_gists, highlight_table, highlight_temp, query_temp, themes_table. Παράλληλα, υπάρχουν και δύο πολιτικές DEMO_POLICY και DEMO_THEME_POLICY, από τις οποίες η πρώτη είναι για απλές ερωτήσεις, ενώ η δεύτερη είναι για ερωτήσεις θεμάτων. Η πρώτη έχει συσχετιστεί με τον πίνακα articles, ενώ η δεύτερη με τον themes_table. Οι σελίδες είναι προσπελάσιμες μέσω του URL <http://localhost/ora>.

Στην πρώτη σελίδα ο χρήστης έχει τρεις επιλογές, την επιλογή για απλή ερώτηση, ερώτηση θέματος και την εξαγωγή περιλήψεων/εννοιών των εγγράφων του πίνακα articles. Τόσο η πρώτη όσο και η δεύτερη επιλογή οδηγούν στην ίδια σελίδα, με τη διαφορά ότι για τις ερωτήσεις θεμάτων οι τελεστές επαύξεσης και NEAR δεν είναι διαθέσιμοι και ότι χρησιμοποιείται ο πίνακας themes_table. Στη δεύτερη σελίδα, υπάρχει μόνο



Εικόνα 5.5: Πεδίο εισαγωγής ερώτησης

ένα διαθέσιμο πεδίο κειμένου (Εικόνα 5.5) στο οποίο ο χρήστης μπορεί να εισάγει την ερώτησή του. Σε αυτή την περίπτωση δεν είναι απαραίτητο οι όροι να περικλείονται με

Λογικοί τελεστές AND & OR EQUIV = (term1=term2)	Τελεστές αλλαγής σκορ ACCUMULATE . (term1,term2) MINUS - (term1-term2) WEIGHT *n (term*n)	Τελεστές ελέγχου αποτελεσμάτων THRESHOLD expression>n MAX expression (1<=n<=B553B) FIRST/NEXT expression#m-n (m<=n)	
Τελεστές επαύξεσης STEM (\$term) SOUNDEX (Iterm) FUZZY (?term)	Τελεστές NEAR NEAR((word1, word2, ..., wordn)[, MAX SPAN[, ORDER]]) Χαρακτήρες διαφυγής { } \	Χαρακτήρες μίσκας % _ Χαρακτήρες γκρουπ () []	

Εικόνα 5.6: Περιγραφή τελεστών

μονά εισαγωγικά ('). Στο κάτω μέρος της σελίδας (Εικόνα 5.6) υπάρχει περιγραφή των διαθέσιμων τελεστών. Αφού ο χρήστης εισάγει τη συνθήκη (χρησιμοποιώντας έναν ή περισσότερους από τους τελεστές που περιγράφονται παρακάτω) οδηγείται στην τρίτη σελίδα, αυτή με την παρουσίαση των εγγράφων που ικανοποιούν τη συνθήκη του χρήστη. Η

σελίδα αυτή εκτελεί αρκετές SQL δηλώσεις. Πρέπει να σημειωθεί ότι χρησιμοποιούνται πάντα ερωτήσεις δύο βημάτων, δηλαδή υπάρχει κάποιος προσωρινός πίνακας που αποθηκεύονται τα αποτελέσματα. Έτσι, καλείται πρώτα η διαδικασία CONTAINS από το πακέτο CTX_QUERY ως εξής :

```
ctx_query.contains (όνομα πολιτικής, συνθήκη, 'QUERY_TEMP')
```

όπου *όνομα πολιτικής* είναι το όνομα της πολιτικής (DEMO_POLICY ή DEMO_THEME_POLICY) και *συνθήκη* είναι η συνθήκη που εισήγαγε ο χρήστης. Η πρώτη παράμετρος της διαδικασίας είναι το όνομα της πολιτικής, ενώ η τρίτη είναι το όνομα του πίνακα που θα αποθηκευθούν τα αποτελέσματα (QUERY_TEMP στη συγκεκριμένη περίπτωση). Στη συνέχεια εκτελείται η δήλωση :

```
select count(*) from query_temp
```

για την ανάκτηση του αριθμού των αποτελεσμάτων. Πρέπει να σημειωθεί ότι η CONTAINS φροντίζει πρώτα να διαγράψει όλες τις εγγραφές του πίνακα QUERY_TEMP. Η επόμενη SQL δήλωση που εκτελείται είναι αυτή που ανακτά τις εγγραφές του πίνακα QUERY_TEMP, τα έγγραφα δηλαδή που ικανοποιούν τη συνθήκη του χρήστη :

```
select a.article_id a_id, a.author auth, a.section sect,
       b.score score, to_char(a.pub_date, 'dd/mm/yyyy') p_date,
       substr(a.title, 1, 100) || '...' title
from articles a, query_temp b
where to_number(b.textkey) = a.article_id order by score desc, a_id
```

Όπως φαίνεται, ανακτώνται ο αριθμός ταυτότητα του εγγράφου, το όνομα του συγγραφέα, ένα όνομα που είναι η κατηγορία στην οποία ανήκει το έγγραφο (για παράδειγμα νέα, οικονομικά, κτλ.), το σκορ που παράχθηκε κατά την εκτέλεση της CONTAINS, η ημερομηνία του εγγράφου, καθώς και ένα τμήμα του τίτλου (για να μην εμφανίζονται τεράστιοι τίτλοι), ταξινομημένα κατά φθίνουσα σειρά του σκορ κάθε εγγράφου και μετά με βάση τον αριθμό ταυτότητα.

Μέσα στο βρόχο που ανακτά κάθε εγγραφή καλείται η CTX_QUERY.HIGHLIGHT για τον τονισμό (*highlighting*) του όρου στο έγγραφο. Η κλήση αυτής της διαδικασίας απαιτεί δύο ενδιάμεσους πίνακες, τον HIGHLIGHT_TEMP και HIGHLIGHT_TABLE. Ο πρώτος πίνακας χρησιμοποιείται για την αποθήκευση της πληροφορίας για το ποια λέξη θα τονιστεί (χρησιμοποιώντας την απόσταση της λέξης μέσα στο κείμενο) και ο δεύτερος πίνακας είναι αυτός που περιέχει το έγγραφο και το τονισμένο κείμενο. Εδώ φαίνεται και η χρήση των παραμέτρων (αντικείμενο Parameter) για το ADO (λίστα κώδικα 5.3). Αρχικά, δημιουργείται ένα αντικείμενο (όνομα sqlHighlight) του τύπου Command, που θα καλεί την CTX_QUERY.HIGHLIGHT (αποθηκευμένη διαδικασία). Η ιδιότητα CommandType λαμβάνει την τιμή adCmdStoredProc ώστε να δείχνει ότι πρόκειται για αποθηκευμένη διαδικασία. Στη συνέχεια, γίνεται προσθήκη των απαραίτητων παραμέτρων της διαδικασίας στη συλλογή των παραμέτρων του αντικειμένου :

```
Set sqlHighlight=Server.CreateObject("ADODB.Command")
sqlHighlight.CommandType=adCmdStoredProc
sqlHighlight.CommandText="ctx_query.highlight"
sqlHighlight.Parameters.Append sqlHighlight.CreateParameter("cspec",
adVarChar, adParamInput, 30, PolicyName)
.
.
.
sqlHighlight.Execute
```

Λίστα κώδικα 5.3: Κλήση της ctx_query.highlight

Η μέθοδος `CreateParameter` δημιουργεί ένα αντικείμενο του τύπου `Parameter` και η `Parameters.Append` φροντίζει να προσθέσει το αντικείμενο αυτό στη συλλογή των πα-

Αριθμός αποτελεσμάτων : 2					
Έγγραφο	Σκορ	Συγγραφέας	Τίτλος	Κατηγορία	Ημερομηνία
11	11	David Stipp	Marketing & Media: Fanning and 2 Other Top Editors Quit Christian Science Monitor in a Dispute...	BUSINESS	15/12/1995
12	5	William M. Bulkeley	Technology: Honeywell Bull New Computer Called Fastest...	BUSINESS	15/12/1995

Εικόνα 5.7: Λίστα εγγράφων αποτελέσματος

ραμέτρων του αντικειμένου `sqlHighlight`. Το τονισμένο κείμενο φαίνεται μόνο στην περίπτωση που ο χρήστης κάνει κλικ στον αριθμό ταυτότητα του εγγράφου που όπως και στην περίπτωση της DB2 και εδώ αποτελεί ένα σύνδεσμο (Εικόνα 5.7). Αν ο χρήστης επιλέξει ένα έγγραφο θα οδηγηθεί στη σελίδα που δείχνει αναλυτικά τις πληροφορίες του εγγράφου που επέλεξε και με το κείμενο τονισμένο με έντονο κόκκινο χρώμα με τον όρο που εισήγαγε. Σε σχέση με μια ερώτηση θέματος, η μόνη διαφορά είναι στον τονισμό, όπου τονίζεται όχι μόνο ένας μεμονωμένος όρος, αλλά συνήθως μια ολόκληρη πρόταση η οποία θεωρείται ότι τον περιγράφει τον όρο του χρήστη. Κατά τα άλλα, δεν υπάρχουν διαφορές.

Στην περίπτωση που ο χρήστης αρχικά επιλέξει τις περιλήψεις/έννοιες των εγγράφων, τότε οδηγείται στη δεύτερη σελίδα που περιέχει μια λίστα με τα έγγραφα του πίνακα `articles` (Εικόνα 5.8). Αφού ο χρήστης επιλέξει το έγγραφο που επιθυμεί και πατήσει

1 - OTC Focus: Composite Index Falls to 5 1/2

Επιλέξτε από τη λίστα το έγγραφο και πατήστε "Εμφάνιση" για να παρουσιαστούν σε μορφή λίστας τα θέματα/ιδέες που αναπτύσσονται σε αυτό.

Εμφάνιση

Εικόνα 5.8: Λίστα εγγράφων του πίνακα `articles`

«Εμφάνιση» θα οδηγηθεί στην τρίτη σελίδα, όπου εμφανίζονται τα θέματα/έννοιες που αναπτύσσονται μέσα στο έγγραφο που επέλεξε. Σε αυτή τη σελίδα εκτελούνται αρκετές εργασίες από την πλευρά της Oracle για την εξαγωγή των θεμάτων/εννοιών. Αρχικά, καλείται η συνάρτηση `request_themes_func` του πακέτου `request_themes` με παράμετρο τον αριθμό ταυτότητα του εγγράφου που επέλεξε ο χρήστης. Η συνάρτηση επιστρέφει 1 για επιτυχή εκτέλεση και 0 για αποτυχημένη. Για τη διαδικασία εξαγωγής των θεμάτων/εννοιών χρησιμοποιούνται οι πίνακες `article_themes` και `article_gists`. Μέσα στη συνάρτηση, αρχικά γίνεται διαγραφή όλων των εγγράφων που περιέχουν οι πίνακες που αναγράφονται παραπάνω και στη συνέχεια καλείται η `set_completion_callback` του πακέτου `ctx_ling` (όπως φαίνεται στη λίστα κώδικα 5.4).

```
ctx_ling.set_completion_callback('request_themes.done_themes');
ctx_ling.request_themes('DEMO_POLICY', to_char(in_doc_id),
    'ARTICLE_THEMES');
ctx_ling.request_gist('DEMO_POLICY', to_char(in_doc_id),
    'ARTICLE_GISTS');
```

```
ctx_ling.submit;
```

Λίστα κώδικα 5.4: Διαδικασίες που εκτελούνται σε PL/SQL κώδικα μέσα στη συνάρτηση request_themes_func

Η διαδικασία αυτή ορίζει τη διαδικασία που θα κληθεί, όταν η εξαγωγή των θεμάτων/εννοιών ολοκληρωθεί από το διακομιστή ConText. Ο λόγος για τον οποίο πρέπει να κληθεί η `set_completion_callback` είναι ότι η κλήση των `request_themes` και `request_gist` του πακέτου `ctx_ling` επιστρέφουν τον έλεγχο κατευθείαν, χωρίς να έχει ολοκληρωθεί η διαδικασία εξαγωγής των θεμάτων/εννοιών. Δηλαδή η διαδικασία αυτή μπαίνει σε μια προσωρινή ουρά λόγω του γεγονότος ότι είναι χρονοβόρα και ίσως δεν είναι επιθυμητό να περιμένει ο χρήστης. Η διαδικασία `submit` προωθεί όλες τις εργασίες που έχουν τοποθετηθεί στην προσωρινή ουρά στην κύρια ουρά εξυπηρέτησης των διακομιστών ConText. Σε αυτό το σημείο, ο πρώτος διαθέσιμος διακομιστής ConText με ρόλο γλωσσολογικής επεξεργασίας αρχίζει την εξαγωγή των θεμάτων/εννοιών.

Επειδή όπως αναφέρθηκε και παραπάνω, όλες οι διαδικασίες που αναφέρονται παραπάνω επιστρέφουν τον έλεγχο απευθείας, πρέπει να υπάρχει κάποιος μηχανισμός ώστε να ολοκληρωθεί η διαδικασία εξαγωγής των θεμάτων/εννοιών και μετά να παρουσιαστεί η σελίδα με τα θέματα/έννοιες του εγγράφου στο χρήστη. Έτσι χρησιμοποιείται ένας μηχανισμός με πέρασμα μηνυμάτων και που υλοποιείται με τη χρησιμοποίηση τριών διαδικασιών του πακέτου `dbms_alert` της Oracle. Η πλήρης περιγραφή τους είναι εκτός της τρέχουσας εργασίας.

Το τελικό αποτέλεσμα είναι ότι ο χρήστης βλέπει με τη μορφή λίστας τα θέματα/έννοιες μαζί με το βάρος που είναι ένας πραγματικός αριθμός (με μέγιστο το εκατό), που αναπτύσσονται στο έγγραφο που επέλεξε αρχικά (Εικόνα 5.9). Τα θέματα τα οποία αναπτύσσονται ικανοποιητικά και έχουν παραγράφους που τα αναπτύσσουν, αποτελούν συνδέσμους που μπορεί να επιλέξει ο χρήστης και να δει το αντίστοιχο κομμάτι κειμένου που αναπτύσσει

Θέμα/ιδέα	Βάρος
stocks	9,41
Nasdaq	8,82
indexes	6,47
prices	5,88
rose	4,71

Εικόνα 5.9: Λίστα θεμάτων/εννοιών

το θέμα που επέλεξε. Η ανάκτηση των θεμάτων και των αντίστοιχων βαρών γίνεται με την SQL δήλωση :

```
select theme, round(weight, 2) weight
from article_normal
order by weight desc
```

όπου `article_normal` είναι μια όψη (*view*). Μάλιστα, η όψη αυτή βασίζεται σε μια άλλη όψη, στην `article_vectors`. Οι δύο όψεις ορίζονται ως εξής :

```
article_vectors : select pk,sum(weight) vector_weight
                  from article_themes
                  group by pk
article_normal  : select themes.pk pk,themes.theme theme,
                  (themes.weight/vector.vector_weight)*100 weight
                  from article_themes themes,article_vectors vector
                  where themes.pk=vector.pk
```

ΚΕΦΑΛΑΙΟ 6

ΕΠΙΛΟΓΟΣ

6.1 Γενική ανασκόπηση

Αρχικά η παρούσα εργασία έκανε μια εισαγωγή στα συστήματα ανάκτησης πληροφορίας. Έδωσε έναν ορισμό, εξέτασε τους τρόπους αξιολόγησης ενός ΣΑΠ, τις μετρικές της ακρίβειας και ανάκτησης, εξέτασε σε γενικές γραμμές τον τρόπο λειτουργίας ενός ΣΑΠ, έγινε μια συσχέτιση με τα συστήματα διαχείρισης βάσεων δεδομένων και τέλος έγινε μια αναφορά στις ψηφιακές βιβλιοθήκες και στις αποθήκες δεδομένων.

Στο δεύτερο κεφάλαιο εξετάστηκαν οι σύγχρονες μηχανές αναζήτησης του διαδικτύου. Δόθηκε ένας ορισμός, εξετάστηκαν ο τρόπος λειτουργίας και τα χαρακτηριστικά τους, έγινε διάκριση μεταξύ των εννοιών «μηχανή αναζήτησης» και «θεματικός κατάλογος» και υπήρξε μια κατηγοριοποίησή τους καθώς και αναφορά ενδεικτικών μηχανών αναζήτησης σε κάθε κατηγορία. Στη συνέχεια έγινε αναφορά στη βασική σύνταξη ερωτήσεων που υποστηρίζει το σύνολο των μηχανών και τέλος έγινε μια επισκόπηση του μεγέθους τόσο των μηχανών αναζήτησης όσο και των θεματικών καταλόγων, ενώ υπήρξε μια πρόβλεψη για το μέλλον τους.

Από το τρίτο κεφάλαιο και μετά, αρχίζει η εξέταση των δυνατοτήτων των σύγχρονων σχεσιακών συστημάτων διαχείρισης βάσεων δεδομένων DB2 και Oracle σε σχέση με τις δυνατότητες ευρετηριοποίησης ελεύθερου ή ημι-δομημένου κειμένου και ανάκτησης πληροφορίας από αυτό. Έτσι, το τρίτο κεφάλαιο εξέτασε την προσθήκη για την DB2, τον Text Extender. Εξετάστηκε τόσο από τεχνικής απόψεως, πώς υπάρχει σε ένα περιβάλλον διακομιστή DB2, πώς γίνεται η ενεργοποίηση ενός στιγμιότυπου βάσης DB2 για χρησιμοποίησή του από τον Extender (διαχειριστικές εργασίες), όσο και από θεωρητικής απόψεως, πώς γίνεται η συντακτική ανάλυση του κειμένου και πώς αξιοποιούνται οι δυνατότητές του μέσω SQL ερωτήσεων.

Το τέταρτο κεφάλαιο εξέτασε το ConText cartridge της Oracle. Όπως και στο τρίτο κεφάλαιο, έτσι και στο κεφάλαιο αυτό έγινε εξέταση του cartridge τόσο από τεχνικής απόψεως, πώς συνεργάζεται ο διακομιστής ConText με τους διακομιστές της βάσης, πώς γίνεται ενεργοποίηση του cartridge για τη χρησιμοποίησή του, πώς γίνεται εκκίνηση και τερματισμός των διακομιστών ConText, όσο και από θεωρητικής απόψεως, περιγράφονται τα είδη των ερωτήσεων και οι τελεστές που μπορούν να χρησιμοποιηθούν στις SQL ερωτήσεις, καθώς και οι συναρτήσεις και διαδικασίες που χρησιμοποιούνται για την αναζήτηση.

Στο πέμπτο κεφάλαιο περιγράφηκε το πρότυπο περιβάλλον ιστοσελίδων που δημιουργήθηκε για την επίδειξη των δυνατοτήτων του Text Extender και του ConText cartridge. Έγινε σύντομη αναφορά στην τεχνολογία των asp σελίδων, στη συνέχεια περιγράφηκε το μοντέλο ADO και τέλος έγινε περιγραφή των ιστοσελίδων.

6.2 Δυσκολίες και προβλήματα

Κατά τη διάρκεια εκπόνησης της εργασίας, ο γράφων συνάντησε ορισμένες δυσκολίες ή προβλήματα που περιγράφονται παρακάτω.

Όσον αφορά στις asp σελίδες, τα αντικείμενα τύπου Recordset (σετ εγγραφών) έχουν την ιδιότητα `RecordCount` που επιστρέφει τον αριθμό των εγγραφών του σετ, αφού έχει ανοιχθεί. Ωστόσο στην πράξη, η ιδιότητα `RecordCount` ενός σετ εγγραφών επέστρεφε -1 που σύμφωνα με την τεκμηρίωση του Option Pack για το μοντέλο ADO σημαίνει ότι για κάποιο λόγο δεν μπορεί να επιστραφεί ο αριθμός των εγγραφών του σετ. Λόγω αυτού του γεγονότος, εκτελούνταν μια SQL δήλωση του τύπου `select count(*) from ...` για την ανάκτηση του αριθμού των εγγραφών. Αυτό ίσχυε τόσο για τη DB2, όσο και για την Oracle και ίσως μπορεί να εξηγηθεί με το γεγονός ότι οι ODBC οδηγούς για τις συ-

γκεκριμένες βάσεις δεν υποστήριζαν αυτή την ιδιότητα, γιατί οι ODBC οδηγοί δεν προσφέρουν όλες τις δυνατότητες και ιδιότητες που αναφέρει το μοντέλο ADO.

Παράλληλα, ο ODBC οδηγός που προσφέρει η Oracle με την έκδοση 8.0.5 εμφανίζει σημαντικά προβλήματα. Κατά τη διάρκεια δημιουργίας δοκιμαστικών σελίδων που εμφάνιζαν εγγραφές ενός πίνακα της βάσης, η σελίδα ήταν κενή διότι ο οδηγός ODBC δεν επέστρεφε εγγραφές! Χρησιμοποιώντας τον ODBC οδηγό που προσφέρει η Microsoft, η ίδια σελίδα δούλεψε κανονικά. Αναβάθμιση του ODBC οδηγού της Oracle στην έκδοση 8.0.5.8 έλυσε το συγκεκριμένο πρόβλημα, αλλά υπήρξαν προβλήματα όσον αφορά στην υποβολή των ConText ερωτήσεων, συγκεκριμένα ο συντακτικός αναλυτής του ConText εμφάνιζε λάθος. Τελικά αποφασίστηκε η χρησιμοποίηση του οδηγού που δίνει η Microsoft για τη σύνδεση με Oracle, έκδοση 2.573.4202.00, ο οποίος δεν (φαίνεται) να εμφανίζει προβλήματα.

Όσον αφορά το ConText, ο συντακτικός αναλυτής εμφάνιζε λάθος όταν η ερώτηση περιλάμβανε το λογικό τελεστή NOT, κάτι το οποίο δεν έπρεπε να συμβαίνει. Για αυτό το λόγο και μόνο, στις σελίδες για το ConText, δεν εμφανίζεται ο τελεστής NOT μαζί με τους άλλους λογικούς τελεστές. Δεν βρέθηκε λύση για το πρόβλημα αυτό.

Ένα άλλο σχετικό πρόβλημα με το ConText είναι ότι η τεκμηρίωση αναφέρει ότι ο τελεστής WITHIN υποστηρίζει τη σύνταξη WITHIN SENTENCE/PARAGRAPH για αναζήτηση μέσα στην ίδια πρόταση ή παράγραφο. Στην πράξη ωστόσο, δεν αναγνωρίζονταν οι λέξεις κλειδιά SENTENCE και PARAGRAPH, αλλά το ConText νόμιζε ότι γινόταν αναφορά σε προκαθορισμένα τμήματα και εμφάνιζε μήνυμα λάθους ότι δεν υπάρχουν (και πολύ σωστά) τέτοια προκαθορισμένα τμήματα. Έτσι δεν χρησιμοποιήθηκε η συγκεκριμένη δυνατότητα.

6.3 Εντυπώσεις

Ο Text Extender της DB2 είναι αρκετά αυστηρός στην ανάλυση της ερώτησης. Οι όροι πρέπει να περικλείονται με διπλά εισαγωγικά μέσα στη συμβολοσειρά τύπου LONG VARCHAR, διαφορετικά εμφανίζει λάθος. Αντίθετα, το ConText cartridge της Oracle είναι λιγότερο αυστηρό, οι όροι μπορεί να περικλείονται ή όχι με μονά εισαγωγικά (') και για το λόγο αυτό υπάρχει περισσότερη ελευθερία από την πλευρά του χρήστη για τη διατύπωση πολυπλοκότερων ερωτήσεων.

Η Oracle προσφέρει καλύτερη και αναλυτικότερη τεκμηρίωση σε σχέση με εκείνη της DB2, όσον αφορά στο ConText και Text Extender αντίστοιχα. Από την άλλη, η IBM προσφέρει κατευθείαν πίνακες παραδείγματα για να ξεκινήσει κανείς με τον Extender, ενώ η Oracle υστερεί σε σχέση με τη DB2 στο συγκεκριμένο τομέα, χωρίς να τα προσφέρει όλα από την αρχή και να πρέπει ο χρήστης να επέμβει διότι από το χρήστη επίδειξης ctdemo λείπουν ορισμένα στοιχεία.

Όσον αφορά στις δυνατότητες, η Oracle φαίνεται να είναι δυνατότερη, μιας και προσφέρει μια πληθώρα διαδικασιών και συναρτήσεων και ορισμένους δυνατούς τελεστές, για παράδειγμα υποστηρίζει τις αποθηκευμένες ερωτήσεις, καθώς επίσης και τρεις διαφορετικούς τύπους εκτέλεσης των ερωτήσεων, ο καθένας με τα μειονεκτήματα και πλεονεκτήματά του. Βέβαια, ίσως δεν μπορεί να γίνει απευθείας σύγκριση μεταξύ των δυο βάσεων, από τη στιγμή που η DB2 δεν παρέχει μια γλώσσα προγραμματισμού για τη δημιουργία αποθηκευμένων διαδικασιών, όπως η PL/SQL της Oracle, που λύνει τα χέρια του προγραμματιστή. Το γεγονός βέβαια ότι το ConText παρέχει περισσότερες δυνατότητες σε σχέση με τον Text Extender σημαίνει και ότι είναι πολυπλοκότερο προγραμματιστικά. Επίσης, η δυνατότητα ύπαρξης πολλών διακομιστών ConText με τον ίδιο ή διαφορετικό ρόλο φαίνεται να προσφέρει καλύτερες πιθανότητες ανάκαμψης.

Σχετικά με τις asp σελίδες, οι εντυπώσεις είναι θετικές. Αρνητικό είναι το γεγονός ότι η τεχνολογία αυτή είναι της Microsoft και μόνο, με αποτέλεσμα να προσανατολίζεται μόνο σε δικά της προϊόντα, για παράδειγμα asp σελίδες υποστηρίζει μόνο ο Internet Information Server της Microsoft. Επίσης, η ενσωματωμένη υποστήριξη των γλωσσών σεναρίων VBScript και JScript βοηθά τη μετάβαση των προγραμματιστών που είναι ήδη εξοικειωμένοι με τη Visual Basic και τη JavaScript αντίστοιχα, στην τεχνολογία των asp σελίδων. Ωστόσο, ο γράφων θα προτιμούσε ένα πιο ευέλικτο τρόπο χειρισμού των λαθών για τη VBScript.

Όσον αφορά στο προγραμματιστικό μοντέλο ADO, δείχνει να είναι σε καλό δρόμο για την ευρεία χρησιμοποίησή του για συνδέσεις με βάσεις δεδομένων, καθώς είναι απλό, χωρίς να μπερδεύει τον προγραμματιστή με μεγάλο αριθμό αντικειμένων.

Τέλος, όσον αφορά στο γράφοντα της εργασίας, οι εμπειρίες που αποκόμισε υπήρξαν πολύ καλές. Ασχολήθηκε με ποικίλα ζητήματα, τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο. Εμπειρίες οι οποίες θα αποδειχθούν πολύτιμες για το υπόλοιπο της καριέρας του στην πληροφορική.

ΠΑΡΑΡΤΗΜΑ

Διαχειριστικές εντολές του Text Extender για τον πελάτη

Η κονσόλα εντολών για τον Text Extender ξεκινά από το εικονίδιο DB2TX Command Line Processor. Παρακάτω γίνεται αναφορά στο τι επιτελεί κάθε εντολή, για την πλήρη σύνταξη κάθε εντολής, ο ενδιαφερόμενος πρέπει να ανατρέξει στο έβδομο κεφάλαιο του *Text Extender Administration And Programming*.

CHANGE INDEX SETTINGS

Αλλάζει τη συχνότητα ενημέρωσης ενός ευρετηρίου, αφού η βάση έχει ενεργοποιηθεί, και του ελάχιστου αριθμού εγγράφων που μπαίνουν στην ουρά για ενημέρωση.

CHANGE TEXT CONFIGURATION

Αλλάζει τις εξ' ορισμού ρυθμίσεις του κειμένου όταν ενεργοποιείται μια βάση. Μπορεί να αλλάχθει ο τύπος ευρετηρίου, η συχνότητα ενημέρωσής του, ο κατάλογος που θα αποθηκεύονται τα αρχεία του, μια παράμετρος που δηλώνει αν το ευρετήριο θα ενημερώνεται ρητά με εντολή του χρήστη ή όχι, ο αριθμός εγγραφών μετά τον οποίο θα γίνεται οριστικοποίηση στη βάση, ο κωδικός CCSID του κειμένου, η γλώσσα, καθώς και η μορφοποίηση του κειμένου.

CONNECT

Με την `CONNECT` γίνεται σύνδεση με τη βάση. Αν δοθεί μια άλλη εντολή χωρίς να έχει ανοιχθεί σύνδεση με τη βάση, τότε αυτόματα γίνεται σύνδεση στη βάση που με όνομα αυτό που έχει η μεταβλητή περιβάλλοντος `DB2DBDFT`.

DELETE INDEX EVENTS

Διαγράφει τα γεγονότα ευρετηριοποίησης από τον πίνακα ημερολογίου ενός ευρετηρίου.

DISABLE DATABASE

Απενεργοποιεί μια βάση για χρήση από τον Text Extender. Απενεργοποιεί όλους τους πίνακες που είχαν ευρετηριοποιηθεί, καταργούνται η όψη `TEXTCOLUMNS`, οι σκανδάλες, καθώς και όλα τα ευρετήρια που χρησιμοποιούσε ο Text Extender.

DISABLE TEXT COLUMN

Απενεργοποιεί μια στήλη κειμένου για χρήση από τον Text Extender. Καταργούνται το ευρετήριο για τη στήλη, ο πίνακας ημερολογίου οι σκανδάλες που ενημέρωναν τον πίνακα ημερολογίου.

DISABLE TEXT FILES

Απενεργοποιεί την ευρετηριοποίηση εξωτερικών αρχείων κειμένου. Καταργεί το ευρετήριο και τις σκανδάλες που ενημερώνανε τον πίνακα ημερολογίου.

DISABLE TEXT TABLE

Απενεργοποιεί ένα πίνακα για χρήση από τον Text Extender. Αν οι στήλες κειμένου του πίνακα είχαν ενεργοποιηθεί με την εντολή `ENABLE TEXT COLUMN`, διαγράφονται όλα τα αντίστοιχα ευρετήρια. Αν ο πίνακας διατηρούσε κοινό ευρετήριο για όλες τις στή-

λες, ο πίνακας δηλαδή είχε ενεργοποιηθεί με την `ENABLE TEXT TABLE`, τότε καταργείται το κοινό ευρετήριο. Επίσης καταργείται ο πίνακας ημερολογίου, καθώς και οι αντίστοιχες σκανδάλες που ενημερώνανε τον πίνακα.

ENABLE DATABASE

Ενεργοποιεί μια βάση για χρησιμοποίηση από τον Text Extender και εμφανίζει πληροφορίες σχετικά με τους πίνακες, τους ορισμένους από το χρήστη τύπους, διαδικασίες και συναρτήσεις. Επίσης, δημιουργούνται οι όψεις `TEXTINDEXES` και `TEXTCOLUMNS`.

ENABLE TEXT COLUMN

Ενεργοποιεί μια στήλη για χρησιμοποίηση από τον Text Extender. Δέχεται παραμέτρους εκείνες που περιγράφηκαν και στην `CHANGE TEXT CONFIGURATION`. Προσθέτει το χειριστή για τη στήλη. Αν ο πίνακας δεν έχει προηγουμένως ενεργοποιηθεί, τότε ενεργοποιείται και δημιουργείται το ευρετήριο για τη στήλη. Επίσης, δημιουργείται ο πίνακας ημερολογίου, καθώς και οι σκανδάλες που ενημερώνουν τον πίνακα αυτό.

ENABLE TEXT FILES

Ενεργοποιεί την αναζήτηση σε εξωτερικά αρχεία κειμένου. Δέχεται τις παραμέτρους που δέχεται και η `ENABLE TEXT COLUMN` και προσθέτει το χειριστή για τη στήλη που δίνεται ως παράμετρος. Επίσης, δημιουργεί τον πίνακα ημερολογίου, αλλά ο χρήστης πρέπει να δημιουργήσει μόνος του τις σκανδάλες που ενημερώνουν τον πίνακα, διότι η DB2 δεν έχει τον έλεγχο των αρχείων και δεν μπορεί να ξέρει πότε έχει συμβεί μια αλλαγή σε αυτά. Αφού δοθεί η εντολή αυτή, στη συνέχεια πρέπει να αρχικοποιηθούν οι χειριστές με την εντολή `INIT_TEXT_HANDLE`.

ENABLE TEXT TABLE

Δημιουργεί ένα κοινό ευρετήριο για όλες τις στήλες κειμένου ενός πίνακα. Αν δεν δοθεί αυτή η εντολή και γίνει ενεργοποίηση καθεμιάς στήλης με την `ENABLE TEXT COLUMN`, τότε ο πίνακας έχει πολλαπλά ευρετήρια. Δημιουργείται επίσης και ο πίνακας ημερολογίου, καθώς και οι αντίστοιχες σκανδάλες.

GET ENVIRONMENT

Εμφανίζει τις τιμές των μεταβλητών περιβάλλοντος `DB2INSTANCE`, `DB2DBDFT`, `DB2TX_INSTOWNER` και `DB2TX_INSTOWNERHOMEDIR` (οι δύο τελευταίες ισχύουν μόνο για Unix συστήματα).

GET INDEX SETTINGS

Εμφανίζει τις ρυθμίσεις ενός ευρετηρίου, ο τύπος του καθώς και με ποια επιλογή έχει δημιουργηθεί, η συχνότητα ενημέρωσής του και ο κατάλογος αποθήκευσης των αρχείων του ευρετηρίου.

GET INDEX STATUS

Εμφανίζει τις εξής πληροφορίες σχετικά με ένα ευρετήριο : την κατάσταση αναζήτησης (αν μπορεί να γίνει αναζήτηση), την κατάσταση ενημέρωσης (αν μπορεί να ενημερωθεί το ευρετήριο), τον αριθμό των εκκρεμών προς ευρετηριοποίηση κειμένων, τον αριθμό των ήδη ευρετηριοποιημένων εγγράφων, καθώς και τυχόν λάθη.

GET STATUS

Εμφανίζει την κατάσταση της βάσης, αν είναι ενεργοποιημένη ή όχι, καθώς και τους πίνακες που είναι ενεργοποιημένοι, μαζί με το αν περιλαμβάνουν κοινό ή ξεχωριστό ευρετήριο για κάθε στήλη.

GET TEXT CONFIGURATION

Εμφανίζει τις εξ' ορισμού ρυθμίσεις για το κείμενο, το CCSID, τη γλώσσα, τη μορφοποίηση, το είδος ευρετηρίου, τη συχνότητα ενημέρωσης, τον κατάλογο αποθήκευσης των αρχείων ευρετηρίων, καθώς και τον αριθμό των εγγραφών για οριστικοποίηση.

GET TEXT INFO

Εμφανίζει το CCSID, τη γλώσσα και τη μορφοποίηση του κειμένου για ένα χειριστή.

QUIT

Τερματίζει την κονσόλα εντολών του Text Extender.

RESET INDEX STATUS

Όταν η κατάσταση ενός ευρετηρίου δείχνει 'Search not available' ή 'Update not available', έχει προκύψει λάθος κατά την ευρετηριοποίηση, που αποτρέπει την χρησιμοποίηση του ευρετηρίου. Η εντολή αυτή επαναφέρει την κατάσταση του ευρετηρίου στο κανονικό, ώστε να μπορεί να συνεχιστεί η εργασία. Πριν γίνει αυτό, πρέπει να γίνει έλεγχος πρώτα των μηνυμάτων λαθών που υπάρχουν στον πίνακα ημερολογίου.

UPDATE INDEX

Δίνει εντολή για την ευρετηριοποίηση. Μετά το πέρας της εντολής, το ευρετήριο έχει ευρετηριοποιήσει όλα τα κείμενα της στήλης με την οποία έχει συσχετιστεί.

Διαχειριστικές εντολές του Text Extender για τον διακομιστή

Οι εντολές που περιγράφονται παρακάτω αποτελούν τις διαχειριστικές εντολές του Text Extender για τον διακομιστή. Δίνονται από το παράθυρο εντολών των Windows NT. Για την πλήρη περιγραφή των εντολών, ο ενδιαφερόμενος πρέπει να ανατρέξει στο ένατο κεφάλαιο του *Text Extender Administration And Programming*.

TXICRT

Δημιουργεί ένα στιγμιότυπο του Text Extender. Απαιτεί την ύπαρξη ενός στιγμιότυπου της DB2.

TXIDROP

Καταργεί ένα στιγμιότυπο του Text Extender μαζί με όλα τα ευρετήρια που περιλαμβάνει. Πρέπει να δοθεί αφού έχει απενεργοποιηθεί η βάση για χρησιμοποίηση από τον Text Extender.

TXNADD

Δημιουργεί ένα πρόσθετο διακομιστή Text Extender. Όταν δοθεί η TXSTART, τότε αυτόματα εκκινεί το νέο διακομιστή.

TXNCHECK

Ελέγχει για την ακεραιότητα του Text Extender στη βάση.

TXNDROP

Διαγράφει ένα διακομιστή Text Extender.

TXSAMPLE

Δημιουργεί τους πίνακες επίδειξης, τους φορτώνει με δεδομένα (έγγραφα στην αγγλική γλώσσα) και ενεργοποιεί τις αντίστοιχες στήλες κειμένου των πινάκων αυτών.

TXSTART

Εκκινεί τους διακομιστές Text Extender.

TXSTATUS

Εμφανίζει την κατάσταση των διακομιστών Text Extender, αν έχουν ξεκινήσει ή όχι.

TXSTOP

Σταματά τους διακομιστές Text Extender.

TXTRACE

Χρησιμοποιείται για τη δημιουργία πληροφορίας *ανίχνευσης (trace)*, που μπορεί να χρησιμοποιηθεί για την ανίχνευση σφαλμάτων. Η πληροφορία αυτή γράφεται σε αρχείο είτε σε δυαδική μορφή, είτε σε μορφή κειμένου. Οι παράμετροι που δέχεται σχετικά με την ανίχνευση είναι το είδος των γεγονότων που θα καταγράφονται, καθώς και η πηγή, για παράδειγμα μια αποθηκευμένη διαδικασία, μια συνάρτηση, από τη συντακτική ανάλυση του Text Extender, κτλ.

TXVERIFY

Δημιουργεί μια βάση επίδειξης και την ενεργοποιεί για χρήση από τον Text Extender.

ΓΛΩΣΣΑΡΙΟ

Αγγλικό

A

Access control list
Add-on
Alias
Attribute

Λίστα ελέγχου πρόσβασης
Προσθήκη
Όνομα είδωλο
Χαρακτηριστικό

B

Batch
Browser

Μαζικός/ή/ό
Πρόγραμμα πλοήγησης

C

Cache memory
Client
Collection
Command line interpreter
Command line tool
Command window
Commit
Configuration
Cursor

Λανθάνουσα μνήμη
Πελάτης
Συλλογή
Μεταφραστής γραμμής εντολών
Εργαλείο γραμμής εντολών
Παράθυρο εντολών
Οριστικοποιώ, οριστικοποίηση
Παραμετροποίηση
Δρομέας

D

Data mining
Data warehouse
Demo
Descriptor
Drop
Dual

Εξόρυξη δεδομένων
Αποθήκη δεδομένων
Επίδειξη
Περιγραφικός
Καταργώ
Διπλός/ή/ό

E

Event
Exact match
Expansion
Extended
Extension

Γεγονός
Πλήρης ταύτιση
Επαύξηση
Εκτεταμένος/η/ο
Επέκταση

F

Feature extraction
Format
Function

Εξαγωγή χαρακτηριστικών
Μορφοποίηση, μορφοποιώ
Συνάρτηση

G

H

Hit list

Λίστα επιτυχιών

I	
Index	Ευρετήριο, ευρετηριοποιώ
Information Retrieval System	Σύστημα ανάκτησης πληροφορίας
Initialized	Αρχικοποιημένος/η/ο
Instance	Στιγμιότυπο
Interactive	Διαδραστικός/ή/ό
Item	Μονάδα πληροφορίας
J	
Join	Σύζευξη
K	
L	
Linguistic	Γλωσσολογικός
Link	Σύνδεσμος
Loader	Φορτωτής
M	
Matching	Ταύτιση
Method	Μέθοδος
Module	Επέκταση
N	
Normalization	Κανονικοποίηση
O	
Object oriented	Αντικειμενοστραφής
Optical Character Recognition Program	Πρόγραμμα οπτικής αναγνώρισης χαρακ- κτήρων
Overhead	Κόστος επεξεργασίας
P	
Parsing	Συντακτική ανάλυση
Partial match	Μερική ταύτιση
Partition	Διαμερισμός, διαμερίζω
Pattern	Σχέδιο
Pointer	Δείκτης
Policy	Πολιτική
Precise	Ακριβές
Precision	Ακρίβεια
Preference	Επιλογή
Procedural	Διαδικασιακός/ή/ό
Procedure	Διαδικασία
Proper name	Κύριο όνομα
Property	Ιδιότητα
Proximity	Εγγύτητα
Q	

Queue	Ουρά εξυπηρέτησης
R	
Rank	Διαβάθμιση
Recall	Ανάκτηση
Record set	Σετ εγγραφών
Refine	Τροποποιώ
Regular expression	Κανονική έκφραση
Relation	Σχέση
Relevant	Σχετιζόμενος/η/ο
Resources	Πόροι
Role	Ρόλος
S	
Scope	Εμβέλεια
Script	Σενάριο
Scripting language	Γλώσσα σεναρίου
Selective dissemination of information	Επιλεκτική διανομή της πληροφορίας
Server	Διακομιστής
Service	Υπηρεσία
Session	Σύνοδος
Stemming	Αποκοπή κοινότυπων καταλήξεων
Stored procedure	Αποθηκευμένη διαδικασία
String	Συμβολοσειρά
T	
Tag	Παραπομπή
Term	Όρος
Thesaurus	Θησαυρός
Thread	Νήμα
Threshold	Κατώφλι
Token	Τεκμήριο
Trace	Ανίχνευση
Transaction	Συναλλαγή
U	
Utility	Βοήθημα
V	
View	Όψη
W	
Web page	Ιστοσελίδα
Web site	Τόπος
Wildcard character	Χαρακτήρας μάσκα
Word	Λέξη
X	
Y	

Z

Ελληνικό

A

Ακρίβεια	Precision
Ακριβές	Precise
Ανάκτηση	Recall
Ανίχνευση	Trace
Αντικειμενοστραφής	Object oriented
Αποθηκευμένη διαδικασία	Stored procedure
Αποθήκη δεδομένων	Data warehouse
Αποκοπή κοινότυπων καταλήξεων	Stemming
Αρχικοποιημένος/η/ο	Initialized

B

Βοήθημα	Utility
---------	---------

Γ

Γεγονός	Event
Γλώσσα σεναρίου	Scripting language
Γλωσσολογικός	Linguistic

Δ

Δείκτης	Pointer
Διαβάθμιση	Rank
Διαδικασία	Procedure
Διαδικασιακός/ή/ό	Procedural
Διαδραστικός/ή/ό	Interactive
Διακομιστής	Server
Διαμερισμός, διαμερίζω	Partition
Διπλός/ή/ό	Dual
Δρομέας	Cursor

Ε

Εγγύτητα	Proximity
Εκτεταμένος/η/ο	Extended
Εμβέλεια	Scope
Εξαγωγή χαρακτηριστικών	Feature extraction
Εξόρυξη δεδομένων	Data mining
Επαύξηση	Expansion
Επέκταση	Module, Extension
Επίδειξη	Demo
Επιλεκτική διανομή της πληροφορίας	Selective dissemination of information
Επιλογή	Preference
Εργαλείο γραμμής εντολών	Command line tool
Ευρετήριο, ευρετηριοποιώ	Index

Z

H

Θ

Θησαυρός

Thesaurus

I

Ιδιότητα

Property

Ιστοσελίδα

Web page

K

Κανονική έκφραση

Regular expression

Κανονικοποίηση

Normalization

Καταργώ

Drop

Κατώφλι

Threshold

Κόστος επεξεργασίας

Overhead

Κύριο όνομα

Proper name

Λ

Λανθάνουσα μνήμη

Cache memory

Λέξη

Word

Λίστα ελέγχου πρόσβασης

Access control list

Λίστα επιτυχιών

Hit list

M

Μαζικός/ή/ό

Batch

Μέθοδος

Method

Μερική ταύτιση

Partial match

Μεταφραστής γραμμής εντολών

Command line interpreter

Μονάδα πληροφορίας

Item

Μορφοποίηση, μορφοποιώ

Format

N

Νήμα

Thread

Ξ

O

Όνομα είδωλο

Alias

Οριστικοποιώ, οριστικοποίηση

Commit

Όρος

Term

Ουρά εξυπηρέτησης

Queue

Όψη

View

Π

Παράθυρο εντολών

Command window

Παραμετροποίηση

Configuration

Παραπομπή

Tag

Πελάτης

Client

Περιγραφικός	Descriptor
Πλήρης ταύτιση	Exact match
Πολιτική	Policy
Πόροι	Resources
Πρόγραμμα οπτικής αναγνώρισης χαρα- κτῆρων	Optical Character Recognition Program
Πρόγραμμα πλοήγησης	Browser
Προσθήκη	Add-on

P

Ρόλος	Role
-------	------

Σ

Σενάριο	Script
Σετ εγγραφών	Record set
Στιγμιότυπο	Instance
Σύζευξη	Join
Συλλογή	Collection
Συμβολοσειρά	String
Συναλλαγή	Transaction
Συνάρτηση	Function
Σύνδεσμος	Link
Σύνοδος	Session
Συντακτική ανάλυση	Parsing
Σύστημα ανάκτησης πληροφορίας	Information Retrieval System
Σχέδιο	Pattern
Σχέση	Relation
Σχετιζόμενος/η/ο	Relevant

T

Ταύτιση	Matching
Τεκμήριο	Token
Τόπος	Web site
Τροποποιώ	Refine

Υ

Υπηρεσία	Service
----------	---------

Φ

Φορτωτής	Loader
----------	--------

X

Χαρακτήρας μάσκα	Wildcard character
Χαρακτηριστικό	Attribute

Ψ

Ω

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Σχήματα Ψηφιακών Υπογραφών για τη Διαχείριση Ελευθέρου Κειμένου, διδακτορική διατριβή Δημητρίου Δέρβου, Θεσσαλονίκη, Ιανουάριος 2000
2. Information Retrieval, Algorithms And Heuristics, David Grossman, Ophir Frieder, Kluwer Academic Publishers, 1998
3. Information Retrieval Systems, Theory And Implementation, Gerald Kowalski, Kluwer Academic Publishers, 1997
4. IBM DB2 Text Extender v5.2 Administration and Programming
5. IBM DB2 v5.2 SQL Reference
6. Oracle ConText cartridge QuickStart Release 2.4
7. Oracle ConText cartridge Application Developer's Guide Release 2.4
8. Oracle ConText cartridge Administrator's Guide Release 2.4
9. Oracle v8.0.5 SQL Reference
10. Oracle v8.0.5 PL/SQL User's Guide and Reference
11. Search Engine Watch : Tips about Internet Search Engines & Search Engine Submission (internet web site at : <http://searchenginewatch.com/>)
12. Internet tutorials (internet web site at : <http://www.albany.edu/library/internet/>)
13. Σύστημα καταχώρησης και αρχειοθέτησης πτυχιακών εργασιών του τμήματος Πληροφορικής, πτυχιακή Φεύγα Αθανάσιου, Ιανουάριος 2000, Θεσσαλονίκη
14. Windows NT 4 Option Pack Documentation