ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ & ΔΙΟΙΚΗΣΗΣ

ΜΕΛΕΤΗ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ ΣΤΟ ΔΙΑΔΙΚΤΥΟ ΚΑΘΩΣ ΚΑΙ ΤΩΝ ΤΕΧΝΙΚΩΝ ΤΟΥΣ - ΑΝΑΠΤΥΞΗ ΕΝΟΣ ΜΟΝΤΕΛΟΥ - ΠΡΟΤΥΠΟΥ ΕΝΙΑΙΑΣ ΑΝΑΖΗΤΗΣΗΣ

Διατριβή που υπεβλήθη για τη μερική ικανοποίηση των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης

υπό

ΓΕΩΡΓΑΚΗ ΚΩΝΣΤΑΝΤΙΝΟ

ΧΑΝΙΑ, 2004

Η διατριβή του Γεωργάκη Κωνσταντίνου εγκρίνεται:

- ✓ ***Αναπληρωτής Καθηγητής Ματσατσίνης Νικόλαος*** **……..**

- ✓ ***Καθηγητής Ζοπουνίδης Κωνσταντίνος*** **……..**

- ✓ ***Αναπληρωτής Καθηγητής Μυγδαλάς Αθανάσιος*** **……..**

ΠΕΡΙΕΧΟΜΕΝΑ

'αφιερωμένη στους πολυαγαπημένους μου γονείς'

"Η Επιστήμη Κάνει τον Σοφό
Και η Λογική τον Άνθρωπο"
Λαϊκή Μούσα

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ευκαιρία της παρούσας διατριβής θα θελα να ευχαριστήσω ιδιαίτερα τον επιβλεποντά Καθηγητή μου κ. Ματσατσίνη Νικόλαο για την πολύτιμη βοήθεια και τη συνεχή καθοδήγησή του. Τον ευχαριστώ και για τον χρόνο που διέθεσε όλον αυτό τον καιρό, προκειμένου να έρθει εις πέρας η παρούσα μεταπτυχιακή διατριβή με τον καλύτερο δυνατό τρόπο.

Επίσης θεωρώ χρέος μου να ευχαριστήσω τους καθηγητές Μυγδαλά Α. και Ζοπουνίδη Κ. μέλη της τριμελούς μου επιτροπής.

Ένα πολύ μεγάλο και θερμό ευχαριστώ στους γονείς μου και τον αδερφό μου για την αμέριστη συμπαράσταση και υποστήριξή τους τόσο ψυχολογική όσο και υλική καθ' όλη τη διάρκεια των σπουδών μου.

Θα ήθελα τέλος να ευχαριστήσω όλους τους φίλους μου που με βοήθησαν να ξεπεράσω τις δύσκολες καταστάσεις κατά τη διάρκεια της εκπόνησης της παρούσας εργασίας. Ιδιαίτερα οφείλω ένα πολύ μεγάλο ευχαριστώ στις αγαπητές μου φίλες Μαρία Σταμπουλή και τη Σοφία Περογιαννάκη, που στις δύσκολες στιγμές ήταν αυτές που πάντα με στήριζαν και με γέμιζαν με κουράγιο και αυτοπεποίθηση για την ολοκλήρωση της παρούσας διατριβής.

Κώστας

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΗΛΕΚΤΡΟΝΙΚΟΣ ΜΗΧΑΝΙΚΟΣ ΚΑΙ ΜΗΧΑΝΙΚΟΣ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΠΩΝΥΜΟ:	**Γεωργάκης**
ΟΝΟΜΑ:	**Κωνσταντίνος**
ΗΜΕΡΟΜΗΝΙΑ ΓΕΝΝΗΣΗΣ:	**25 Δεκεμβρίου 1977**
E-mail:	**kgeorgakis@hotmail.com**
ΜΟΝΙΜΗ ΔΙΕΥΘΥΝΣΗ:	**Πλάτωνος 21** **Ανατολή Ιωαννίνων**
ΤΡΕΧΟΥΣΑ ΔΙΕΥΘΥΝΣΗ:	**Καποδιστρίου 73** **Χανιά- Κρήτης**
ΤΗΛΕΦΩΝΟ:	**26510- 68288** **Κιν. 697-4804861**

Σπουδές

Οκτώβριος 2001 – Σήμερα:	***Μεταπτυχιακός Φοιτητής του Τμήματος Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, στον τομέα Οργάνωσης και Διοίκησης (M.Sc.),*** *Χανιά*
Σεπτεμβριος 2001:	***Εισαγωγή στο επιδοτούμενο Μεταπτυχιακό Πρόγραμμα του Πανεπιστημίου Ιωαννίνων στο τμήμα Πληροφορικής,*** *Ιωάννινα*
Οκτώβριος 1995 – Ιούνιος 2001:	***Δίπλωμα Ηλεκτρονικού Μηχανικού και Μηχανικού Ηλεκτρονικών Υπολογίστων του Πολυτεχνείου Κρήτης, τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών,*** *Χανιά* *Βαθμός Πτυχίου: 7.61 (*Λιαν Καλώς) *Εκπόνηση διπλωματικής εργασίας με τίτλο: "Αντιμετώπιση ISI (Inter- Symbol Inference) σε Δίαυλο Σταθερών Παραμέτρων Παρουσία μη- Γκαουσιανού Θορύβου CLASS A (Non Gaussian Noise) "*
1993- 1995:	***Λύκειο Mannheim (Γερμανία)*** *Βαθμός Απολυτηρίου: 20 (*Άριστα)

Επαγγελματική Εμπειρία

Σεπτέμβριος 2003– Σήμερα:	***Καθηγητής Πληροφορικής (ΠΕ19) στη Δευτεροβάθμια Εκπαίδευση, 2ο-3ο Ενιαίο Λύκειο,*** *Χανιά*
Σεπτέμβριος 2001– Ιούνιος 2003:	Στρατιωτική Θητεία στον ελληνικό Στρατό Ξηράς ως *Δόκιμος Έφεδρος Αξιωματικός (ΔΕΑ)* του Τεχνικού Σώματος με ειδικότητα Τεχνίτης Τηλεπικοινωνιών
Δεκέμβριος 2001 - Φεβρουάριος 2002:	***Φοίτηση στη* Σχολή Εκπαίδευσης Τεχνικών Τηλεπικοινωνιών (Σ.Ε.Τ.ΤΗΛ), *του Ελληνικού Στρατού Ξηράς, με ειδικότητα του Τεχνίτη Τηλεπικοινωνιών,*** *Πύργος*
Ιούλιος 1999 – Αύγουστος 1999	***Πρακτική άσκηση σύμφωνα με το πρόγραμμα* ΕΠΕΑΕΚ *στις εγκαταστάσεις της* ΔΕΗ Χανίων.**

Λοιπά Στοιχεία

- *Μέλος ΤΕΕ από τον Ιούνιο 2003 (Αρ. Μητρώου: 95696)*

ΠΕΡΙΛΗΨΗ

Οι δυνατότητες πληροφόρησης που παρέχει το Διαδίκτυο είναι απεριόριστες και οι πληροφορίες για οτιδήποτε χρειάζεται ο χρήστης σίγουρα υπάρχουν κάπου εκεί έξω. Είναι όμως χρήσιμες μόνο αν καταφέρουμε να τις βρούμε. Πώς όμως και με ποιον τρόπο γίνεται η αναζήτηση, το φιλτράρισμα, η αξιολόγηση και η παρουσίαση των πληροφοριών στον χρήστη; Οι μηχανές αναζήτησης είναι η λύση στο πρόβλημα και αποτελούν τους οδηγούς μας στο Internet. Μέχρι σήμερα έχει αναπτυχθεί ένας πολύ μεγάλος αριθμός μηχανών αναζήτησης, η κάθε μια από τις οποίες ακολουθεί δικές της τεχνικές αναζήτησης. Κάθε μηχανή αναζήτησης διέπεται από τους δικούς της κανόνες, που αφορούν τη σύνταξη, τον τρόπο εμφάνισης και αναζήτησης, τις πληροφορίες που αναζητά, τις τεχνικές αξιολόγησης κα. Επιπλέον, διαθέτει και συντηρεί τις δικές της βάσεις δεδομένων. Υπάρχουν όμως κάποιοι γενικοί κανόνες και τεχνικές αναζήτησης που πρέπει να ακολουθήσουμε, αν θέλουμε να φτάσουμε στα επιθυμητά αποτελέσματα.

Σκοπός της εργασίας αυτής είναι η καταγραφή των ιδιαίτερων χαρακτηριστικών τους όπως τεχνικών αναζήτησης και φιλτραρίσματος, αναζητούμενων πληροφοριών, τεχνικών αξιολόγησης και των απαραίτητων για αυτό πληροφοριών, κα. Θα προσπαθήσουμε επίσης να διαπιστώσουμε αν οι τεχνικές που εφαρμόζονται από τις μηχανές αναζήτησης διέπονται από κοινές αρχές ή εφαρμόζουν όμοιες τεχνικές με στόχο την ανάπτυξη ενός κοινού μοντέλου-προτύπου ενιαίας αναζήτησης.

ΕΙΣΑΓΩΓΗ

Η εποχή μας χαρακτηρίζεται από σημαντικές εξελίξεις στο χώρο της τεχνολογίας, οι οποίες ιδιαίτερα στον τομέα της *Τεχνολογίας της Πληροφορίας* ***(Information Technology, IT)*** είναι ραγδαίες, με αποτέλεσμα πλέον, να παρέχεται σε οιονδήποτε το επιθυμεί, η δυνατότητα πρόσβασης σε τεράστιους όγκους πληροφορίας. Το μόνο που απαιτείται είναι ένας υπολογιστής, ο οποίος να παρέχει σύνδεση στο δ*ιαδίκτυο* ***(Internet)*** και κάποιο *πρόγραμμα περιήγησης* **(*Browser*),** με τη βοήθεια του οποίου εμφανίζονται τα περιεχόμενα των διαφόρων *δικτυακών τοποθεσιών* ***(sites)*** του διαδικτύου.

Επίσης, η παγκόσμια κοινωνία αποτελεί πλέον μια πραγματικότητα που κανείς δεν μπορεί να αγνοήσει ή να αμφισβητήσει. Το διαδίκτυο και ο *Παγκόσμιος Ιστός* ***(World Wide Web, WWW)*** αναπτύσσονται με ιλιγγιώδεις ρυθμούς και χρησιμοποιούνται πλέον σε όλους τους τομείς και τις εκφράσεις της ανθρώπινης δραστηριότητας. Τρία εκατομμύρια δικτυωμένοι υπολογιστές και περισσότερες από ένα δισεκατομμύριο σελίδες *online*, προκαλούν νέα δεδομένα στον τομέα των επικοινωνιών και του εμπορίου, αλλαγές στον τρόπο με τον οποίο καθημερινά εργαζόμαστε, επικοινωνούμε και ανταλλάζουμε πληροφορίες. Εκεί ακριβώς έγκειται και η τεράστια απήχηση του διαδικτύου: στον τρόπο που μας παρέχει πρόσβαση στην πληροφορία, στον τρόπο που η πληροφορία αυτή ανακτάται και αναδιανέμεται σε εκατομμύρια υπολογιστών και χρηστών ανά την υφήλιο κι όλα αυτά σε χρόνους εξαιρετικά μικρούς. Η δημιουργία του παγκόσμιου ιστού στις αρχές της δεκαετίας του '90 έθεσε τις πρώτες βάσεις, ώστε όλο το υλικό που διοχετευόταν μέσα στο *Internet* να είναι στη διάθεση όλων όσων είχαν δυνατότητα σύνδεσης με αυτό.

Το διαδίκτυο περιλαμβάνει κάθε είδους πληροφορία και είναι ίσως το πρώτο

μέρος από το οποίο ξεκινά κάποιος την αναζήτηση στοιχείων για οποιονδήποτε και ο,τιδήποτε: από την αναζήτηση μίας φωτογραφίας ή ενός μουσικού αρχείου, μέχρι την αναζήτηση άλλων ανθρώπων με κοινά ενδιαφέροντα, προβληματισμούς, χόμπι, κ.ο.κ. Το διαδίκτυο είναι σαφέστατα η μεγαλύτερη βιβλιοθήκη του κόσμου, την οποία μπορεί να επισκεφθεί ο καθένας ανά πάσα στιγμή.

Διαβάζοντας τα παραπάνω εύκολα θα μπορούσε να θεωρήσει κανείς το *Δίκτυο* ***(Web)***, ως την ιδανική βιβλιοθήκη, όπου ο καθένας μπορεί να έχει πρόσβαση σε εκατομμύρια γνωστικές πηγές από την οθόνη του υπολογιστή του, να αναζητά και να ανακτά γρήγορα και εύκολα τις συγκεκριμένες πληροφορίες που τον ενδιαφέρουν. Αλλά τα πράγματα δεν έχουν ακριβώς έτσι.

Αυτό οφείλεται στο γεγονός ότι, το διαδίκτυο είναι χαώδες, εξαιτίας της ετερογενούς, αδόμητης και μη λογοκριθείσας φύσης του. Είναι ανοργάνωτο και άναρχο, χωρίς κανέλα πρότυπο τυποποίησης. Συνεπώς, αφενός ο τεράστιος όγκος των πληροφοριών του, ο οποίος αυξάνεται καθημερινά με εντυπωσιακό και ανεξέλεγκτο ρυθμό, αφετέρου η έλλειψη οργάνωσης των πληροφοριών που περιέχει -εφόσον κανείς δεν το ελέγχει- κατά τέτοιο τρόπο ώστε να εξασφαλίζεται άμεσα και εύκολα η πρόσβασή τους, δημιουργούν προβλήματα στους χρήστες όσον αφορά τον εντοπισμό και την πρόσβαση των πληροφοριών που επιθυμούν. Το διαδίκτυο εξάλλου, περιέχει εκατοντάδες εκατομμύρια *ιστοσελίδες* ***(web pages)*** που έχουν δημιουργηθεί από απλούς χρήστες, εταιρείες, οργανισμούς κ.λ.π., οι οποίες μέρα με τη μέρα αυξάνονται όλο και περισσότερο, ενώ από τις ήδη υπάρχουσες, άλλες αλλάζουν τοποθεσία (διεύθυνση), άλλες τροποποιούνται και άλλες καταργούνται τελείως. Είναι σαφές λοιπόν, ότι η χαρτογράφηση ενός τέτοιου δικτύου είναι κάτι παραπάνω από δύσκολη. Η πραγματική πρόκληση βρίσκεται ακριβώς στην επιλογή, στην απομόνωση της ποιοτικής πληροφορίας μέσα από τον κυκεώνα των εκατομμυρίων πληροφοριών που έχει καθένας σήμερα στη διάθεση του.

Προκειμένου λοιπόν να ικανοποιηθεί η ανάγκη των χρηστών του διαδικτύου, για τον εντοπισμό και την πρόσβαση των επιθυμητών πληροφοριών, αναπτύχθηκαν οι λεγόμενες *«μηχανές αναζήτησης»* ***(Search Engines)***, οι οποίες σκοπό έχουν τη διευκόλυνση του χρήστη στην αναζήτηση πληροφοριών, παρέχοντας του τη δυνατότητα χρήσης είτε θεματικών καταλόγων είτε κάποιων *λέξεων ή φράσεων-κλειδιών* (***key words***) για τον εντοπισμό τους.

Η αναγκαιότητα των μηχανών αναζήτησης και γενικότερα της καταγραφής των σελίδων και των πληροφοριών του Web σε βάσεις δεδομένων είχε γίνει αντιληπτή από

πολύ νωρίς. Η ραγδαία εξάπλωση του διαδικτύου, όμως, σε συνδυασμό με τις καθημερινές ανακατατάξεις στις σελίδες (κατάργηση ή αλλαγή διευθύνσεων) δυσχέραινε το έργο. Γι' αυτό το λόγο δημιουργήθηκε ένα σύστημα που έλεγχε συνέχεια το Web και κατέγραφε σε βάσεις δεδομένων όλες τις σελίδες, που ανήκαν στη βάση, ή τις αλλαγές που συντελούνταν σε αυτές. Το σύστημα αυτό ονομάστηκε ***Web Crawling***.

Παράλληλα, άρχισαν να δημιουργούνται κατηγορίες θεμάτων και να κατηγοριοποιούνται οι σελίδες και οι πληροφορίες του διαδικτύου. Αυτές οι κατηγορίες θεμάτων είναι τα λεγόμενα *θεματικά ευρετήρια* ***(Subject Dictionaries)***. Η λογική αυτής της ταξινόμησης δεν διαφέρει από τη λογική που χρησιμοποιείται από μία βιβλιοθήκη. Δηλαδή, αρχικά δημιουργούνται κάποιες γενικές κατηγορίες, στη συνέχεια υποκατηγορίες, κ.ο.κ. και τελικά οι πληροφορίες ταξινομούνται κατάλληλα ανάλογα με το θεματικό τους περιεχόμενο. Με αυτό τον τρόπο η ανάκτηση πληροφοριών γίνεται γρήγορα και απλά.

Οι μηχανές αναζήτησης και τα θεματικά ευρετήρια αποτελούν αναμφισβήτητα σήμερα το βασικό εργαλείο τόσο για τον αρχάριο όσο και για τον πιο εξοικειωμένο χρήστη του διαδικτύου. Πλέον, το 85% των χρηστών του *web* κάνουν χρήση των μηχανών αναζήτησης, από απλούς καταναλωτές και επαγγελματίες, μέχρι ερευνητές και επιστήμονες.

Στο διαδίκτυο βέβαια, η ανάκτηση των κατάλληλων πληροφοριών δεν είναι τόσο απλή υπόθεση. Υπάρχουν περιπτώσεις κατά τις οποίες κάποιες πληροφορίες καταγράφονται σε περισσότερες από μία κατηγορίες, με αποτέλεσμα η ανεύρεση τους να γίνεται δυσκολότερη.

Σε κάποιες άλλες περιπτώσεις, η μέθοδος των καταλόγων αποδεικνύεται ελλιπής για την κάλυψη των αναγκών των χρηστών, όπως, για παράδειγμα, όταν οι χρήστες δεν γνωρίζουν τον τρόπο με τον οποίο έχουν κατηγοριοποιηθεί οι πληροφορίες που αναζητούν. Σε αυτό το σημείο λοιπόν, γίνεται απαραίτητη η αναζήτηση πληροφοριών με βάση λέξεις ή φράσεις-κλειδιά που ορίζει ο χρήστης. Η μηχανή αναζήτησης αναλαμβάνει να ελέγξει τη βάση δεδομένων της και να εμφανίσει όλες τις ιστοσελίδες που περιλαμβάνουν τις λέξεις-κλειδιά. Αυτού του είδους η ελεύθερη αναζήτηση επιτρέπει μεν, την πρόσβαση σε μεγαλύτερο όγκο πληροφοριών, είτε χρήσιμων είτε άχρηστων, απαιτεί δε περισσότερο χρόνο από την πλευρά του χρήστη προκειμένου να διαχωρίσει τις χρήσιμες πληροφορίες, αυτές δηλαδή που τον ενδιαφέρουν, από τις άχρηστες.

Η επιλογή βέβαια των κατάλληλων λέξεων-κλειδιών και η χρήση των διαφόρων

τεχνικών που υποστηρίζουν οι μηχανές αναζήτησης, προκειμένου αφενός να περιοριστεί το εύρος των αποτελεσμάτων της αναζήτησης και αφετέρου να μειωθούν στο ελάχιστο τα άχρηστα αποτελέσματα, παρουσιάζει ακόμα αρκετές δυσκολίες.

Σε ότι αφορά το παραπάνω πρόβλημα, την ευστοχία των αποτελεσμάτων, την ουσιαστική προσέγγιση δηλαδή των απαιτήσεων του χρήστη ένας άλλος επιστημονικός χώρος καλείται να δώσει τη λύση. Πρόκειται για το *Φιλτράρισμα Πληροφοριών* ***(Information Filtering)****,* που χρησιμοποιεί διάφορες τεχνικές με σκοπό την επιλογή και απομόνωση της ποιοτικής μόνο πληροφορίας. Σε αντίθεση με τις παραδοσιακές μηχανές αναζήτησης και τα θεματικά ευρετήρια που δεν γνωρίζουν τίποτα για το χρήστη τους, τα συστήματα που χρησιμοποιούν τεχνικές από το χώρο του *Information Filtering* "μελετούν" το χρήστη τους και ασχολούνται με τις μακροπρόθεσμες ανάγκες του. Με βάση το προφίλ του χρήστη που διαμορφώνεται από τα συστήματα αυτά πραγματοποιείται αναζήτηση που είναι όσο το δυνατόν περισσότερο συγκεκριμένη και παραμετροποιημένη.

Σκοπός της παρούσας μεταπτυχιακής διατριβής είναι η καταγραφή των ιδιαίτερων χαρακτηριστικών των μηχανών αναζήτησης, όπως τεχνικών αναζήτησης και φιλτραρίσματος, αναζητούμενων πληροφοριών, τεχνικών αξιολόγησης και των απαραίτητων για αυτό πληροφοριών. Θα γίνει προσπάθεια επίσης να διαπιστωθεί αν οι τεχνικές που εφαρμόζονται από τις μηχανές αναζήτησης διέπονται από κοινές αρχές ή εφαρμόζουν όμοιες τεχνικές με στόχο την ανάπτυξη ενός κοινού μοντέλου-προτύπου ενιαίας αναζήτησης.

Πιο συγκεκριμένα, η δομή αυτής της εργασίας έχει ως εξής: Στο πρώτο κεφάλαιο θα αναφερθεί η αναζήτηση των πληροφοριών στο World Wide Web και θα γίνει προσπάθεια κατανόησης του λόγου που, ειδικά στη σημερινή εποχή, είναι απαραίτητες οι μηχανές αναζήτησης. Στο δεύτερο μέρος αυτού του κεφαλαίου θα γίνει μια γενική αναφορά στις μηχανές αναζήτησης. Θα γίνει λόγος για την προϊστορία τους, για τα είδη τους και για τα βασικά μέρη από τα οποία αποτελούνται.

Στο δεύτερο κεφάλαιο θα αναπτυχθούν οι μέθοδοι και οι τεχνικές αναζήτησης πληροφοριών στο Διαδίκτυο. Ενδεικτικά αναφέρονται: *Αναζήτηση με Περιήγηση* ***(Browsing)***, *Αναζήτηση με Χρήση Θεματικών Ευρετηρίων* ***(Subject Index Search ή Directory Search)***, *Αναζήτηση με Λέξεις-Κλειδιά* ***(Keyword Search)***, *Συνδυαστική Αναζήτηση* ***(Combined Search Engine/ Directory Search)***, *Αναζήτηση με Βάση το Προφίλ του Χρήστη* ***(User's Profile Searching)***.

Στο τρίτο κεφάλαιο μελετώνται τα εργαλεία αναζήτησης πληροφοριών από το

διαδίκτυο. Αρχικά γίνεται λόγος για τις μηχανές αναζήτησης, για τα χαρακτηριστικά τους, για το πως δουλεύουν και ποιες τεχνικές χρησιμοποιούν. Στη συνέχεια θα αναπτυχθούν οι *Μηχανές Πολλαπλής Αναζήτησης* ***(Multi or Meta-Search Engine)***, τα χαρακτηριστικά τους και τα πλεονεκτήματά τους. Η ίδια αναφορά θα γίνει και για τους *Θεματικούς Καταλόγους- Θεματικά Ευρετήρια* ***(Subject catalogues- Subject directories)*** και για τις *Υβριδικές Μηχανές Αναζήτησης* ***(Hybrid Search Engines)***, καθώς και για τις *«Πύλες» και «Εικονικές» Βιβλιοθήκες* ***(Gateways and Virtual Libraries)***.

Στο τέταρτο κεφάλαιο παρουσιάζεται κάθε μια από τις κύριες μηχανές αναζήτησης ξεχωριστά. Θα αναφερθεί η αρχιτεκτονική τους, το πώς δουλεύουν αλλά και τα ιδιαίτερα τεχνικά χαρακτηριστικά της κάθε μιας. Στο τέλος του κεφαλαίου θα γίνει μια σύγκριση αυτών των μηχανών αναζήτησης.

Στο πέμπτο κεφάλαιο σχολιάζουμε τις πιο γνωστές μηχανές πολλαπλής αναζήτησης. Θα εξεταστεί η αρχιτεκτονική τους και τα χαρακτηριστικά τους. Στο τέλος του κεφαλαίου θα γίνει μια σύγκριση αυτών των μηχανών αναζήτησης.

Στο έκτο κεφάλαιο θα αναπτυχθούν οι θεματικοί κατάλογοι-θεματικά ευρετήρια. Θα εξεταστούν τα πιο γνωστά, όπως το Yahoo (www.yahoo.com) και το Excite (www.excite.com). Στο τέλος του κεφαλαίου θα γίνει μια σύγκριση αυτών των μηχανών αναζήτησης.

Το έβδομο κεφάλαιο είναι αφιερωμένο στις πιο γνωστές μηχανές αναζήτησης με βάση το προφίλ του χρήστη. Γίνεται αναφορά για τις Amalthia (www.amalthia.com) και Lawbot (www.lawbot.com).

Στο όγδοο κεφάλαιο, έχοντας αναλύσει και συγκρίνει στα παραπάνω κεφάλαια τα εργαλεία αναζήτησης που χρησιμοποιούνται στο Διαδίκτυο, αναπτύσσεται το Μοντέλο-Πρότυπο Ενιαίας Αναζήτησης.

Η μεταπτυχιακή διατριβή κλείνει με το κεφάλαιο εννέα, όπου αναφέρονται τα συμπεράσματα της εργασίας, οι προβληματισμοί που ανέκυψαν και οι μελλοντικές επεκτάσεις.

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

1.1. Πρόλογος

Δημιούργημα των τελευταίων τριών δεκαετιών, το διαδίκτυο συνεχίζει, λίγο μετά την αυγή της νέας χιλιετίας, να γιγαντώνεται, να μεταμορφώνεται και να εξελίσσεται, σε μια προσπάθεια να καθιερωθεί ως το απόλυτο μέσο επικοινωνίας για τους πολίτες αυτού του πλανήτη. Για πολλούς η εποχή της παγκοσμιοποίησης έχει ήδη ξεκινήσει και οι συνέπειες της κάνουν αισθητή την παρουσία τους σε κάθε έκφραση της ανθρώπινης δραστηριότητας, είτε πρόκειται για επαγγελματική και κοινωνική, είτε για προσωπική δραστηριοποίηση: έκρηξη των επικοινωνιών, συγχωνεύσεις και εξαγορές μεταξύ γιγαντιαίων πολυεθνικών οργανισμών και χρηματοπιστωτικών ιδρυμάτων, μετασχηματισμοί των πολιτικών και θρησκευτικών συστημάτων ανά τον κόσμο, σε μια συνεχή προσπάθεια για επιβίωση, επικράτηση και προσαρμογή στα νέα δεδομένα που θέτει η παγκοσμιοποίηση.

Βασική προϋπόθεση για την υλοποίηση αυτής της διαδικασίας είναι η επικοινωνία. Η απρόσκοπτη επικοινωνία και ανταλλαγή των πληροφοριών σε κάθε σημείο του πλανήτη, ο τρόπος με τον οποίο η πληροφορία μεταδίδεται, αποθηκεύεται, ανακτάται και επαναχρησιμοποιείται, όπως και τα τεχνικά μέσα με τα οποία επιτυγχάνεται η επικοινωνία είναι αδιαμφισβήτητα δημιουργήματα του διαδικτύου. Η παγκοσμιοποίηση δείχνει πλέον απτή, και το διαδίκτυο χωρίς καμία αμφιβολία

διαδραμάτισε ίσως τον σημαντικότερο ρόλο στην πραγμάτωσή της. Το διαδίκτυο είναι το όχημα που θα μας οδηγήσει στην παγκόσμια κοινωνία της άμεσης πληροφόρησης, στην κοινωνία της εμπορικής ανταλλαγής προϊόντων από και προς κάθε γωνία του πλανήτη, στην κοινωνία όπου η εργασία, η διασκέδαση, ο τρόπος που οι άνθρωποι επικοινωνούν μεταξύ τους θα γνωρίσουν μια νέα εποχή.

1.2. Αναζήτηση Πληροφοριών στο World Wide Web (www)

Το *δίκτυο* ***(web)*** έφερε επανάσταση στον τρόπο με τον οποίο ο άνθρωπος αποκτούσε πρόσβαση στην πληροφορία και άνοιξε νέους ορίζοντες στις μεθόδους που αυτή ανακτάται και διαχέεται.

Ειδικότερα σε τομείς όπως οι ηλεκτρονικές βιβλιοθήκες, η εκπαίδευση, το εμπόριο, η ψυχαγωγία ή ακόμη η φαρμακευτική και η ιατρική επιστήμη, η συνεισφορά του διαδικτύου στον τρόπο με τον οποίο αναζητείται, ανακτάται και διαχέεται η πληροφορία είναι τεράστια. Το διαδίκτυο έφερε πραγματική επανάσταση και άλλαξε για πάντα τον τρόπο με τον οποίο γίνεται η διαχείριση των πληροφοριών, είτε αυτές προέρχονται από έντυπες είτε από ηλεκτρονικές πηγές. Τα εργαλεία αναζήτησης επιτρέπουν την πρόσβαση σε τεράστιους όγκους πληροφοριών με τρόπο φιλικό και εύχρηστο ακόμη και για τον αρχάριο χρήστη. Οποιοσδήποτε έχει πρόσβαση στο διαδίκτυο αποκτά ταυτόχρονα πρόσβαση στη μεγαλύτερη βιβλιοθήκη του κόσμου, στη μεγαλύτερη συγκέντρωση πληροφοριών που γνώρισε ποτέ η ανθρωπότητα.

Σίγουρα το διαδίκτυο δεν αποτελεί πανάκεια για τα ερωτήματα και τις έρευνες όλων των ειδών. Σύμφωνα με τελευταίους υπολογισμούς (Κωνσταντινίδης, 2000), μόνο ένα μικρό ποσοστό, περίπου 14-15%, του συνόλου των περιεχομένων του διαδικτύου καλύπτεται από τα διάφορα εργαλεία αναζήτησης. Ειδικές επιχειρηματικές πληροφορίες, περιεχόμενα περιοδικών και άλλων έντυπων αποτελούν εμπορικά προϊόντα μεγάλης αξίας. Εξάλλου, ας μην ξεχνάμε ότι αυτή τη στιγμή το διαδίκτυο αποτελεί την ταχύτερα αναπτυσσόμενη επιχείρηση του κόσμου με πωλήσεις δισεκατομμυρίων δολαρίων, ενώ ένα μεγάλο κομμάτι του -ίσως το μεγαλύτερο- είναι σαφώς εστιασμένο στην εμπορική εκμετάλλευση, είτε πρόκειται για πληροφορίες είτε για προϊόντα. Εντούτοις, κανείς δεν μπορεί να αμφισβητήσει ότι η προσφορά του διαδικτύου δεν έγκειται στον όγκο των πληροφοριών που βρίσκονται διαθέσιμες -εξάλλου και σήμερα υπάρχουν αρκετές πηγές που διαθέτουν τεράστιο όγκο αποθηκευμένων πληροφοριών-, αλλά κυρίως στον τρόπο με τον οποίο έχει κάποιος πρόσβαση σε αυτές. Δεν υπάρχει αμφιβολία ότι στο άμεσο μέλλον το διαδίκτυο θα αποτελέσει την κύρια πηγή για τις αγορές μας και το βασικό σημείο αναφοράς για οποιαδήποτε απορία έχουμε, για οποιαδήποτε πληροφορία αναζητάμε και για όποια έρευνα θελήσουμε να εκπονήσουμε.

Πριν από μερικά χρόνια η διαφοροποίηση ανάμεσα στις τότε λιγοστές μηχανές αναζήτησης ήταν ελάχιστη ή μηδενική. Την τελευταία διετία το διαδίκτυο έχει

αναδειχθεί σε σημαντικό φορέα-παροχέα πληροφοριών, είτε απευθύνεται στον επαγγελματία, είτε στον καταναλωτή, είτε στον απλό περιηγητή. Μέρα με την ημέρα οι μηχανές αναζήτησης πληθαίνουν, κατά συνέπεια ο ανταγωνισμός γίνεται ολοένα εντονότερος. Νέα χαρακτηριστικά, νέες τεχνολογίες και υπηρεσίες στην υπηρεσία του χρήστη, νέα προϊόντα και ευκολίες, μεγαλύτερη κάλυψη, όλα σε μια προσπάθεια να καθιερωθούν σε μια αγορά έντονα ανταγωνιστική. Το αποτέλεσμα των παραπάνω είναι καλύτερες υπηρεσίες και προϊόντα για τον τελικό χρήστη, που πλέον μπορεί να επιλέξει τη μηχανή που καλύπτει καλύτερα τις πληροφοριακές ανάγκες του, με τρόπο εύχρηστο και φιλικό.

Ξεκινώντας την προσπάθεια να αναζητήσουμε πληροφορίες στο δίκτυο, εκείνο που σίγουρα θα πρέπει να γνωρίζουμε είναι ότι οι μηχανές αναζήτησης και γενικότερα τα εργαλεία αναζήτησης είναι εκείνα που θα πρέπει να αποτελέσουν το σημείο αναφοράς και εκκίνησης κάθε προσπάθειας. Οποιαδήποτε άλλη προσπάθεια χωρίς να είναι γνωστή η συγκεκριμένη ηλεκτρονική διεύθυνση ή κάποιο άλλο σημείο αναφοράς είναι καταδικασμένη.

Τα δεδομένα που προαναφέρθηκαν-έλλειψη συγκεντρωτικών ευρετηρίων, έλλειψη τυποποίησης των περιεχομένων και κοινώς αποδεκτού συστήματος ταξινόμησης- αποτελούν σίγουρα ανασταλτικούς παράγοντες για μια απόπειρα περιήγησης/ αναζήτησης χωρίς τη χρήση των κατάλληλων εργαλείων.

Δε θα ήταν υπερβολή να υποστηρίξουμε πως χωρίς τη χρήση των ειδικών αυτών εργαλείων η εξερεύνηση στο δίκτυο θα ήταν σαν να προσπαθούμε να βρούμε ένα βιβλίο στη συλλογή μιας βιβλιοθήκης όταν τα βιβλία βρίσκονται στα ράφια χωρίς καμία σειρά ή σύστημα, ενώ παράλληλα λείπει και ο κατάλογος των περιεχομένων της βιβλιοθήκης. Βέβαια, θα πρέπει να τονίσουμε πως η πληροφορία που περιέχεται στο δίκτυο ίσως δεν ακολουθεί τα τυποποιημένα ταξινομικά συστήματα και νόρμες, ωστόσο ακολουθεί κάποιους στοιχειώδεις κανόνες, που σχετίζονται τόσο με τη μορφή με την οποία εισάγονται και δημοσιεύονται οι πληροφορίες αυτές όσο και με τον τρόπο με τον οποίο τις αντιλαμβάνεται ο άνθρωπος. Για παράδειγμα, με διαφορετικό τρόπο εισάγεται και αναζητείται ένα αρχείο κειμένου ή ένα αρχείο ήχου ή κινούμενης εικόνας και διαφορετικά ένα *e-mail* ή μια ιστοσελίδα. Αντίστοιχα, με διαφορετικό τρόπο αντιλαμβανόμαστε τις διαφορετικές αυτές μορφές των πληροφοριών, κατά συνέπεια με διαφορετικό τρόπο τις αναζητούμε. Καταλήγουμε λοιπόν στο συμπέρασμα ότι, έχοντας μια αρκετά σαφή εικόνα του είδους της πληροφορίας που αναζητούμε, αυξάνουμε τις πιθανότητες μιας επιτυχημένης αναζήτησης στα περιεχόμενα του παγκόσμιου ιστού.

Στο δίκτυο υπάρχουν τρεις βασικές πληροφοριακές πηγές (Κωνσταντινίδης, 2000):

- *οι κύριες σελίδες πληροφοριών* ***(Primary Information Sites/Pages)***
- *οι κύριες σελίδες παραπομπών* ***(Primary Link Sites/Pages)***
- *οι απλές σελίδες παραπομπών* ***(Simple Link Sites/Pages)***

Τα κύρια sites-σελίδες πληροφοριών αποτελούν τις «αποθήκες», όπου είναι συγκεντρωμένη η πληροφορία. Στην πλειονότητα των περιπτώσεων πρόκειται για βάσεις δεδομένων, συλλογικούς καταλόγους βιβλιοθηκών, κ.λ.π. Αυτές οι βάσεις δεδομένων είναι οργανωμένες στο πλαίσιο ενός καταλόγου βιβλιοθήκης, με θεματικά οργανωμένη συλλογή, ιεραρχική ταξινόμηση των περιεχομένων και συγκεντρωτικό ευρετήριο με δυνατότητα αναζήτησης.

Οι κύριες σελίδες-sites παραπομπών αποτελούν στην ουσία καταλόγους με παραπομπές προς κύριες πηγές πληροφοριών, οργανωμένες θεματικά στην πλειονότητα των περιπτώσεων. Οι σελίδες αυτές αποτελούν τα πλέον χρήσιμα σημεία αναφοράς για το λόγο ότι συγκεντρώνουν σε σχετικά μικρή έκταση έναν τεράστιο όγκο πληροφοριών, οι οποίες παρέχονται έμμεσα. Οι κύριες σελίδες παραπομπών-δεικτών αποτελούν τις κατεξοχήν καταχωρίσεις στα αρχεία *«βιβλιοδεικτών»* ***(Bookmarks)***.

Τέλος, οι απλές σελίδες-sites παραπομπών αποτελούν μικρότερους σε έκταση καταλόγους με παραπομπές και δείκτες σε άλλες σελίδες, που με τη σειρά τους παραπέμπουν στα κύρια sites-σελίδες πληροφοριών.

Αυτές οι τρεις βασικές πηγές σης περισσότερες περιπτώσεις λειτουργούν συμπληρωματικά, ώστε να εξασφαλίσουν στο χρήστη τη μεγαλύτερη δυνατή κάλυψη του ζητούμενου θέματος. Στο *web* και στο *Internet* υπάρχουν χιλιάδες πηγές πληροφοριών και εκατομμύρια διαθέσιμες πληροφορίες. Σε ορισμένες περιπτώσεις η αναζήτηση και η ανάκτηση των πληροφοριών αυτών γίνονται με χαρακτηριστική ευκολία, ενώ σε άλλες περιπτώσεις απαιτούνται ειδικές τεχνικές και μέθοδοι αναζήτησης.

1.2.1. Όγκος Πληροφοριών

Επιχειρώντας μια αναδρομή σε παλαιότερες εποχές, όταν η πληροφορία και η πληροφόρηση γενικότερα αποτελούσαν προνόμιο των λιγοστών που είχαν τα μέσα να τις αποκτήσουν, μπορούμε ευκολότερα σήμερα να συνειδητοποιήσουμε τον αντίκτυπο του διαδικτύου στην καθημερινή ζωή μας, όπου όλοι ανεξαιρέτως γινόμαστε πλέον

δέκτες εκατοντάδων χιλιάδων μηνυμάτων και πληροφοριών, ειδήσεων και εικόνων που, ηθελημένα ή αθέλητα, λαμβάνουμε κατά την περιήγηση μας στον κυβερνοχώρο. Η φράση *«η πληροφορία προσδίδει δύναμη»* έχει χάσει πια την αρχική σημασία της, αφού όλοι όσοι έχουν πρόσβαση στο διαδίκτυο έχουν στη διάθεση τους τη μεγαλύτερη πηγή πληροφοριών που γνώρισε ποτέ η ανθρωπότητα. Το χάσμα ανάμεσα στην πληροφορία και τη γνώση που αυτή συνεπάγεται δείχνει να μεγαλώνει καθημερινά, καθώς ο όγκος των πληροφοριών που άτακτα καταχωρίζονται καθημερινά στο διαδίκτυο μάλλον δυσχεραίνει παρά διευκολύνει την πορεία προς την εξεύρεση ποιοτικών πληροφοριών. Έτσι, η πραγματική πρόκληση βρίσκεται ακριβώς στην επιλογή, στην απομόνωση της ποιοτικής πληροφορίας- με τη χρήση των κατάλληλων εργαλείων- μέσα από τον κυκεώνα των εκατομμυρίων πληροφοριών που έχει καθένας σήμερα στη διάθεση του και στη μετουσίωση της σε γνώση.

Όταν πλέον η ποσότητα των πληροφοριών που έχουμε στη διάθεση μας γίνει τόση ώστε να είναι δύσκολο να τη διαχειριστούμε, μια σειρά από «παρενέργειες» κάνουν την εμφάνιση τους:

- ελλιπής κατανόηση των πληροφοριών
- αδυναμία αξιολόγησης των πληροφοριών
- αδυναμία εξεύρεσης-ανάκτησης των πληροφοριών.

Τα συμπτώματα αυτά τα τελευταία χρόνια έχουν αποκτήσει επίσημη ονομασία, είναι γνωστά ως ***Information Overload***. Τα αποτελέσματα σε μια τέτοια κατάσταση είναι ο εκνευρισμός, η άσκοπη σπατάλη χρόνου και η κακή ή ελλιπής χρήση των διαθέσιμων πληροφοριακών πηγών. Είναι γεγονός ότι ο παγκόσμιος ιστός και οι παρελκυόμενες υπηρεσίες και προϊόντα που προσφέρει, όπως *το ηλεκτρονικό ταχυδρομείο* ***(e-mail)*** και το ***Usenet***, έχουν επιτείνει την κατάσταση αυτή, και πραγματικά είναι λίγοι εκείνοι που συνειδητοποιημένα κατορθώνουν να οργανώσουν τη ροή και να ελέγξουν την ποιότητα των πληροφοριών που λαμβάνουν και που τελικά αφομοιώνουν. Είναι μάλλον σίγουρο ότι η κατάσταση δεν πρόκειται να βελτιωθεί, τουλάχιστον σε ό,τι αφορά την ποσότητα των πληροφοριών που διοχετεύονται καθημερινά μέσω του παγκόσμιου ιστού και των υπηρεσιών του. Οι χρήστες του διαδικτύου θα συνεχίσουν να αυξάνονται καθημερινά, όπως θα συνεχίσουν να αυξάνονται και οι εταιρείες εκείνες που θα δουν το διαδίκτυο ως το καινούριο πολυκατάστημα μέσω του οποίου θα προσεγγίσουν τον καταναλωτή.

Δυστυχώς, με τον όγκο των διαθέσιμων πληροφοριών που προσφέρεται, το να δοθεί μόνιμη λύση στο πρόβλημα της υπερπροσφοράς των πληροφοριών θα ήταν μάλλον ουτοπικό. Προς το παρόν, παρά τα *σχετικά προγράμματα* ***(Information Agents)***

που έχουν αναπτυχθεί, την εξέλιξη που έχει επιτευχθεί στα εργαλεία αναζήτησης και έρευνας και τις προσπάθειες μερικού ελέγχου των πληροφοριών που καταχωρούνται και δημοσιεύονται στο διαδίκτυο, κανείς δεν είναι σε θέση να ισχυριστεί ότι διαθέτει μια μόνιμη λύση, ικανή να λύσει το πρόβλημα. Έτσι, το μόνο σημείο όπου όλοι δείχνουν να συμφωνούν είναι μια κουλτούρα την οποία θα πρέπει να υιοθετήσει ο χρήστης απέναντι στην πληροφόρηση γενικότερα, προκειμένου να αντεπεξέλθει, έστω και μερικώς, στα προβλήματα που προκαλεί η αλόγιστη συσσώρευση πληροφορίας.

1.2.2. Παγκόσμιος Ιστός (World Wide Web, www)

Όταν στα τέλη της δεκαετίας του '80 ο κόσμος, που μόλις είχε αρχίσει να συνειδητοποιεί την επίδραση των οικιακών προσωπικών υπολογιστών, άρχισε να ακούει για ένα νέο δίκτυο υπολογιστών, που είχε υλοποιηθεί στο πλαίσιο κάποιου ερευνητικού προγράμματος, σίγουρα δε θα μπορούσε να φανταστεί το μέγεθος που θα κατέληγε να έχει αυτό ύστερα από μία δεκαετία και τις επιπτώσεις -έμμεσες ή άμεσες- στην κοινωνία, την οικονομία, τον πολιτισμό, τον ίδιο τον άνθρωπο.

Σίγουρα κανείς δεν μπορούσε να αναλογιστεί το γιγαντισμό του διαδικτύου υστέρα από μία δεκαετία. Το διαδίκτυο, και ειδικά ο παγκόσμιος ιστός, είναι απίστευτα δημοφιλείς στα σπίτια και στα γραφεία. Λόγω της απουσίας κεντρικού ελέγχου οι στατιστικές για το δίκτυο είναι σε κάποιο βαθμό ανακριβείς. Είναι αδιαμφισβήτητο, όμως, ότι το δίκτυο είναι τεράστιο σε αριθμό χρηστών, δικτυακών τοποθεσιών, και ιστοσελίδων. Παραδείγματος χάριν, μια εκτίμηση του ελάχιστου αριθμού υπολογιστών συνδεδεμένων στο διαδίκτυο είναι πάνω από 16 εκατομμύρια *(Internet Domain Survey, 1997)*. Ομοίως, από τα 220 εκατομμύρια ανθρώπων στις Ηνωμένες Πολιτείες και τον Καναδά πάνω από την ηλικία των 16, το 23% (πάνω από 50 εκατομμύρια) υπολογίζεται ότι χρησιμοποιεί το Διαδίκτυο και το 17% (πάνω από 37 εκατομμύρια) τον παγκόσμιο ιστό (CommerceNet/ Nielsen, 1997). Αυτοί οι αριθμοί θεωρούνται αληθής, όπως επίσης και το γεγονός ότι, μερικοί αναλυτές υποστηρίζουν ότι, ο ιστός διπλασιάζεται σε μέγεθος κάθε 100 έως 125 ημέρες (Morgan, 1996).

Ο παγκόσμιος ιστός έχει επιφέρει νέα δεδομένα στον τομέα των επικοινωνιών και του εμπορίου, αλλαγές στον τρόπο με τον οποίο εργαζόμαστε, επικοινωνούμε και ανταλλάσσουμε πληροφορίες. Η παγκόσμια κοινωνία, που πολλοί είχαν οραματιστεί με τον ένα ή τον άλλο τρόπο, αποτελεί πλέον μια πραγματικότητα που κανείς δε θα μπορούσε να αγνοήσει ή να αμφισβητήσει.

Ο παγκόσμιος ιστός γιγαντώνεται με ιλιγγιώδεις ρυθμούς και χρησιμοποιείται πλέον σε όλους τους τομείς και τις εκφράσεις της ανθρώπινης κοινωνίας. Από τον επιστήμονα ερευνητή που θα αναζητήσει τις τελευταίες εξελίξεις στο χώρο του και τον καταναλωτή που θα αναζητήσει τα πιο πρόσφατα προϊόντα που κυκλοφορούν στην αγορά έως τον εκπαιδευτικό που θα περιηγηθεί στις τελευταίες ανακοινώσεις συνεδρίων για την εκπαίδευση, ο παγκόσμιος ιστός και το διαδίκτυο δείχνουν να κυριαρχούν ολοένα εντονότερα στην καθημερινή ζωή μας σε όλες τις εκφάνσεις της. ανεξάρτητα από ηλικία, μορφωτικό επίπεδο, χώρα και γλώσσα. Ο παγκόσμιος ιστός και το *Web* διαμορφώνουν μια νέα πραγματικότητα, όπου η πληροφορία κατέχει κυρίαρχο ρόλο στην καθημερινότητα μας. Εκεί ακριβώς έγκειται η τεράστια απήχηση του διαδικτύου: στον τρόπο που μας παρέχει πρόσβαση στην πληροφορία, στον τρόπο που η πληροφορία αυτή ανακτάται και αναδιανέμεται σε εκατομμύρια υπολογιστών και χρηστών ανά την υφήλιο, σε χρόνους που κάποτε η αναφορά τους αποτελούσε προϊόν επιστημονικής φαντασίας.

Αν και πολλές διαδικτυακές εφαρμογές και υπηρεσίες είναι διαθέσιμες σήμερα, η πρωταρχική χρήση του διαδικτύου (εκτός από το ηλεκτρονικό ταχυδρομείο) είναι η ανάκτηση των πληροφοριών. Με το πλεονέκτημα της εύκολης χρήσης των εργαλείων δημιουργίας μιας ιστοσελίδας, κάθε άτομο έχει τοποθετήσει πληροφορίες σχεδόν για κάθε θέμα στο διαδίκτυο. Φυσικά, με μια τέτοια ποικιλομορφία των περιεχομένων του, και με τον τεράστιο όγκο των πληροφοριών που βρίσκονται στο διαδίκτυο, ο δρόμος της ανάκτησης των πληροφοριών είναι μακρύς και αβέβαιος.

1.2.3. Σημερινά Μεγέθη

Σήμερα υπάρχουν δισεκατομμύρια σελίδες. Μια ζωντανή, τεράστια βιβλιοθήκη, που τα περιεχόμενα της ανανεώνονται σχεδόν καθημερινά και οι σελίδες της καλύπτουν κάθε τομέα της ανθρώπινης γνώσης, κάθε ανθρώπινη δραστηριότητα και έκφραση. Μια ιδανική, εικονική, παγκόσμια βιβλιοθήκη, αποδεσμευμένη από γεωγραφικά όρια, θρησκείες και πολιτικά καθεστώτα, προορισμένη να προσφέρει στους χρήστες της έναν κοινό τόπο αναζήτησης, μια κοινωνία πληροφοριών όπου ο καθένας θα μπορεί να επικοινωνεί, να ωφελεί με τις γνώσεις του τους συν-αναγνώστες του και να ωφελείται με τη σειρά του από τις πληροφορίες και τη γνώση που αυτοί μπορούν να του προσφέρουν. Αυτές οι σελίδες, θα μπορούσαν να αποτελέσουν ίσως το μοναδικό σημείο αναφοράς και αναζήτησης πληροφοριών, ψυχαγωγίας ή απλά περιήγησης,

ανεξαρτήτως αντικειμένου, αφού είναι δύσκολο να σκεφτούμε κάτι για το οποίο δεν υπάρχει έστω μία αναφορά στον κυβερνοχώρο.

Χαρακτηριστικές είναι οι διαπιστώσεις του Wurman, που αναφέρει ότι: *«μία εβδομαδιαία έκδοση των "New York Times" περιέχει περισσότερες πληροφορίες από όσες ένας μέσος άνθρωπος θα μπορούσε να δεχτεί κατά τη διάρκεια της ζωής του (!) στην Αγγλία του 17ου αιώνα»* (Κωνσταντινίδης, 2000). Στη σημερινή εποχή, ο άνθρωπος αντιμετωπίζει μια αντίστροφη κατάσταση, όπου από την παντελή έλλειψη πληροφοριών έχει πλέον κατακλυστεί από τον όγκο των πληροφοριών με τις οποίες βομβαρδίζεται καθημερινά μέσω των εφημερίδων, των περιοδικών, του διαδικτύου, του e-mail, της τηλεόρασης, του ραδιοφώνου και των άλλων μέσων που οι παραγωγοί και διαθέτες των πληροφοριών χρησιμοποιούν για να φτάσουν τους αποδέκτες τους σε όλα τα μήκη και τα πλάτη του πλανήτη. Έτσι, το διαδίκτυο και το web θα αποτελέσουν -και για κάποιους ήδη αποτελούν- την παγκόσμια κοινωνία, όπου ιδέες και αγαθά θα διακινούνται χωρίς περιορισμούς. Επομένως, είναι τουλάχιστον ουτοπικό να κάνουμε λόγο για μείωση των πληροφοριακών πηγών, για μείωση των πληροφοριών που διακινούνται στο διαδίκτυο, για ποιοτικό έλεγχο και αξιολόγηση των περιεχομένων. Η εποχή της πληροφορίας έχει φτάσει.

Θα αναφέρουμε ενδεικτικά κάποιες μετρήσεις για τη γιγάντωση του διαδικτύου και πως πραγματικά φτάσαμε στη σημερινή πραγματικότητα. Το εθνικό κέντρο εφαρμογών supercomputing (National Center for Supercomputing Applicatios - NCSA) υπολογίζει ότι ο αριθμός χρηστών του διαδικτύου αυξήθηκε από 1 εκατομμύριο έως 25 εκατομμύρια σε μια πενταετία, μέχρι τον Ιανουάριο του 1997 (Schatz, 1997). Ένας οργανισμός μετρήσεων του διαδικτύου, ο Matrix Information and Directory Services (MIDS), παρέχει διάφορα στατιστικά για τους χρήστες διαδικτύου: υπολόγισε ότι τον Απρίλιο του 1998 υπήρχαν 57 εκατομμύρια χρήστες του διαδικτύου παγκοσμίως, και ότι ο αριθμός τους θα αυξανόταν σε 377 εκατομμύρια μέχρι το 2000. Ο Morgan Stanley δίνει την εκτίμηση των 150 εκατομμυρίων το 2000, ενώ ο Killen δίνει την εκτίμηση των 250 εκατομμυρίων το 2000. Ο Nua υπολογίζει τον αριθμό των χρηστών στα 201 εκατομμύρια παγκοσμίως το Σεπτέμβριο του 1999, και πιο συγκεκριμένα ανά περιοχή: 1,72 εκατομμύρια χρήστες στην Αφρική, 33,61 στην Ασία/Ειρηνικό, 47,15 στην Ευρώπη, 0,88 στη Μέση Ανατολή, 112,4 στον Καναδά και στις ΗΠΑ και 5,29 στη Λατινική Αμερική. Από τις παραπάνω μετρήσεις βγαίνει το συμπέρασμα για μια μεγάλη και συνεχή αύξηση των χρηστών (κυρίως εκθετική) παρότι οι ακριβείς μετρήσεις διαφέρουν.

Τα περισσότερα δεδομένα σε σχέση με το μέγεθος της πληροφορίας στο διαδίκτυο παρουσιάζουν τεράστια αύξηση, και τα μεγέθη καθώς και οι αριθμοί εμφανίζονται να αυξάνονται εκθετικά. Ο Lynch (Lynch, 1997) έχει τεκμηριώσει την εκρηκτική αύξηση των ***hosts*** του διαδικτύου, όπου ο συγκεκριμένος αριθμός διπλασιάζεται κατά προσέγγιση κάθε έτος. Παραδείγματος χάριν, υπολογίζει ότι ήταν 1,3 εκατομμύρια τον Ιανουάριο του 1993, 2,2 εκατομμύρια τον Ιανουάριο του 1994, 4,9 εκατομμύρια τον Ιανουάριο του 1995, και 9,5 εκατομμύρια τον Ιανουάριο του 1996. Και άλλες μελέτες που έχουν πραγματοποιηθεί καταλήγουν στο ίδιο συμπέρασμα για την αύξηση των *hosts*. Το τελευταίο σύνολο των στοιχείων του είναι 12,9 εκατομμύρια hosts τον Ιούλιο του 1996.

Ο αριθμός των δημόσια προσιτών σελίδων αυξάνεται επίσης με μεγάλο ρυθμό. Τον Ιανουάριο του 1997 υπήρξαν 80 εκατομμύρια δημόσιες ιστοσελίδες, και αυτός ο αριθμός θα διπλασιαζόταν ετησίως (Smith, 1997). Τον Νοέμβριο του 1997 ο συνολικός αριθμός ιστοσελίδων υπολογίστηκε πάνω από 200 εκατομμύρια (Bharat και Broder, 1998). Εάν και οι δύο εκτιμήσεις για τον αριθμό των ιστοσελίδων είναι σωστές, τότε το ποσοστό αύξησης είναι υψηλότερο από τις προβλέψεις του Smith, δηλ., αυτό σημαίνει ότι η αύξηση είναι υπερδιπλάσια. Σε μια άλλη εκτίμηση, ο Monier, ανώτερος τεχνικός υπάλληλος της AltaVista, υπολόγισε ότι το μέγεθος των προσιτών πληροφοριών του ιστού, έχει αυξηθεί από 50 εκατομμύρια σελίδες σε 100.000 ιστοσελίδες το 1995, σε 100 έως και 150 εκατομμύρια σελίδες σε 600.000 ιστοσελίδες τον Ιούνιο του 1997 (Monier, 1998).

Λαμβάνοντας υπόψη το μεγάλο αριθμό ιστοσελίδων, είναι γεγονός ότι οι χρήστες του διαδικτύου αυξάνονται χρησιμοποιώντας τις μηχανές αναζήτησης και τις υπηρεσίες της αναζήτησης για να βρίσκουν συγκεκριμένες πληροφορίες. Σύμφωνα με τους Brin και Paige, η World Wide Web Worm ισχυρίζεται ότι έχει χειριστεί έναν μέσο όρο 1.500 ερωτήσεων ημερησίως τον Απρίλιο του 1994, και η AltaVista υποστηρίζει ότι έχει εξυπηρετήσει 20 εκατομμύρια ερωτήσεις τον Νοέμβριο του 1997.

Τα αποτελέσματα της έρευνας χρηστών του ιστού, τον Απρίλιο 1998, δείχνουν ότι περίπου το 86% των ανθρώπων βρίσκουν μια χρήσιμη ιστοσελίδα μέσω των μηχανών αναζήτησης, και το 85% τις βρίσκουν μέσω των αναφορών σε σελίδες. Οι άνθρωποι χρησιμοποιούν τώρα τις μηχανές αναζήτησης όλο και περισσότερο καθώς σερφάρουν στο διαδίκτυο για την ανεύρεση πληροφοριών.

1.3. Οι Μηχανές Αναζήτησης (Search Engines)

Οι δυνατότητες που παρέχει το διαδίκτυο είναι απεριόριστες και οι πληροφορίες για οτιδήποτε χρειαζόμαστε σίγουρα υπάρχουν κάπου εκεί έξω. Είναι όμως χρήσιμες μόνο αν καταφέρουμε να τις βρούμε: πώς και με ποιον τρόπο γίνεται η εύρεση αυτών των πληροφοριών στο διαδίκτυο; Οι μηχανές αναζήτησης είναι η λύση στο πρόβλημα, είναι οι οδηγοί μας στο διαδίκτυο. Είναι αυτές που μας επιτρέπουν να εκμεταλλευτούμε αποδοτικά τις δυνατότητες και τις πληροφορίες που μας παρέχει το διαδίκτυο.

Το διαδίκτυο είναι αχανές και ο όγκος πληροφοριών του αυξάνεται καθημερινά με εντυπωσιακό ρυθμό. Με τον ίδιο ρυθμό αυξάνονται και οι υπηρεσίες *on-line*, οι οποίες ενημερώνουν το χρήστη σε πραγματικό χρόνο για τις εξελίξεις σε διάφορους τομείς. Ένα παράδειγμα είναι το χρηματιστήριο ή οι διεθνείς αγώνες ποδοσφαίρου. Μία σύνδεση *on-line* μέσω διαδικτύου επιτρέπει να παρακολουθούμε τις εξελίξεις στη Σοφοκλέους ή στο Μπερναμπέου της Ισπανίας σε πραγματικό χρόνο. Είναι πλέον γνωστό ακόμα και στον αδαή χρήστη ότι το διαδίκτυο περιλαμβάνει κάθε είδους πληροφορία και είναι ίσως το πρώτο μέρος από το οποίο ξεκινά κάποιος την αναζήτηση στοιχείων για οποιονδήποτε και οτιδήποτε: από μία απλή συνταγή μαγειρικής μέχρι τα ταξίδια στο διάστημα, και από τη ζωή του Mozart μέχρι τις πληροφορίες και τις φωτογραφίες για αρχαία νομίσματα. Το διαδίκτυο είναι σίγουρα η μεγαλύτερη βιβλιοθήκη του κόσμου που μπορεί να επισκεφθεί ο καθένας ανά πάσα στιγμή.

Σίγουρα κάπου εκεί έξω βρίσκονται επίσης αρκετοί άνθρωποι με τους οποίους έχουμε κοινά ενδιαφέροντα, προβληματισμούς, χόμπι κ.λ.π., ασχολούνται με το αντικείμενο που μας ενδιαφέρει και κατέχουν πληροφορίες που θα θέλαμε να γνωρίζουμε. Πρέπει να θυμόμαστε πάντα, ότι το διαδίκτυο δεν το ελέγχει κανείς και άρα οι πληροφορίες του δεν είναι οργανωμένες με τέτοιον τρόπο, ώστε όλες να είναι άμεσα και χωρίς κόπο προσβάσιμες στον οποιονδήποτε. Υπάρχουν εκατοντάδες εκατομμύρια σελίδες που έχουν δημιουργηθεί από απλούς χρήστες, εταιρείες, οργανισμούς κ.λπ., οι οποίες μέρα με τη μέρα αυξάνονται όλο και περισσότερο, ενώ από τις ήδη υπάρχουσες, άλλες αλλάζουν τοποθεσία (διεύθυνση), άλλες τροποποιούνται και άλλες καταργούνται τελείως. Η χαρτογράφηση ενός τέτοιου δικτύου είναι κάτι παραπάνω από δύσκολη. Παρ' όλα αυτά, είναι γεγονός ότι υπάρχουν οι τρόποι για να βρούμε αυτό που θέλουμε, όταν το θέλουμε. Η εύρεση των πληροφοριών στο διαδίκτυο επιτυγχάνεται με τις μηχανές αναζήτησης.

1.3.1. Η προϊστορία των Μηχανών Αναζήτησης

Οι μηχανές αναζήτησης έχουν μικρή προϊστορία, λιγότερο από μια δεκαετία, και σε αυτό το τμήμα της εργασίας θα γίνει μια πολύ συνοπτική περίληψη αυτής της προϊστορίας.

Πριν από τις μηχανές αναζήτησης, επικρατούσε χάος στο διαδίκτυο. Εάν κάποιος ήθελε να αναζητήσει κάποια πληροφορία στο διαδίκτυο έπρεπε να ξέρει την ακριβή διεύθυνσή του. Το πρώτο πραγματικά σημαντικό βήμα για να αποφευχθεί-εξαλειφθεί το χάος, αλλά και προς έναν βαθμό οργάνωσης των περιεχομένων του διαδικτύου ήταν η ανάπτυξη των ***"Gopher"*** (Randolph Hoch, 2001), συλλογές διευθύνσεων του διαδικτύου βασισμένες σε υπολογιστή που καταχωρούνταν με τη βοήθεια ενός *μενού* ***(menu)*** (Ο όρος "*Gopher*" προέρχεται από τη μασκότ του πανεπιστημίου της Μινεσότα, όπου εκεί δημιουργήθηκε το πρώτο *"Gopher"* του διαδικτύου). Τα *"Gopher"* δεν ήταν βασισμένα σε *HTML* και το ευρετήριό τους στηριζόταν στους τίτλους των αρχείων ή στις πολύ συνοπτικές περιγραφές, αλλά εάν γνώριζε κάποιος πως να φτάσει σε έναν *"Gopher"* θα του επέτρεπε να κάνει *"download"* τα επιλεγμένα αρχεία.

Το *"Gopher"* επικράτησε για μερικά έτη πριν επισκιαστεί από τη γρήγορη ανάπτυξη του *Παγκόσμιου Ιστού* ***(World Wide Web)***, ο οποίος επέτρεπε τη χρήση των *υπερσυνδέσμων* ***(Hyperlinks)***, την *ευρετηρίαση όλου του κειμένου* ***(Full-Text Searching)***, των *γραφικών φυλλομετρητών* ***(Graphical Browser)***, όπως επίσης της εύχρηστης και ιδιαίτερα της *διαλογικής τεχνολογίας* ***(Interactive Technology)*** αλλά και της ανάπτυξης των μηχανών αναζήτησης.

Η πρώτη μηχανή αναζήτησης που αναπτύχθηκε ήταν η ***WebCrawler***, η οποία προήλθε από το πανεπιστήμιο της Ουάσιγκτον και άρχισε να χρησιμοποιείται από τον Απρίλιο του 1994. Μέσα σε ένα χρόνο τρεις ανταγωνιστές εμφανίστηκαν στο προσκήνιο: η μηχανή αναζήτησης ***Lycos***, η μηχανή αναζήτησης ***Infoseek***, και η μηχανή αναζήτησης ***OpenText***. Στα τέλη του 1995 οι ***AltaVista*** και ***Excite*** έκαναν την εμφάνισή τους. Είναι ενδιαφέρον να τονιστεί, ότι κατά ένα μεγάλο μέρος η αναζήτηση γινόταν με παρόμοιο τρόπο με αυτόν στις σημερινές μηχανές αναζήτησης, όπως με τη χρήση των *λογικών τελεστών* ***(Boolean)***, με τη χρήση της *αποκοπής* ***(Truncation)*** κ.λ.π. Δυστυχώς -και ο αντίκτυπος αυτών συνεχίζεται μέχρι και σήμερα- καμία από αυτές τις μηχανές αναζήτησης δεν εκμεταλλεύθηκε την *τεχνολογία αναζήτησης* ***(Searching Technology)***. Επιπλέον, ούτε οι μηχανές αναζήτησης, αλλά ούτε και οι θεματικοί

κατάλογοι, δεν εκμεταλλεύθηκαν την εκτενή *θεωρία ταξινόμησης* ***(Subject Classification Theory)*** και την πρακτική των τελευταίων εκατό χρόνων. Αυτά τα σημεία σχετίζονται με έναν πολύ πρακτικό τρόπο δεδομένου, ότι ο επαγγελματίας περιηγητής πρέπει να αναγνωρίσει το γεγονός ότι οι περισσότερες μηχανές αναζήτησης αναπτύχθηκαν και αναπτύσσονται για τον καθημερινό περιηγητή και όχι για εκείνους που στοχεύουν να αποκτήσουν τα πλεονεκτήματα των περιπλοκότερων προσεγγίσεων και τεχνικών.

Η μηχανή αναζήτησης ***HotBot*** εμφανίστηκε το 1996 και η ***Northern Light*** το 1997 (Randolph Hoch, 2001). Η *HotBot* εμφάνισε μια πιο περίπλοκη, αλλά συγχρόνως μια εύχρηστη *διεπαφή* ***(interface)*** και ήταν συνδεδεμένη με μια πολύ μεγάλη βάση δεδομένων (μέχρι το τέλος του 1997, η βάση της θεωρείτο η μεγαλύτερη διαθέσιμη). Η *Northern Light* επέφερε μια ολοκλήρωση στην αναζήτηση του ιστού και των πληροφοριών που περιέχει. Η μηχανή αναζήτησης ***Google*** εμφανίστηκε το 1998, και η ταξινόμηση αρχείων της μαζί με ένα εξαιρετικά απλό *interface* συνδυάστηκαν αποτελεσματικά για να παραγάγουν μια μηχανή που κατάφερε γρήγορα να αποκτήσει δημοτικότητα και μεταξύ των περιστασιακών, αλλά και των μακροπρόθεσμων ερευνητών. Εν τω μεταξύ, η κούρσα για τη μεγαλύτερη σε αριθμό σελίδων μηχανή αναζήτησης είχε μειωθεί κάπως, μέχρι την εμφάνιση το 1999 της μηχανής αναζήτησης ***Fast Search***, η οποία αποτελούνταν από μια βάση δεδομένων πάνω από 200 εκατομμύρια αρχεία. Αυτό το κίνητρο, μαζί με κάποια άλλα χαρακτηριστικά, σήμανε ότι ο ανταγωνισμός για το μέγεθος είχε αρχίσει πάλι, με τέσσερις μηχανές να έχουν φτάσει τα 200 εκατομμύρια αρχεία μέχρι τον Ιανουάριο του 2000.

Μεταξύ των "πρώτων" μηχανών αναζήτησης, η *Open Text* έκανε το πρώτο μεγάλο βήμα. Στις αρχές του 1998 δεν ήταν πλέον διαθέσιμη, είχε καταργηθεί. Πιθανώς, θα υπάρξουν περισσότερες εξαφανίσεις μηχανών αναζήτησης κατά τη διάρκεια των επόμενων χρόνων, και ίσως να έχουμε την εμφάνιση τουλάχιστον μιας ή δύο κύριων (σημαντικών) μηχανών αναζήτησης. Στο μεταξύ, οι αλλαγές στις υπάρχουσες μηχανές αναζήτησης θα συνεχίζονται, αν και πολλές από αυτές είναι κατά ένα μεγάλο μέρος αρκετά επιφανειακές παρά ένα αναπόσπαστο τμήμα της θεωρίας αναζήτησης του web. Μπορούμε να ελπίσουμε ότι οι δημιουργοί αυτών των εργαλείων θα κινηθούν προς την ενίσχυση των δυνατοτήτων αναζήτησης, υπάρχουν ενδείξεις ότι οι ανταγωνιστικές τάσεις θα βοηθήσουν σε αυτό. Και έτσι σε μερικές περιπτώσεις, θα είναι ένα βήμα προς τη σωστή κατεύθυνση, εάν η μηχανή αναζήτησης εκπληρώνει τις υποσχέσεις της.

Όπως και ο υπόλοιπος επιχειρηματικός κόσμος, οι εταιρείες των μηχανών αναζήτησης είναι εξαιρετικά ευαίσθητες στις τάσεις της εποχής με αποτέλεσμα να επηρεάζονται από αυτές. Το 1996 και το 1997, η τάση ήταν να σιγουρευτεί κάποιος, ότι η μηχανή αναζήτησης που χρησιμοποιούσε ή κατασκεύαζε ήταν μια "προηγμένη" έκδοση σε σχέση με τις υπόλοιπες, ανεξάρτητα από το εάν η προηγμένη έκδοση είχε πραγματικά περισσότερες δυνατότητες από τις υπόλοιπες.

Μεγαλύτερης σε σημασία όσων αφορά τα οφέλη, το 1998 έφερε την "προσωποποίηση" της πύλης. Η ιδέα της "προσωποποίησης της πύλης" ή της "*πύλης Ιστού"* ***(Web Gateway)*** ήταν έκδηλη στις εντοπισμένες και στις επιλεγμένες από τους χρήστες κατηγορίες ειδήσεων που εμφανίζονται στην αρχική σελίδα, όπως τον τοπικό καιρό και τα προγράμματα των τοπικών καναλιών της τηλεόρασης, την προσωπική παρακολούθηση των χαρτοφυλακίων, τα προσωπικά ημερολόγια, κ.λ.π. Η επιθυμία των δημιουργών των μηχανών αναζήτησης να ακολουθήσουν το παράδειγμα αυτό και της συνειδητοποίησης, ότι αυτή η προσέγγιση θα μπορούσε να επιφέρει έσοδα λόγω διαφήμισης, αυτά τα δύο πολύ στενά συνδεδεμένα πρότυπα έγιναν γρήγορα το καθολικό πρότυπο των επιχειρήσεων για τις σημαντικότερες μηχανές αναζήτησης. Αν και πολλοί χρήστες δεν το είχαν συνειδητοποιήσει ακόμα, αυτή η προσέγγιση της προσωποποίησης ήταν ένα σημαντικό βήμα προόδου από την πλευρά, ότι έτσι επιτεύχθηκε να φτάσει ο ιστός στο επίπεδο μιας οικογένειας και γενικά να είναι πάντα προσιτός, απλός, να χρησιμοποιείται πιο συχνά, και, επιπλέον, να παρέχει τα συγκεκριμένα και προφανή οφέλη.

Τα έτη 1999 και 2000 υπήρξε μεγάλη προσπάθεια στην έννοια της *πύλης* ***(portal)***. Κατά τη διάρκεια του πρώτου έτους, τα προστιθέμενα εργαλεία (όπως οι κατάλογοι, κ.λ.π.) κυρίως σχεδιάστηκαν στην αρχική σελίδα με την ελπίδα, ότι οι χρήστες θα τα χρησιμοποιούσαν. Το 1999 υπήρξε μια σημαντική μετατόπιση προς την αυτόματη ενσωμάτωση του περιεχομένου των σελίδων στα τελικά αποτελέσματα -ταυτοχρόνως που η βάση δεδομένων του ιστού της μηχανής αναζήτησης ευρετηριάζεται, ψάχνει τον θεματικό κατάλογο, τον κατάλογο της επιχείρησης, κ.λ.π., και παρουσιάζει αυτά τα αποτελέσματα μαζί με τα κανονικά αποτελέσματα της αναζήτησης. Αυτή η ύπαρξη (ολοκλήρωση) των εργαλείων (πόρων) έχει βελτιώσει σημαντικά την ποιότητα των αποτελεσμάτων της αναζήτησης αφού παρέχει στον ερευνητή ιδιαίτερα σχετικά με το θέμα της αναζήτησης αποτελέσματα, χωρίς να πρέπει να εκτελεσθεί η αναζήτηση χωριστά σε διαφορετικά εργαλεία.

Το επόμενο βήμα έχει να κάνει τόσο με τους χρήστες όσο και με τους

δημιουργούς των μηχανών αναζήτησης. Τα εργαλεία που λαμβάνουν την προσοχή των χρηστών θα διατηρηθούν, θα ενισχυθούν, θα αντιγραφούν, και θα αξιολογηθούν. Το πρόβλημα, αρχικά με τις μηχανές αναζήτησης του ιστού, είναι ότι το πιθανό πρόσωπο που θα τις χρησιμοποιήσει δεν είναι ο τυπικός (καθημερινός) χρήστης των μηχανών αναζήτησης. Ο τυπικός χρήστης θα ενδιαφερόταν λιγότερο για τα περιπλοκότερα και για τα προσανατολισμένα στην έρευνα χαρακτηριστικά γνωρίσματα. Ο βαθμός στον οποίο αυτό ισχύει είναι πολύ εμφανής, εάν εξετάσουμε τις αναζητήσεις ενός τυπικού χρήστη. Η μηχανή αναζήτησης *Lycos* παρέχει έναν ενδιαφέρον, εν τούτοις μερικές φορές πολύ συμπιεσμένο, κατάλογο αγαπημένων διευθύνσεων αναζητήσεων. Σε μια εβδομάδα, οι 50 κυριότερες αναζητήσεις περιλαμβάνουν 46 αναζητήσεις που αναφέρονται στην ψυχαγωγία, τον αθλητισμό, ή στις κατηγορίες των παιχνιδιών. Η σχετικότητα αυτού δεν είναι ένα ζήτημα σνομπισμού των πληροφοριών, αλλά η ανάγκη να αντιμετωπιστεί η πραγματικότητα και η πρωταρχική θέση που υποστηρίζει, ότι οι περισσότερες μηχανές αναζήτησης δεν κερδίζουν χρήματα από τον ερευνητή που χρησιμοποιεί τον ιστό για επαγγελματικούς λόγους. Το μόνο θετικό είναι ότι το σύνολο των χρηστών αυξάνεται, καθώς και ο αριθμός των ανθρώπων που χρησιμοποιούν τις μηχανές αναζήτησης για επαγγελματικούς λόγους, για επενδύσεις, όπως επίσης και για τη βασική εκπαίδευσή τους σε θέματα επιστήμης όπως της ανθρωπότητας, των επιχειρήσεων, και της ιατρικής, ίσως να αυξάνεται γρηγορότερα αυτός ο αριθμός, δηλαδή των πιο διανοητικών αναζητήσεων. Υπάρχουν πολλοί περισσότεροι λόγοι για τους δημιουργούς των μηχανών αναζήτησης να δώσουν μεγαλύτερη προσοχή στον ακραίο ερευνητή. Αλλά και ο σοβαρός ερευνητής πρέπει επίσης να χρησιμοποιήσει τα κυριότερα χαρακτηριστικά γνωρίσματα μιας μηχανής έτσι ώστε εκείνα τα χαρακτηριστικά γνωρίσματα να παραμείνουν και να ενισχυθούν.

1.3.2. Τα είδη των Μηχανών Αναζήτησης

Τα εργαλεία που χρησιμοποιούνται για την ανεύρεση πληροφοριών στο διαδίκτυο διαχωρίζονται στις παρακάτω κατηγορίες:

- *Μηχανές Αναζήτησης* ***(Search Engines):*** Οι μηχανές αναζήτησης αποτελούν το βασικό εργαλείο τόσο για τον αρχάριο όσο και για τον πιο εξοικειωμένο χρήστη του διαδικτύου. Οι μηχανές αναζήτησης διακρίνονται στις παρακάτω κατηγορίες:
  - ✓ *Αναζήτηση Ελευθέρου Κειμένου* ***(Free Text Search Engines)****:* Μπορεί

κάποιος να χρησιμοποιήσει μια σειρά από λέξεις συνδέοντάς τις με όρους όπως το AND και το OR ή με τελεστές της πρόσθεσης (+) και της αφαίρεσης (-) ή χρησιμοποιώντας διπλά εισαγωγικά στην αρχή και στο τέλος της φράσης. Δηλαδή χρησιμοποιούν για την αναζήτηση τους τελεστές Boolean. Παραδείγματα τέτοιων μηχανών αναζήτησης είναι οι Alta Vista, HotBot, Go.com, Google, Lycos, Northernlight, WebCrawler.

- ✓ *Θεματικοί Κατάλογοι- Θεματικά Ευρετήρια* ***(Directory or Index Based Search Engines)***: Αυτού του είδους οι μηχανές αναζήτησης είναι πιο εύκολες στη χρήση τους μιας και βασίζονται στην ιεραρχική προσέγγιση των πληροφοριών, ξεκινώντας από πιο γενικά θέματα και τίτλους για να καταλήξει στα ζητούμενα του χρήστη. Παραδείγματα τέτοιων μηχανών αναζήτησης είναι οι Excite, Metaplus, Yahoo.
- ✓ *Μηχανές Πολλαπλής Αναζήτησης* ***(Multi or Meta Search Engines)***: Οι οποίες ονομάζονται και ***meta-crawlers***, δίνουν τη δυνατότητα ταυτόχρονης αναζήτησης σε διάφορα εργαλεία αναζήτησης και όχι σε μια απλή βάση δεδομένων. Παραδείγματα τέτοιων μηχανών αναζήτησης είναι οι Search.com, Ixquick, WebSearch.com.
- ✓ *Αναζήτηση Φυσικής Γλώσσας* ***(Natural Language Search Engines)****:* Αυτές οι μηχανές αναζήτησης λύνουν το πρόβλημα της εύρεσης της πληροφορίας χρησιμοποιώντας διαφορετική προσέγγιση. Δεν χρησιμοποιούν τους τελεστές Boolean, αλλά οι ερωτήσεις διατυπώνονται από τον χρήστη χρησιμοποιώντας τη φυσική γλώσσα. Παραδείγματα τέτοιων μηχανών αναζήτησης είναι η AskJeeves.
- ✓ *Μηχανές Αναζήτησης Συγκεκριμένων Θεμάτων* ***(Subject Specific Search Engines)****:* Αυτές οι μηχανές αναζήτησης δεν επιχειρούν να καταχωρήσουν σε ευρετήριο όλο το δίκτυο. Αντίθετα, εστιάζουν στην αναζήτηση των δικτυακών τόπων ή σελίδων εντός ενός προκαθορισμένου τομέα θεμάτων, γεωγραφικής περιοχής ή είδους πόρου. Επειδή αυτές οι ειδικευμένες μηχανές αναζήτησης στοχεύουν στην εκτενή κάλυψη εντός ενός μόνο τομέα και στο εύρος κάλυψης μεταξύ των θεμάτων, συχνά μπορούν να καταχωρήσουν σε ευρετήριο έγγραφα που δεν περιλαμβάνονται ακόμη και στις μεγαλύτερες βάσεις δεδομένων των μηχανών αναζήτησης. Γι' αυτό το λόγο, είναι συχνά χρήσιμο σημείο εκκίνησης για συγκεκριμένες αναζητήσεις.

- *Πύλες και Εικονικές Βιβλιοθήκες* ***(Gateways and Virtual Libraries):*** Καθώς το διαδίκτυο έχει μεγαλώσει, τόσο επιτακτικότερη κρίνεται και η ανάγκη για εύρεση της πληροφορίας σε αυτό. Βέβαια αυτό οδηγεί σε δύο κύρια προβλήματα- στην ανεύρεση της πληροφορίας και στην εκτίμηση αυτής όταν βρεθεί. Η εικονική βιβλιοθήκη είναι η απάντηση σε αυτά τα δύο ερωτήματα. Είναι σχεδιασμένες έτσι, ώστε να προσφέρουν γρήγορους και εύκολους τρόπους για την εύρεση ποιοτικής πληροφορίας, η οποία θα βοηθήσει το χρήστη στην εργασία του. Συγκεκριμένα οι πύλες και οι εικονικές βιβλιοθήκες είναι συλλογές πηγών πληροφορίας υψηλής ποιότητας ενός συγκεκριμένου θεματικού τομέα, που αφορούν ένα καθορισμένο κοινό.

- *Οι Ευφυείς Πράκτορες* ***(Intelligent Agents):*** Είναι κοινώς αποδεκτό ότι με την πάροδο του χρόνου η αποτελεσματική χρήση του διαδικτύου γίνεται όλο και πιο δύσκολη. Είναι λοιπόν προφανές ότι ο εντοπισμός και η πρόσβαση σε πληροφορίες, η ανάκτηση, το φιλτράρισμα και η αξιολόγησή τους είναι μια ιδιαίτερα δύσκολη διαδικασία για να πραγματοποιηθεί επιτυχώς από έναν ανθρώπινο χρήστη. Οι μηχανές αναζήτησης στην παραδοσιακή τους μορφή αποδεικνύονται υπό αυτές τις συνθήκες αναποτελεσματικές και η ανάγκη χρήσης σύγχρονων τεχνολογιών γίνεται απαιτητική. Οι ευφυείς πράκτορες προσπαθούν να προσαρμοστούν στο προφίλ και στις ανάγκες των χρηστών τους και να πραγματοποιήσουν επιτυχώς την παραπάνω διαδικασία κάνοντας ότι ακριβώς θα έκαναν και οι χρήστες τους αν είχαν τον απαραίτητο χρόνο.

Οι περισσότερες από τις μηχανές αναζήτησης που αναφέρθηκαν παραπάνω θα μελετηθούν πολύ αναλυτικά στα επόμενα κεφάλαια αυτής της εργασίας.

1.3.3. Τα Βασικά Μέρη των Μηχανών Αναζήτησης

Η μηχανή αναζήτησης μπορεί να θεωρηθεί ότι αποτελείται γενικά από πέντε κύρια μέρη (Randolph Hock, 2001):

- *Το ειδικό λογισμικό* ***(robot, spider, crawler κ.λ.π.)***
- *Η βάση δεδομένων ή αλλιώς ο κατάλογος* ***(database of information)***
- *Το πρόγραμμα ευρετηρίασης και το ευρετήριο* ***(the indexing program and the index)***

- *Η μηχανή ανάκτησης, μηχανή αναζήτησης, το ειδικό πρόγραμμα* ***(retrieval engine)*** και
- *Η γραφική διεπαφή* ***(the graphical HTML (Hyper-text Markup Language) interface)***

➢ **<u>Το Ειδικό Λογισμικό</u>**

Το ειδικό λογισμικό (robot, spider, crawler κ.λ.π.) είναι προγράμματα που επισκέπτονται σελίδες στον ιστό για να:

1. προσδιορίσουν τις νέες σελίδες-διευθύνσεις που πρόκειται να προστεθούν στη μηχανή αναζήτησης και

2. να προσδιορίσουν σελίδες-διευθύνσεις που έχουν ήδη εξερευνηθεί και έχουν αλλάξει.

Το ειδικό αυτό λογισμικό συγκεντρώνει πληροφορίες για το περιεχόμενο των σελίδων από τις διευθύνσεις που επισκέπτεται και παρέχει αυτές τις πληροφορίες στη *βάση δεδομένων* της μηχανής αναζήτησης. Πολλά θα μπορούσαν να ειπωθούν για το πώς γίνεται αυτό, αλλά για τον ερευνητή αυτά δεν έχουν σημασία αν και γίνεται κατανοητό, γιατί μερικές μηχανές αναζήτησης βρίσκουν ορισμένες σελίδες που άλλες μηχανές αναζήτησης δεν τις εμφανίζουν, ακόμα και στην περίπτωση που η σελίδα βρίσκεται στη βάση δεδομένων της δεύτερης μηχανής αναζήτησης. Σε πολλές μηχανές αναζήτησης, οι δημοφιλέστερες σελίδες (όπως εκείνες που τις επισκέπτονται πολύ συχνά οι χρήστες ή εκείνες που έχουν πολλούς *συνδέσμους* ***(links)***) εξερευνούνται λεπτομερώς και συχνότερα από τους *crawlers* από ότι οι λιγότερο-δημοφιλείς σελίδες.

Το ειδικό αυτό λογισμικό μπορεί να προγραμματιστεί να εξερευνεί τις ιστοσελίδες σε *βάθος* ***(depth)*** ή σε *εύρος* ***(breadth)***, ή και τα δύο. Εκείνο που προγραμματίζεται για σε βάθος εξερεύνηση όχι μόνο προσδιορίζει τις κύριες περιοχές-σελίδες, αλλά προσδιορίζει και τις συνδεόμενες σελίδες σε αυτές. Αυτό το λογισμικό που προγραμματίζεται γιατί σε εύρος εξερεύνησης των περιοχών, ενδιαφέρεται χαρακτηριστικά για την εύρεση περισσότερο των κύριων σελίδων, αλλά όχι απαραιτήτως και για τον προσδιορισμό όλων των συνδεόμενων σελίδων μιας κύριας σελίδας. Δεδομένου ότι οι μηχανές αναζήτησης στην σημερινή εποχή έχουν εξελιχθεί πάρα πολύ και έχουν γίνει ακόμη περισσότερο ανταγωνιστικές, έχει υπάρξει μια τάση συγχώνευσης του βάθους και του εύρους.

- **Η Βάση Δεδομένων της Μηχανής**

Η συνολική συλλογή της πληροφορίας που αποθηκεύεται για καθεμιά από τις σελίδες του ιστού αποτελεί τη *βάση δεδομένων της μηχανής αναζήτησης* ***(the engine's database)***. Η συλλογή αποτελείται από σελίδες που έχουν προσδιοριστεί από τους *crawlers*, αλλά όλο και περισσότερο περιλαμβάνει επίσης σελίδες που προσδιορίζονται με άλλες πηγές ή τεχνικές. Ένας πολύ μεγάλος αριθμός σελίδων που προστίθεται στις μηχανές αναζήτησης προέρχεται από την απευθείας αίτηση καταχώρησης των δημιουργών της ιστοσελίδας. Εάν εξεταστεί η αρχική σελίδα οποιασδήποτε μηχανής αναζήτησης, θα υπάρχει πιθανώς ένας σύνδεσμος που επιτρέπει στον καθένα μας να καταχωρήσει μια σελίδα στη συγκεκριμένη μηχανή αναζήτησης. Εφ' όσον η σελίδα δεν αποτελεί περίπτωση ***'spamming'***, οι υποβληθείσες σελίδες θα προστεθούν πιθανώς στη βάση δεδομένων της μηχανής αναζήτησης. Όλοι ή οι περισσότεροι σχεδιαστές των μηχανών αναζήτησης εξετάζουν τις υποβληθείσες σελίδες για την περίπτωση του ***spam*** (προγράμματα που χρησιμοποιούνται από τους προγραμματιστές προσπαθώντας να εξαπατήσουν για να αυξήσουν παράνομα τις πιθανότητες ανάκτησης μιας σελίδας). Μια υπηρεσία μπορεί επίσης να εφαρμόσει και άλλα κριτήρια, αλλά με την εξαίρεση του *spam*, οι πιθανότητες είναι πολύ καλές ότι η υποβληθείσα σελίδα θα καταλήξει στη βάση δεδομένων της μηχανής αναζήτησης.

Άλλες πηγές μπορούν επίσης να τροφοδοτήσουν τη βάση δεδομένων της μηχανής αναζήτησης. Η βάση δεδομένων μπορεί, παραδείγματος χάριν, να περιέχει σελίδες με τους υπαγόμενους τίτλους από ένα θεματικό ευρετήριο όπως το Open Directory ή το Yahoo.

Είναι μερικές φορές εύκολο να ξεχάσει κάποιος ότι όταν χρησιμοποιεί μια μηχανή αναζήτησης, δεν ψάχνει αυτή άμεσα τον ιστό, αλλά ψάχνει μια βάση δεδομένων που περιέχει τα αρχεία της, τα οποία περιγράφουν ένα μέρος εκείνων των σελίδων που υπάρχουν στον ιστό. Γνωρίζοντας αυτό μπορεί να τον βοηθήσει να αποφύγει τις μη ρεαλιστικές προσδοκίες για αυτό που μια μηχανή αναζήτησης μπορεί πραγματικά να επιτύχει.

- **Το Πρόγραμμα Ευρετηρίασης και το Ευρετήριο**

Από την άποψη ποίες σελίδες θα ανακτηθούν πραγματικά από μια ερώτηση ενός χρήστη, η ευρετηρίαση μπορεί να κατέχει σημαντικότερο ρόλο από τη διαδικασία του *Crawling*. Το *πρόγραμμα ευρετηρίασης* ***(the indexing program)*** εξετάζει τις πληροφορίες που αποθηκεύονται στη βάση δεδομένων και δημιουργεί τις κατάλληλες

καταχωρήσεις στο *ευρετήριο* ***(index)***. Όταν υποβάλλουμε μια ερώτηση, αυτό χρησιμοποιείται προκειμένου να προσδιοριστούν τα αρχεία που είναι σχετικά με την ερώτηση του χρήστη.

Οι περισσότερες μηχανές αναζήτησης υποστηρίζουν ότι συντάσσουν ευρετήριο εξετάζοντας όλες τις λέξεις από κάθε σελίδα. Το σημαντικότερο είναι αυτό που οι μηχανές αναζήτησης επιλέγουν να θεωρήσουν ως 'λέξη'. Μερικές έχουν έναν κατάλογο από ***'stop words'*** (μικρές, κοινές λέξεις που θεωρούνται αρκετά ασήμαντες και μπορούν να αγνοηθούν) που δεν περιέχονται στο ευρετήριο. Μερικές μηχανές αναζήτησης δεν περιέχουν στο ευρετήριό τους τέτοιες λέξεις, όπως τα άρθρα και οι σύνδεσμοι. Άλλες παραλείπουν λέξεις ευρέως χρησιμοποιημένες αλλά και ενδεχομένως πολύτιμες όπως 'web' και 'internet'. Μερικές φορές και οι αριθμοί παραλείπονται. Βέβαια κατά τη διάρκεια των δύο τελευταίων ετών, γενικά, οι μηχανές αναζήτησης μεταχειρίζονται λιγότερες λέξεις ως ***'stop words'***.

Όλες οι κύριες μηχανές αναζήτησης συντάσσουν το ευρετήριο τους εξετάζοντας τον τίτλο και τη ***URL(Uniform Resource Locator)***. Τα ***Metatags*** ευρετηριάζονται συνήθως, αλλά όχι πάντα (Τα *Metatags* είναι λέξεις, φράσεις, ή προτάσεις που τοποθετούνται σε ένα ειδικό τμήμα του κώδικα ***HTML (Hypertext Markup Language)*** περιγράφοντας το περιεχόμενο της σελίδας). Τα *Metatags* δεν μπορούμε να τα δούμε όταν επισκεπτόμαστε μια σελίδα, ωστόσο μπορούμε να τα δούμε εάν επιθυμούμε να ζητήσουμε από το *φυλλομετρητή* ***(browser)*** να μας παρουσιάσει την ***'page source'***. Είναι κατανοητό το πόσο χρήσιμα αποδεικνύονται τα περιεχόμενα των *metatags* για την ανάκτηση των πληροφοριών. Εντούτοις, μερικές μηχανές αναζήτησης εσκεμμένα δεν ευρετηριάζουν τα *metatags,* επειδή αυτά αποτελούν μέρος της σελίδας που μπορεί πολύ εύκολα να καταχραστεί και να αλλοιωθεί. Αυτή η προσοχή έχει σαν αποτέλεσμα τον μη εντοπισμό πολύτιμων πληροφοριών ευρετηρίασης.

Οι γνώστες της *HTML* γνωρίζουν ότι τα *πλαίσια* ***(frames)*** χρησιμοποιούνται σε εκατομμύρια ιστοσελίδες. (Τα πλαίσια είναι μια συσκευή της *HTML* που μεταχειρίζεται τα διαφορετικά μέρη μιας σελίδας ως ανεξάρτητα *παράθυρα* ***(windows)***). Μερικές μηχανές αναζήτησης δεν συντάσσουν ευρετήριο εξερευνώντας τα πλαίσια, με αυτόν τον τρόπο προκαλείται πιθανή απώλεια σχετικών περιοχών με το θέμα της ερώτησης του χρήστη. Αυτό το μειονέκτημα αντισταθμίζεται από το γεγονός ότι ο δημιουργός της ιστοσελίδας δημιουργεί μια έκδοση σελίδας χωρίς πλαίσια καθώς επίσης και την έκδοση αυτής με πλαίσια. Επιπλέον, με την εξέλιξη της κατασκευής των ιστοσελίδων, τα πλαίσια χρησιμοποιούνται πολύ λιγότερο από ότι στο παρελθόν.

Κατανοώντας αυτούς τους διαφορετικούς τρόπους της πολιτικής της ευρετηρίασης γίνεται αντιληπτό, γιατί οι σχετικές σελίδες, ακόμα και αν είναι καταχωρημένες στη βάση δεδομένων της μηχανής, δεν ανακτώνται μετά από μερικές αναζητήσεις. Εξηγεί επίσης, γιατί μια σελίδα μπορεί να ανακτηθεί από μια μηχανή και όχι από μια άλλη, ακόμα και όταν η ίδια σελίδα βρίσκεται και στις δύο μηχανές.

- **Η Μηχανή Ανάκτησης**

Αυτό είναι το πρόγραμμα που λαμβάνει την ερώτηση ενός χρήστη και ψάχνει έπειτα το ευρετήριο για να προσδιορίσει και να παραδώσει τα αρχεία που ταιριάζουν με την ερώτησή του. Στην πραγματικότητα, δύο σημαντικά γεγονότα συμβαίνουν κατά τη διάρκεια αυτής της διαδικασίας:

1. η μηχανή ανάκτησης προσδιορίζει τα αρχεία που αναφέρονται στην ερώτηση με τη βοήθεια ενός "αλγορίθμου ανάκτησης", και
2. η μηχανή αναζήτησης κατόπιν ταξινομεί τα ανακτημένα αρχεία σε μια συγκεκριμένη σειρά και τα εμφανίζει στο χρήστη.

Αυτά μπορούν να συμβούν λίγο ή πολύ ταυτόχρονα, ή μπορούν να είναι αρκετά ευδιάκριτες διαδικασίες.

Οι *Αλγόριθμοι Ανάκτησης* ***(Retrieval Algorithms)*** είναι προγράμματα που χρησιμοποιούνται για την εφαρμογή των κριτηρίων, ώστε να καθορίσουν ποια αρχεία περιέχουν ιδιαίτερες λέξεις, φράσεις, ή συνδυασμούς αυτού. Μπορούν επίσης να τα ταιριάξουν με άλλα καθορισμένα ως προς τον χρήστη κριτήρια, όπως εάν μια ιδιαίτερη σελίδα περιέχει αρχεία ήχου ή εικόνας.

Το μέρος της μηχανής αναζήτησης που υπολογίζει τη σχετικότητα των αρχείων μπορεί να ενσωματωθεί στον αλγόριθμο ανάκτησης ή μπορεί να είναι μια χωριστή διαδικασία. Ακόμα και όταν είναι μια χωριστή διαδικασία, η διαφορετικότητα μπορεί να μην είναι προφανής στο χρήστη, και συνήθως δεν πρέπει να είναι.

- **Η Γραφική Διεπαφή HTML**

Το τι βλέπουν οι χρήστες όποτε συνδέονται με μια μηχανή αναζήτησης είναι το *HTML-based interface*. Αυτή η διεπαφή συγκεντρώνει τα στοιχεία της ερώτησης από τον χρήστη, και στέλνει τα στοιχεία αυτά στη μηχανή αναζήτησης για να γίνει η ανάκτηση των σελίδων. Η πιο προφανής βέβαια λειτουργία του είναι να παρέχει στο χρήστη έναν τρόπο για να υποβάλλει την ερώτησή του. Ωστόσο το *interface* παρέχει και άλλες λειτουργίες στο χρήστη, όπως της παροχής ενός διαστήματος για

διαφημίσεις, παρέχει την πρόσβαση στα διάφορα χαρακτηριστικά γνωρίσματα, και της παροχής των συνδέσεων στις σελίδες βοήθειας και άλλες πληροφορίες για την υπηρεσία γενικά.

Ο τρόπος που παρουσιάζονται τα αποτελέσματα στο χρήστη τείνει να τυποποιηθεί, αφού οι περισσότερες μηχανές αναζήτησης δίνουν πλέον μαζί με την παραπομπή στη συγκεκριμένη πληροφορία μια μικρή περίληψη καθώς και ένα ποσοστό σχετικότητας σε σχέση με το ζητούμενο όρο, όπως αυτός τέθηκε από το χρήστη.

Ο ***spider*** επιστρέφει σε *sites* τακτικά (π.χ. κάθε εβδομάδα) προκειμένου να ελέγξει για τυχόν αλλαγές και να ενημερώσει τη βάση, εξασφαλίζοντας ότι η κάλυψη του δικτύου είναι ενημερωμένη και εκτεταμένη. Αυτό έχει ως συνέπεια ένα τεράστιο αριθμό αποτελεσμάτων σχεδόν για οποιαδήποτε αναζήτηση.

Εξάλλου η αυτόματη δημιουργία της βάσης δεδομένων της μηχανής αναζήτησης σημαίνει ότι δεν υπάρχει διαχωρισμός όσον αφορά την ποιότητα της πληροφορίας που ανακτάται, κάτι που είναι απαραίτητο, δεδομένου ότι ο καθένας μπορεί να δημοσιεύσει πληροφορίες μέσω του διαδικτύου. Γενικά, η έλλειψη ελέγχου ποιότητας των πόρων του διαδικτύου σημαίνει ότι οι τεράστιες ποσότητες της ανακτώμενης πληροφορίας μπορεί να κυμαίνονται από υψηλής ποιότητας και σχετικό με την αναζήτηση υλικό έως εξαιρετικά αμφιβόλου αξίας πληροφορία.

Αν και οι μηχανές αναζήτησης στοχεύουν στην εκτέλεση της ίδιας λειτουργίας, η κάθε μία την προσεγγίζει με διαφορετικό τρόπο, οδηγώντας μερικές φορές σε εντυπωσιακά διαφορετικά αποτελέσματα. Παράγοντες που επηρεάζουν τα αποτελέσματα περιλαμβάνουν το μέγεθος της βάσης δεδομένων, τη συχνότητα της ενημέρωσης και τις δυνατότητες αναζήτησης. Επίσης, οι μηχανές αναζήτησης διαφέρουν ως προς την ταχύτητα τους, τη σχεδίαση του περιβάλλοντος της αναζήτησης, τον τρόπο εμφάνισης των αποτελεσμάτων και την ποσότητα της βοήθειας που παρέχουν.

1.3.4. Αξιολόγηση των Μηχανών Αναζήτησης

Η αυξανόμενη ποσότητα της διαθέσιμης πληροφορίας στο διαδίκτυο δημιουργεί συνεχώς νέα ενδιαφέροντα προβλήματα στον τομέα της ανάκτησης πληροφοριών ***(IR)***. Εξαιτίας του τεράστιου αριθμού των σελίδων και των συνδέσμων, η περιήγηση δεν

μπορεί να θεωρείται μία επαρκή διαδικασία αναζήτησης, ακόμα και με την εμφάνιση των θεματικών καταλόγων (π.χ. Yahoo!) (Alschuler, 1989). Συνεπώς, απαιτούνται αποτελεσματικοί *query-based* μηχανισμοί για την πρόσβαση της πληροφορίας, ειδικά από το 85% των χρηστών του δικτύου, για τους οποίους οι μηχανές αναζήτησης είναι ένα πρωταρχικό εργαλείο (Lawrence & Giles, 1999, Schwartz, 1998). Δηλαδή, οι μηχανές αναζήτησης είναι απαραίτητες για την εύρεση πληροφορίας στο *World Wide Web*.

Αρκετοί διαφορετικοί τρόποι έχουν προταθεί για την ποσοτική μέτρηση της απόδοσης των κλασσικών συστημάτων ανάκτησης πληροφοριών (π.χ. [Losee 1998], [Manning, Schutze 1999]), το μεγαλύτερο μέρος των οποίων τείνει να αξιολογήσει τις μηχανές αναζήτησης του ιστού. Εντούτοις, οι χρήστες ιστού μπορεί να έχουν την τάση να προτιμούν μερικούς τρόπους μέτρησης της απόδοσης περισσότερο από ότι οι παραδοσιακοί χρήστες των συστημάτων ανάκτησης πληροφοριών. Για παράδειγμα, οι *χρόνοι απόκρισης* ***(response times)*** εμφανίζονται να είναι στην κορυφή της λίστας των σημαντικότερων θεμάτων για τους χρήστες ιστού, καθώς επίσης και ο αριθμός των πιο πολύτιμων-σχετικών σελίδων που εμφανίζονται στην πρώτη σελίδα των ανακτημένων αποτελεσμάτων (δηλ., τα κορυφαία 8, 10, ή 12 αποτελέσματα), έτσι ώστε το κουμπί που μας οδηγεί στην *επόμενη σελίδα* ***(next page)*** δεν θα είναι απαραίτητο να χρησιμοποιηθεί για την ανεύρεση των καλύτερων αποτελεσμάτων.

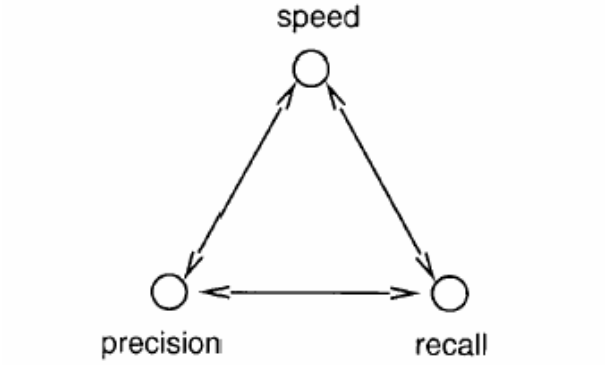
Μερικοί παραδοσιακοί τρόποι μέτρησης της απόδοσης των συστημάτων ανάκτησης πληροφοριών χρησιμοποιούνται με κάποια τροποποιημένη μορφή από τους χρήστες του ιστού. Παραδείγματος χάριν, ένα βασικό πρότυπο των παραδοσιακών συστημάτων ανάκτησης πληροφοριών αναγνωρίζει τη σχέση μεταξύ τριών παραγόντων: της *ταχύτητας* της ανάκτησης των πληροφοριών ***(speed)***, της *ακρίβειας* ***(precision)***, και της *ανάκλησης* ***(recall)***, όπως φαίνεται στο παρακάτω *σχήμα1*. Αυτή η σχέση, μεταξύ των παραγόντων γίνεται όλο και δυσκολότερη καθώς ο αριθμός των εγγράφων και οι χρήστες μιας βάσης δεδομένων αυξάνονται. Στο χώρο της ανάκτησης πληροφοριών, *η ακρίβεια* ορίζεται ως το πηλίκο των σχετικών εγγράφων ως προς το σύνολο των εγγράφων που έχουν ανακτηθεί:

Ακρίβεια (precision): Αριθμός Σχετικών Εγγράφων
Αριθμός Ανακτημένων Εγγράφων

και *ανάκληση* ορίζεται ως το ποσοστό των σχετικών εγγράφων που κατάφεραν να ανακτηθούν από το σύστημα:

Ανάκληση (recall): Αριθμός Σχετικών Ανακτημένων Εγγράφων
Συνολικός Αριθμός Υπάρχοντων Σχετικών Εγγράφων

Οι περισσότεροι χρήστες ιστού που χρησιμοποιούν τις μηχανές αναζήτησης δεν ενδιαφέρονται τόσο πολύ για τη μέτρηση της ακρίβειας με τον παραπάνω τρόπο, αλλά ως ακρίβεια των αποτελεσμάτων που εμφανίζονται στην πρώτη σελίδα του καταλόγου των ανακτημένων εγγράφων. Δεδομένου ότι πραγματικά υπάρχει μικρή πιθανότητα μέτρησης του ποσοστού ανάκλησης για κάθε ερώτηση στη μηχανή αναζήτησης του ιστού -και σε πολλές περιπτώσεις μπορεί να υπάρχουν πάρα πολλές σχετικές σελίδες- ο χρήστης του ιστού θα ενδιαφερόταν για την ανάκτηση και το να είναι σε θέση να προσδιορίσει μόνο τις ιδιαίτερα πολύτιμες σελίδες. Ο Kleinberg (1998) αναγνωρίζει τη σημασία της εύρεσης των πληροφοριών ή των *αρχικών σελίδων* ***(authority pages)***. Ένας χρήστης του ιστού μπορεί να αντικαταστήσει την ανάκληση με μια τροποποιημένη έκδοση στην οποία η ανάκληση σχετίζεται με τις δέκα ή είκοσι πρώτες ταξινομημένες σελίδες (παρά το σύνολο των σχετικών σελίδων).



Σχήμα 1.1: Βασικό πρότυπο αξιολόγησης των παραδοσιακών συστημάτων ανάκτησης πληροφοριών (Kobayashi M. And Takeda K., 2000)

Ο Hearst (1999) σημειώνει ότι το *user-interface*, δηλ., η ποιότητα της αλληλεπίδρασης ανθρώπου-υπολογιστή, πρέπει να ληφθεί υπόψη κατά την αξιολόγηση ενός συστήματος ανάκτησης πληροφοριών. Ο Nielsen (1993) υποστηρίζει τη χρήση των ποιοτικών (παρά των ποσοτικών) μέτρων για την αξιολόγηση των συστημάτων ανάκτησης πληροφοριών. Συγκεκριμένα, προτείνεται η ικανοποίηση του χρήστη με το *σύστημα διεπαφής* ***(system interface)*** καθώς επίσης και η ικανοποίησή του με τα ανακτημένα αποτελέσματα ως σύνολο (παρά ως στατιστικά μέτρα). Ο Westera (1996) προτείνει κάποια σημεία για την αξιολόγηση των μηχανών αναζήτησης, όπως: ενιαία

αναζήτηση των λέξεων κλειδιών, ικανότητα πολλαπλών αναζητήσεων, αναζήτηση φράσεων, αναζήτηση με χρήση των τελεστών Boolean και συνδυασμός των τελεστών Boolean.

Η θεμελιώδης οργανωτική μονάδα του διαδικτύου στην έννοια του ήταν ο *σύνδεσμος υπερκειμένου* ***(hypertext link)*** (Berners-Lee, Cailliau, Luotonen, Nielsen & Secret, 1994). Ωστόσο, η εύρεση συγκεκριμένων εγγράφων μέσω μίας ακολουθίας συνδέσμων υπερκειμένου δεν είναι ιδιαίτερα αποδοτική και εξαρτάται από το αν έχει συμπεριληφθεί μία κατάλληλη ομάδα συνδέσμων από τους διαφόρους συντάκτες των ιστοσελίδων. Η ανάγκη για συγκεκριμένα έγγραφα, καθώς και η περιήγηση γύρω από γενικά θέματα, δημιούργησε έναν αριθμό από μηχανές αναζήτησης και την επακόλουθη αξιολόγηση της λειτουργίας τους.

Οι μηχανισμοί ανάκτησης που έχουν ήδη προταθεί (Leighton & Srivastava, 1999, Gordon & Pathak, 1999) βασίζονται σε συμβατικά μοντέλα *IR*. (Salton, 1989), τα οποία περιλαμβάνουν συγκεντρωτικά ευρετήρια εγγράφων. Αυτές οι μηχανές αναζήτησης δεν μπορούν να κάνουν αποτελεσματική χρήση όλων των διαθέσιμων πληροφοριών (Lawrence & Giles, 1999) και οι περισσότερες από αυτές δεν αναγνωρίζουν τους συνδέσμους υπερκειμένου ως ένα μέσο αύξησης της αποτελεσματικότητας της ανάκτησης.

Παλιότερες εργασίες στο χώρο της ανάκτησης πληροφορίας του διαδικτύου φαίνεται ότι αναγνωρίζουν ότι οι δομές υπερκειμένου μπορούν να είναι σημαντικά πολύτιμες στον εντοπισμό της πληροφορίας (Marchiori, 1997, Kleinberg, 1998, Brin & Page, 1998, Bharat & Henzinger, 1998).

1.3.4.1. Είδη Αξιολογήσεων των Μηχανών Αναζήτησης

Οι αξιολογήσεις των μηχανών αναζήτησης είναι δύο ειδών: *αποδεικτικές/ θεωρητικές* ***(testimonials)*** και *πειραματικές* ***(shootouts)*** (Gordon & Pathak, 1999). Οι *testimonials* γενικά διεξάγονται από τον τύπο ή από οργανισμούς της βιομηχανίας των υπολογιστών, οι οποίοι κάνουν "test drive" και στη συνέχεια συγκρίνουν τις μηχανές αναζήτησης, με βάση την ταχύτητα, την ευκολία χρήσης, τη σχεδίαση και το περιβάλλον τους ή άλλα χαρακτηριστικά, τα οποία είναι εύκολα αντιληπτά στους χρήστες. Ένα άλλο είδος αποδεικτικής αξιολόγησης γίνεται με την έρευνα περισσότερο τεχνικών χαρακτηριστικών των μηχανών αναζήτησης και τη σύγκριση μεταξύ τους σε αυτή τη βάση. Τέτοιες αξιολογήσεις βασίζονται σε χαρακτηριστικά, όπως η ομάδα των

δυνατοτήτων αναζήτησης που έχουν διάφορες μηχανές (π.χ. proximity operators), η κάλυψη τους, ο ρυθμός με τον οποίο οι πρόσφατα δημιουργημένες σελίδες καταχωρούνται στο ευρετήριο και γίνονται διαθέσιμες κατά την αναζήτηση, κ.λ.π.

Αν και οι αποδεικτικές αξιολογήσεις μπορούν να δώσουν στους χρήστες μερικές χρήσιμες πληροφορίες, βοηθώντας τους να αποφασίσουν όσον αφορά το δίλημμα για το ποια μηχανή αναζήτησης είναι καλό να χρησιμοποιήσουν, προτείνουν μόνο έμμεσα ποιες μηχανές αναζήτησης είναι πιο αποτελεσματικές στην ανάκτηση σχετικών ιστοσελίδων.

Στις *shootouts*, από την άλλη, χρησιμοποιούνται σε πραγματικό χρόνο διάφορες μηχανές αναζήτησης για την ανάκτηση ιστοσελίδων και συγκρίνεται η αποτελεσματικότητα τους. Οι *shootouts* μοιάζουν με τις τυπικές αξιολογήσεις ανάκτησης της πληροφορίας, που λαμβάνουν χώρα σε εργαστηριακές συνθήκες, προκειμένου να συγκριθούν διάφοροι αλγόριθμοι κατάταξης, αν και τα *internet shootouts* συχνά λαμβάνουν υπόψη μόνο τα πρώτα 10 με 20 ανακτώμενα έγγραφα, σε αντίθεση με τις παραδοσιακές μελέτες *IR*, που λαμβάνουν υπόψη πολύ περισσότερα.

1.3.4.2. Χαρακτηριστικά Ακριβούς και Περιεκτικής Αξιολόγησης

Προκειμένου να εισαχθούν σε ένα πλαίσιο μερικές από τις παλιότερα δημοσιευμένες μελέτες σχετικές με τις *search engine shootouts*, είναι χρήσιμο να αναφερθούν εφτά χαρακτηριστικά, τα οποία κάνουν μία τέτοια αξιολόγηση πιο ακριβή και πιο περιεκτική (Gordon & Pathak, 1999). Γίνεται αρχικά η υπόθεση ότι ο στόχος μίας τέτοιας έρευνας θα πρέπει να είναι ο καθορισμός της δύναμης του κάθε εργαλείου σε συνθήκες πραγματικής ανάκτησης.

- Πρώτα, οι αναζητήσεις πρέπει να υποκινούνται από τις πραγματικές ανάγκες των χρηστών. Οι ερευνητές που επινοούν, οι ίδιοι, αναζητήσεις για ένα πείραμα, μπορεί να εισάγουν *προκαταλήψεις* ***(biases)*** (για παράδειγμα, με τη σύνθεση αναζητήσεων που ευνοούν μία συγκεκριμένη μηχανή αναζήτησης). Επίσης, δεν μπορούν ποτέ να προσεγγίσουν την απίστευτη ποικιλία της πληροφορίας που αναζητάει ο κόσμος -και χρειάζεται- κατά την αναζήτηση του στο δίκτυο, ούτε μπορούν να εκφράσουν τις λεπτές διαφορές αυτών των αναζητήσεων.
- Δεύτερον, αν ένα πείραμα αναζητά έγγραφα σχετικά με ένα θέμα αναζήτησης, το οποίο έχει προσδιορίσει κάποιος άλλος, αυτή η ανάγκη πληροφόρησης του ατόμου θα πρέπει να συλληφθεί πλήρως και με όσο το δυνατόν περισσότερη συναφή

έκφραση. Μία λίστα από λέξεις-κλειδιά, ακόμα και με *δομική* ***(structuring)*** γραμματική (όπως Boolean ή proximity operators), μπορεί να δώσει μόνο μία γενική προσέγγιση του είδους της πληροφορίας που αναζητά κάποιος.

- Τρίτον, πρέπει να διεξαχθεί ένας επαρκώς μεγάλος αριθμός αναζητήσεων, προκειμένου να γίνουν σημαντικές αξιολογήσεις της αποτελεσματικότητας των μηχανών αναζήτησης.
- Τέταρτον, το *shootout* θα πρέπει να περιλαμβάνει τις περισσότερες μεγάλες μηχανές αναζήτησης.
- Πέμπτον, η αποτελεσματικότητα διαφορετικών μηχανών αναζήτησης πρέπει να αναλυθεί αφού γίνει εκμετάλλευση των ιδιαίτερων χαρακτηριστικών της κάθε μηχανής. Αυτό σημαίνει ότι δε θα πρέπει απαραίτητα να χρησιμοποιηθεί το ίδιο query (πιθανώς με μικρές συντακτικές ή μορφικές παραλλαγές) σε διαφορετικές μηχανές αναζήτησης, για την εύρεση ιστοσελίδων για την ίδια ανάγκη πληροφόρησης. Σε αντίθετη περίπτωση, τα καλύτερα χαρακτηριστικά μίας δεδομένης μηχανής αναζήτησης μπορεί να μη γίνουν ορατά.
- Έκτον, οι κριτικές σχετικότητας πρέπει να γίνουν από το άτομο που χρειάζεται την πληροφορία. Αλλιώς, αν οι πειραματιστές αποφασίζουν για το ποιες ιστοσελίδες είναι σχετικές και ποιες όχι, θα υπάρχουν πολλές λάθος αξιολογήσεις, εξαιτίας τόσο της έλλειψης οικειότητας του πειραματιστή με το θέμα της αναζήτησης όσο και της αδυναμίας του, να γνωρίζει τις ανάγκες του χρήστη, το υπόβαθρο του, την υποκίνησή του, κ.λ.π., χωρίς να υπάρχει πουθενά κοντά επαρκή λεπτομέρεια προκειμένου να αποφασίσει αν έχουν ικανοποιηθεί οι ανάγκες πληροφόρησης του χρήστη ή όχι.

Εν συντομία, από μία *πραγματική* ***(pragmatic)*** άποψη, το ίδιο έγγραφο μπορεί να σημαίνει διαφορετικά πράγματα για διαφορετικούς ανθρώπους. Αυτό διαστρεβλώνει κάθε προσπάθεια για αμερόληπτη «κριτική σχετικότητας», όσον αφορά τη λήψη απόφασης σχετικά με το αν κάποιος άλλος θα εκτιμούσε ένα συγκεκριμένο έγγραφο ως σχετικό. Οι κριτές της σχετικότητας μπορούν να κάνουν μόνο εννοιολογικές ή ακόμα και συντακτικές αξιολογήσεις των εγγράφων και των ερωτημάτων. Ωστόσο, αυτές οι κρίσεις αποτυγχάνουν να λάβουν υπόψη το συγκεκριμένο χρήστη και επομένως αποτυγχάνουν να προσδιορίσουν το κατά πόσο ο χρήστης *πραγματικά θεωρεί* ένα έγγραφο σχετικό.

- Τέλος, πρόσθετα στα παραπάνω κριτήρια, τα *καλά- διεξαχθέντα* ***(well-conducted)***

πειράματα είναι απαραίτητα για την απόκτηση σημαντικών μέτρων απόδοσης. Αυτό σημαίνει:

1) Την ακολούθηση κατάλληλου πειραματικού σχεδιασμού (για παράδειγμα, ανακατεύοντας τη σειρά με την οποία παρουσιάζονται τα έγγραφα στους αξιολογητές, προκειμένου να υπερνικηθούν οι επιρροές λόγω σειράς)

2) Τη συμμόρφωση με αποδεκτές μετρήσεις της *IR* (όπως καμπύλες ανάκλησης - ακρίβειας) προκειμένου τα αποτελέσματα να αξιολογηθούν σε ένα οικείο περιβάλλον

3) Τη χρησιμοποίηση στατιστικών δοκιμών προκειμένου να μετρηθούν ακριβώς, διαφορές στις αποδόσεις μεταξύ των μηχανών αναζήτησης.

Υπάρχει ένας αριθμός αναφερθέντων *shootouts* στη βιβλιογραφία, αλλά κανείς δεν ικανοποιεί τα επτά χαρακτηριστικά που αναφέρθηκαν παραπάνω. Αυτές οι μελέτες παρουσιάζονται στον παρακάτω πίνακα.

	Genuine search?	Information need stated?[a]	Number of searches	Number of search engines	Queries optimized per search engine?	Relevance judged by actual users?	Appropriate experimental design and evaluation?
Leighton, 1995	no	–	8	4	no	no	no
Leighton and Srivastava, 1997	yes	no	15	5	no	no	yes
Ding and Marchionini, 1996	no	–	5	3	yes	no	yes
Chu and Rosenthal, 1996	yes	no	10	3	yes	no	no
Westera, 1996	no	–	5	8	no	no[b]	no
Lebedev, 1997	no	–	8	11	no	no[b]	no
Overton, 1996	no	–	10	8	no	no	no
Schlichting and Nilsen, 1996	yes	yes	5	4	no	yes	no
Feldman, 1997	no	–	7 discussed	7	no	yes	no
Lake, 1997	no	no	40 *facts*	4	yes	evaluation = facts	no
Tomaiuolo and Packer, 1996	some	no	200	5	yes	no	no
Gordon and Pathak (current study)	yes	yes	33	8	yes	yes	yes

Πίνακας 1.1: Προηγούμενες μελέτες αξιολόγησης μηχανών αναζήτησης (Πηγή: Gordon & Pathak, 1999)

Αυτές οι *δοκιμές* ***(tests)*** ποικίλουν με διάφορους τρόπους. Καταρχήν, διαφέρουν όσον αφορά τις ανάγκες πληροφόρησης που θίγουν. Ο Westera (1996), για παράδειγμα, έθεσε μόνο ερωτήματα που είχαν σχέση με το κρασί, ενώ ο Feldman (1997) δοκίμασε μία ομάδα από διαφορετικά ερωτήματα, σε θέματα όπως είναι τα αυτοκίνητα και η ανάκτηση πληροφορίας (Gordon & Pathak, 1999). Σε πολλές μελέτες, τα ερωτήματα

δημιουργήθηκαν από τους πειραματιστές και σε άλλες, τα ερωτήματα προέκυψαν από εκπαιδευτικούς οδηγούς και βιβλία αναφοράς. Φυσικά, σε τέτοιες περιπτώσεις, είναι αδύνατον να υπάρχει μία πλήρως διατυπωμένη ανάγκη πληροφόρησης. Αντίθετα, πρέπει κανείς απλά να χρησιμοποιήσει ή να προσαρμόσει ελάχιστα ένα δοσμένο ερώτημα.

Σε πολλές μελέτες τέθηκε το ίδιο (ή σχεδόν το ίδιο) ερώτημα σε όλες τις μηχανές αναζήτησης. Αν και αυτό μπορεί να μιμείται τη συμπεριφορά αναζήτησης ενός αρχάριου, δε δοκιμάζει τις πραγματικές δυνατότητες μίας μηχανής αναζήτησης. Σε μερικές περιπτώσεις, οι συγγραφείς των μελετών ανάφεραν ότι είχαν συμβουλευθεί τις *Help* και *FAQ* σελίδες κατά την επινόηση των ερωτημάτων τους. Ωστόσο, σε καμία προηγούμενη μελέτη (Gordon & Pathak, 1999), εκτός από του Lake (1997), δεν αναμενόταν από τους ερευνώντες να εξετάσουν τις μηχανές αναζήτησης, εκμεταλλευόμενοι όλες τις δυνατότητες της κάθε μιας.

Επίσης, οι περισσότερες προηγούμενες μελέτες δεν ενέπλεκαν τους χρήστες όσον αφορά τις κρίσεις σχετικότητας. Στην επόμενη παράγραφο υπάρχει και μελέτη που εμπλέκει τους χρήστες για την εξαγωγή των συμπερασμάτων. Αντίθετα, οι κρίσεις σχετικότητας γινόντουσαν ως συνήθως από τους ίδιους τους πειραματιστές. Αν και οι πειραματιστές μπορεί να προσπαθούν αντικειμενικά να δοκιμάσουν το εύρος διαφορετικών ερωτημάτων, που μπορεί διαφορετικοί χρήστες να θέσουν, είναι σχεδόν αδύνατο για αυτούς να καθορίσουν επακριβώς ποια έγγραφα θα ήταν σχετικά. Οι Tomaiuolo and Packer (1996) δοκίμασαν 200 ερωτήματα. Ωστόσο έκαναν οι ίδιοι τις κρίσεις σχετικότητας, συχνά βασισμένοι μόνο στις μικρές περιληπτικές περιγραφές των ιστοσελίδων που παρέχουν οι μηχανές αναζήτησης. Στη μελέτη του Lake (1996), όπου ο σκοπός ήταν να βρεθούν οι ιστοσελίδες που παρείχαν απαντήσεις σε πραγματικές ερωτήσεις, οι κρίσεις από τους πειραματιστές σχετικά με την ακρίβεια της ανακτώμενης πληροφορίας είναι αποδεκτή. Στις μελέτες *IR*, όμως, δίνεται έμφαση στην ικανότητα του συστήματος ανάκτησης να ταυτοποιήσει διάφορα έγγραφα- όχι γεγονότα. Συνεπώς, αν και η μελέτη του Lake (1997) είναι μία ενδιαφέρουσα δοκιμή των χαρακτηριστικών των μηχανών αναζήτησης, δεν αποδίδει την αποτελεσματικότητα των μηχανών αναζήτησης, από το πρίσμα της *IR*.

Άλλες μελέτες, οι μελέτες διέφεραν σημαντικά όσον αφορά την ποσότητα της πειραματικής ακρίβειας που χρησιμοποίησαν. Οι Leighton and Srivastava (1997) και οι Ding and Marchionini (1996) εκτέλεσαν στατιστικές δοκιμές προκειμένου να συγκρίνουν μηχανές αναζήτησης. Επίσης, οι Leighton and Srivastava (1997) έκαναν

«τυφλές» εκτιμήσεις σχετικότητας, εννοώντας ότι (ενεργώντας ως κριτές σχετικότητας) δεν ήξεραν από ποια μηχανή αναζήτησης ήταν ανακτημένο το έγγραφο που αξιολογούσαν. Άλλες μελέτες απέτυχαν να υιοθετήσουν τον ίδιο βαθμό αυστηρότητας, αν και, πολλές από αυτές ήταν λιγότερο επίσημες μελέτες.

Τέλος, μελέτες σχετικά με τις μηχανές αναζήτησης συνέκριναν χαρακτηριστικά όπως, Boolean vs natural language, αριθμό indexed εγγράφων, κ.λ.π. (Liu, 1996, Notess, 1996, Stobart & Kerridge, 1996, Zorn, Emanoil, Marshall & Panek, 1996, Kenk, 1997). Επιπρόσθετα στην εμπειρική τους μελέτη για την απόδοση της ανάκτησης, οι Ding and Marchionini (1996) έκαναν μία περιγραφική σύγκριση και μία εμπειρική αξιολόγηση της απόδοσης της ανάκτησης (ακρίβεια και χρόνος απόκρισης).

Συνεπώς, αν και περιορισμένες σε εύρος, οι προηγούμενες μελέτες σχετικά με τις μηχανές αναζήτησης έχουν αρκετές διαφορές μεταξύ τους. Επίσης, όταν αξιολογούνται ποσοτικά (π.χ. με Signal Detection Analysis), η απόδοση των μηχανών αναζήτησης έχει γενικά βρεθεί ότι είναι χαμηλή.

Οι μετά-μηχανές αναζήτησης αναπτύχθηκαν για να βελτιώσουν την απόδοση της αναζήτησης. Οι Selberg and Etzioni (1995, αναφέρεται από τους Dreilinger & Howe, 1997), βασισμένοι σε εμπειρικές διαπιστώσεις, δήλωσαν ότι καμία απλή μηχανή αναζήτησης δεν είναι πιθανό να επιστρέψει περισσότερα από 45% των σχετικών αποτελεσμάτων. Έτσι οι μετά-μηχανές αναζήτησης θεωρήθηκαν ως οι λύσεις του προβλήματος της χαμηλής ανάκλησης (ανεπαρκής αριθμός σχετικών σελίδων ή ανακτημένων άρθρων). Οι μετα-μηχανές αναζήτησης μπορούν επίσης να παρέχουν ένα ενιαίο περιβάλλον στο χρήστη, αν και τα *interfaces* των *πρώτου-επιπέδου* ***(first level)*** μηχανών αναζήτησης, στις οποίες θέτουν τα ερωτήματα τους μπορεί να ποικίλουν πολύ.

Αν και έχουν γίνει λίγες άμεσες συγκρίσεις της απόδοσης μεταξύ μετα-μηχανών και μηχανών αναζήτησης, τουλάχιστον από μία τέτοια μελέτη, έχει διαπιστωθεί ότι μια μετα-μηχανή αναζήτησης επέστρεφε το μεγαλύτερο αριθμό συνδέσμων, σχετικών με τα θέματα (Gauch, Wang & Gomez, 1996).

Η εμπειρία σχετικά με την έρευνα στο χώρο της ανάκτησης πληροφοριών *(IR)* έχει δείξει ότι είναι εξαιρετικά δύσκολο να βρεθούν κατάλληλα κριτήρια αξιολόγησης. Για παράδειγμα, η ακρίβεια και η ανάκληση είναι τα πιο διαδεδομένα κριτήρια, τα οποία όμως τείνουν να εκφυλιστούν, ενώ εξάλλου είναι εξαιρετικά ευαίσθητα στον τρόπο με τον οποίο καθορίζεται και αξιολογείται η «σχετικότητα», το οποίο είναι από μόνο του ένα σοβαρό θέμα (Harter, 1996). Το γεγονός αυτό οδήγησε σε προτάσεις για

κριτήρια αποτελεσματικότητας, που συνδυάζουν την ανάκληση και την ακρίβεια με διάφορους τρόπους (π.χ. Meadow, 1992).

1.3.4.3. Μελέτες που Λαμβάνουν Υπόψη την Ανθρώπινη Αλληλεπίδραση

Ένα αναπτυσσόμενο μέρος των μελετών εξελίσσει προσεγγίσεις της αξιολόγησης της ανθρώπινης αλληλεπίδρασης με τις μηχανές αναζήτησης, συμπεριλαμβανομένων της *ευχρηστίας* ***(usability)*** και της αποτελεσματικότητας των εργαλείων αναζήτησης του δικτύου. Η μελέτη της Spink (2002) ερευνά μία user-centered προσέγγιση για την αξιολόγηση της μηχανής αναζήτησης Inquirus του δικτύου - ένα εργαλείο μετά - αναζήτησης, αναπτυγμένο από ερευνητές του ***NEC Research Institute***. Στόχος αυτής της μελέτης ήταν η ανάπτυξη μίας user - centered προσέγγισης για την αξιολόγηση, η οποία θα περιλάμβανε:

1. ***Αποτελεσματικότητα:*** βασισμένη στην επιρροή των αλληλεπιδράσεων των χρηστών από τα προβλήματα πληροφόρησής τους και από το στάδιο αναζήτησης της πληροφορίας

2. ***Ευχρηστία:*** η οποία περιλαμβάνει σχεδίαση οθόνης και δυνατότητες του συστήματος για τους χρηστές. Είκοσι δύο εθελοντές αναζήτησαν στο Inquirus ερωτήματα πάνω σε προσωπικά θέματα πληροφόρησης.

Τα ***δεδομένα ανάλυσης*** περιλάμβαναν:

1. Ερωτηματολόγια για τους χρήστες πριν και μετά από την έρευνα
2. Αρχεία εκτέλεσης της αναζήτησης στην Inquirus.

Διαπιστώσεις - κλειδιά περιλάμβαναν:

1. η Inquirus βαθμολογήθηκε υψηλά από τους χρήστες όσον αφορά διάφορα κριτήρια ευχρηστίας
2. όλοι οι χρήστες βιώσαν κάποιο επίπεδο αλλαγής του προβλήματος πληροφόρησης, της αναζήτησης της πληροφορίας και της προσωπικής γνώσης, λόγω της αλληλεπίδρασης με την Inquirus
3. διαφορετικοί χρήστες βιώσαν διαφορετικά επίπεδα αλλαγής
4. το κριτήριο ακρίβεια δε συσχετίστηκε με άλλα βασισμένα στο χρήστη κριτήρια.

Μερικοί χρήστες βιώσαν μεγάλες αλλαγές σε διάφορες βασισμένες στο χρήστη μεταβλητές, όπως το πρόβλημα της πληροφόρησης ή το στάδιο της αναζήτησης της πληροφορίας, με μία αναζήτηση χαμηλής ακρίβειας και αντίστροφα. Αναγνωρίζεται ακόμα η σημασία της ανάπτυξης user - centered προσεγγίσεων για την αξιολόγηση του web και των IR συστημάτων.

ΚΕΦΑΛΑΙΟ 2

ΜΕΘΟΔΟΙ ΚΑΙ ΤΕΧΝΙΚΕΣ ΑΝΑΖΗΤΗΣΗΣ

Η αναζήτηση, είτε πρόκειται για τον παγκόσμιο ιστό είτε για οτιδήποτε άλλο, μπορεί να έχει δύο κύριες μορφές. Καταρχήν, υπάρχει η *γενική αναζήτηση*, κατά την οποία ο χρήστης δε γνωρίζει με βεβαιότητα το ζητούμενο αντικείμενο-πληροφορία και προχωρά στην αναζήτηση συχνότερα με τη μέθοδο της *περιήγησης* ***(browsing)***, προκειμένου να εντοπίσει σχετικές ή ακόμη και ειδικότερες πληροφορίες και αναφορές για το αντικείμενο της ερευνάς του. Σε αντίθεση με τη γενική αναζήτηση, *η ειδική αναζήτηση* αποτελεί μια πιο περίπλοκη διαδικασία, που προϋποθέτει την προηγούμενη ανάλυση του αντικειμένου αναζήτησης, της μεθόδου και της στρατηγικής αναζήτησης που θα χρησιμοποιηθεί, καθώς και, τέλος, της επιλογής του κατάλληλου εργαλείου που θα χρησιμοποιηθεί για την έρευνα.

Σε ό,τι αφορά τη μέθοδο που θα χρησιμοποιηθεί, τέσσερις είναι οι βασικοί τρόποι με τους οποίους ο χρήστης μπορεί να αναζητήσει την όποια πληροφορία:

- ***Αναζήτηση με περιήγηση (Browsing)***
- ***Αναζήτηση με χρήση θεματικών ευρετηρίων (Subject Directories)***
- ***Αναζήτηση με λέξεις-κλειδί (Keyword Search)***
- ***Αναζήτηση με βάση το προφίλ του χρήστη (User's Profile Searching).***

Τέλος, θα μπορούσαν επίσης να αναφερθούν η *συνδυαστική αναζήτηση* ***(Combined Search Engine/Subject Directory)***, που ουσιαστικά αποτελεί έναν

συνδυασμό της αναζήτησης με χρήση μηχανών αναζήτησης και θεματικών ευρετηρίων, και οι *πολλαπλές μηχανές αναζήτησης* ***(Multi Search Engines)***, οι οποίες συγκεντρώνουν σε ένα κοινό ***Interface*** περισσότερες από μία μηχανές αναζήτησης.

2.1. Αναζήτηση με Περιήγηση (Browsing)

Στην αναζήτηση με περιήγηση, ο χρήστης ξεκινά την ερευνά του «περιδιαβάζοντας» τις σελίδες web προκειμένου να οδηγηθεί σε κάποιο αποτέλεσμα, κάποια σχετική πληροφορία που θα αποτελέσει σημείο εκκίνησης στην περαιτέρω έρευνα. Η μέθοδος αυτή χρησιμοποιείται από τους χρήστες που, είτε δεν έχουν απόλυτα σαφή εικόνα του αντικειμένου αναζήτησης, είτε δεν είναι σε θέση να γνωρίζουν τι να περιμένουν ως αποτέλεσμα της αναζήτησης τους. Λόγω της συνεχόμενης και ταχείας αύξησης του όγκου και των αλλαγών στις σελίδες web του διαδικτύου, η εν λόγω μέθοδος ουσιαστικά έχει μετατραπεί στην αναζήτηση με τη χρήση θεματικών ευρετηρίων.

2.2. Αναζήτηση με Χρήση Θεματικών Ευρετηριών (Subject Index Search ή Directory Search)

Στη μέθοδο αναζήτησης με τη χρήση θεματικών ευρετηρίων ή θεματικών καταλόγων, ο χρήστης ή το εργαλείο αναζήτησης ψάχνει για τις πληροφορίες σε θεματικά οργανωμένες βάσεις δεδομένων. Η ιεραρχική δομή με την οποία είναι οργανωμένες οι συγκεκριμένες βάσεις επιτρέπει την αναζήτηση ξεκινώντας από το γενικότερο όρο ή κατηγορία του θέματος. Προχωρώντας σταδιακά στις ειδικότερες υποκατηγορίες, ο χρήστης περιορίζει και εξειδικεύει την αναζήτηση και κατά συνέπεια την πληροφορία που τελικά επιθυμεί να ανακτήσει. Η θεματική αυτή αναζήτηση αποτελεί τον κατεξοχήν τρόπο αναζήτησης στους θεματικούς καταλόγους βιβλιοθηκών, για το λόγο ότι ο χρήστης, προχωρώντας από το γενικότερο προς το ειδικότερο θέμα, έχει τη δυνατότητα να εντοπίσει παράλληλα με τις πληροφορίες που αναζητά μια σειρά συναφών πληροφοριών, που ίσως αποβούν χρήσιμες στην ερευνά του.

2.3. Αναζήτηση με Λέξεις-Κλειδί (Keyword Search)

Η αναζήτηση με βάση τις *λέξεις-κλειδί* ***(keyword search)*** προϋποθέτει τη χρήση των ειδικών εργαλείων που είναι γνωστά ως μηχανές αναζήτησης. Η συγκεκριμένη

αναζήτηση αποτελεί μια παραλλαγή της θεματικής αναζήτησης, που περιγράφηκε στην προηγούμενη ενότητα, καθώς ο χρήστης ανατρέχει στις βάσεις δεδομένων χρησιμοποιώντας θεματικές περιγραφές και λέξεις-κλειδί (keywords) που προσδιορίζουν καλύτερα το ζητούμενο θέμα. Η αναζήτηση αυτού του τύπου επιστρέφει στο χρήστη μια σειρά αποτελεσμάτων και *αναφορών* ***(Hits)***. Η εν λόγω διαδικασία αποτελεί την ευρύτερα χρησιμοποιούμενη μέθοδο αναζήτησης, καθώς προσφέρει μια μεγάλη γκάμα από εργαλεία και μηχανές αναζήτησης, τη χρήση «φυσικής γλώσσας» για τη σύνταξη των ερωτήσεων όπως επίσης δυνατότητες βελτίωσης των αποτελεσμάτων με βάση την ακρίβεια και τη σχετικότητα που έχουν αυτά με το ζητούμενο θέμα.

Η *Προηγμένη Αναζήτηση* ***(Advanced Search)*** με χρήση λέξεων-κλειδιών επιτρέπει στους χρήστες να εισαγάγουν περισσότερες από μια λέξεις-κλειδιά και να συνδυαστούν μεταξύ τους με τη χρήση των τελεστών Boolean (“AND,” “OR,” “NOT”). Κάποιες άλλες πιο περίπλοκες εκδόσεις επιτρέπουν στους χρήστες να ορίζουν διαφορετικά βάρη σε κάθε μια από τις λέξεις-κλειδιά που εισαγάγουν. Η έρευνα του διαδικτύου με βάση τις λέξεις-κλειδιά παρέχεται από τα ακόλουθα εργαλεία αναζήτησης: *Alta Vista* (κατασκευασμένη από την Digital Equipment Corporation), *Excite, Open Text* και *HotBot* (που χρησιμοποιεί τη Μηχανή αναζήτησης *Inktomi*).

2.4. Συνδυαστική Αναζήτηση (Combined Search Engine/ Directory Search)

Η συγκεκριμένη μέθοδος ουσιαστικά αποτελεί ένα εργαλείο αναζήτησης που χρησιμοποιεί δύο συναφείς τρόπους αναζήτησης σε συνδυασμό. Ακολουθώντας την αναζήτηση με βάση το θεματικό κατάλογο, ο χρήστης σταδιακά οδηγείται από τη γενικότερη προς την ειδικότερη κατηγορία. Σε κάθε κατηγορία παρέχεται η δυνατότητα χρήσης μιας μηχανής αναζήτησης, με την οποία ο χρήστης μπορεί να ανατρέξει, κάνοντας χρήση των λέξεων-κλειδί, στα περιεχόμενα της συγκεκριμένης κατηγορίας ή υποκατηγορίας όπου έχει οδηγηθεί μέχρι εκείνη τη στιγμή, με βάση την ιεραρχική αναζήτηση του θεματικού κατάλογου που χρησιμοποιούσε.

Όπου είναι εφικτή, η χρήση της συγκεκριμένης μεθόδου αποφέρει καλύτερα αποτελέσματα, για το λόγο ότι όσο πιο ειδική είναι η κατηγορία για την οποία υποβάλλει ο χρήστης την ερώτηση τόσο πιο περιορισμένο είναι το εύρος της αναζήτησης, κατά συνέπεια και ο όγκος των αποτελεσμάτων που θα προκύψουν. Αυτόν

τον τρόπο αναζήτησης τον παρέχουν οι παρακάτω μηχανές αναζήτησης: *Lycos, Yahoo!, Infoseek, Magellan,* και *WebCrawler (America OnLine.Inc)*. Όλες αυτές οι υπηρεσίες έχουν σχέση με μια υπηρεσία καταλόγου που δημιουργεί χειρωνακτικά τις κατηγορίες κατά θέμα ή τίτλο και ομαδοποιεί τις αρχικές σελίδες κάτω από την κατάλληλη κατηγορία.

2.5. Αναζήτηση με βάση το Προφίλ του Χρήστη (User's Profile Searching)

User profiling μπορεί να οριστεί ως η προσπάθεια να δημιουργηθεί ένα προφίλ (profile) του χρήστη με βάση τα ενδιαφέροντά του και τις συνήθειές του με σκοπό τη βελτίωση της *αλληλεπίδρασης του ανθρώπου με τον υπολογιστή* ***(human computer interaction)***. Τα συστήματα ***User Modeling*** διαφέρουν με τους τρόπους που αποκτούν, χρησιμοποιούν και παρουσιάζουν το προφίλ του χρήστη. Τα προφίλ μπορούν να αποκτηθούν ή να δημιουργηθούν με ποικίλους τρόπους:

1. Άμεσα από τις συνεντεύξεις των χρηστών
2. Από τη *"μηχανική γνώση* ***(knowledge engineers)"*** χρησιμοποιώντας τα στερεότυπα χρηστών (δηλαδή, τη συλλογή ενδιαφερόντων που μοιράζονται οι χρήστες που ανήκουν σε μια συγκεκριμένη ομάδα.) Για παράδειγμα το στερεότυπο των χρηστών "Επιστήμης Πληροφορικής" θα περιελάμβανε μια υποκατηγορία "προγραμματισμός" στο προφίλ τους.
3. *Τεχνικές εκμάθησης μηχανών* ***(machine learning techniques)***, όπου ο *υπεύθυνος δημιουργίας του προφίλ* ***(modeler)*** προσπαθεί να προσδιορίσει στο χρήστη συγκεκριμένη τυπική συμπεριφορά.
4. Το προφίλ δημιουργείται και μέσω παραδειγμάτων, όπου ο χρήστης παρέχει παραδείγματα της συμπεριφοράς του και το λογισμικό τις καταγράφει.
5. *Προφίλ βασισμένα σε κανόνες* ***(Rule-based profiles)***, όπου ο ίδιος ο χρήστης καθορίζει τους κανόνες που ελέγχουν τη συμπεριφορά του προτύπου υπό ορισμένους όρους.

Οι παραπάνω μέθοδοι έχουν τα πλεονεκτήματα και τα μειονεκτήματά τους, αλλά γενικά οι πιο επιτυχημένες είναι εκείνες που προσπαθούν να αναλύσουν τις πληροφορίες, όχι μόνο σε επίπεδο λέξης-κλειδιών, αλλά μερικές φορές βασίζονται στα συμφραζόμενα και σε σημασιολογικό επίπεδο. Τα προφίλ των χρηστών μπορούν να αντιπροσωπευθούν χρησιμοποιώντας ένα ευρύ φάσμα των τεχνικών, από την απλή

τεχνική των λέξεων-κλειδιών, έως την *τεχνητή νοημοσύνη* ***(artificial intelligence)***.

Είναι χρήσιμο να σημειωθεί ότι ένας τυπικός χρήστης έχει πολλαπλάσια και επικαλυπτόμενα μερικές φορές ενδιαφέροντα. Οι κατηγορίες ενδιαφερόντων στα στερεότυπα προφίλ των χρηστών πρέπει να είναι πολύ καλά διαχωρισμένες και ο χρήστης πρέπει να επιλέξει ο ίδιος τις κατηγορίες προκειμένου να δημιουργήσει το προφίλ του. Οι συνεντεύξεις χρηστών είναι πολύ χρονοβόρες, μερικές φορές οι χρήστες αποτυγχάνουν στον κατάλληλο προσδιορισμό των κατηγοριών ενδιαφερόντων τους και το κόστος συντήρησης των προφίλ είναι πολύ υψηλό. Η χρήση των τεχνικών εκμάθησης μηχανών για τη δημιουργία και τη διατήρηση των προφίλ χρηστών στις εφαρμογές φιλτραρίσματος της πληροφορίας (information filtering) είναι επιτακτικό.

ΚΕΦΑΛΑΙΟ 3
ΤΑ ΕΡΓΑΛΕΙΑ ΑΝΑΖΗΤΗΣΗΣ

Ως *εργαλεία αναζήτησης* ***(search tools)*** θα μπορούσαμε να θεωρήσουμε τα προγράμματα εκείνα που επιτρέπουν την αναζήτηση και την ανάκτηση πληροφοριών. Τα κυριότερα εργαλεία αναζήτησης θα μπορούσαν να χωριστούν σε τρεις βασικές κατηγορίες:

- Τις *Μηχανές Αναζήτησης* ***(Search Engines)*** και τις παραλλαγές τους
- Τους *Θεματικούς Καταλόγους* ***(Subject Directories)***
- *Τις Πύλες και Εικονικές Βιβλιοθήκες* ***(Gateways and Virtual Libraries)***

Και τα τρία αυτά εργαλεία προαπαιτούν την ύπαρξη ενός συστήματος *ευρετηριασμού* ***(indexing)***. Η δημιουργία ενός *ευρετηρίου* ***(index)*** γίνεται είτε από τον άνθρωπο, είτε από τον υπολογιστή. Στην περίπτωση των υπολογιστών, τα προγράμματα αυτά ονομάζονται *«ρομπότ»* ***(Robot)***, *«αράχνες»* ***(spiders)***, *«περιηγητές»* ***(wanderers)*** ή *«σκουλήκια»* ***(worms)***. Ο τρόπος που λειτουργούν είναι απλός: τα μικρά αυτά προγράμματα αναλαμβάνουν να συλλέξουν και στη συνέχεια να ευρετηριάσουν τις πληροφορίες που θα βρίσκουν σε μια σελίδα *web* κάθε φορά που την επισκέπτονται, δημιουργώντας έτσι μια βάση δεδομένων. Στο ευρετήριο που δημιουργείται αποθηκεύονται στοιχεία όπως η ηλεκτρονική διεύθυνση της σελίδας *(URL)*, η επικεφαλίδα ή ο τίτλος της ιστοσελίδας καθώς και μια σειρά άλλων λέξεων που περιέχονται στη σελίδα αυτή. Κάθε τέτοιο πρόγραμμα έχει το δικό του τρόπο λειτουργίας και δεν υπάρχει τυποποίηση στα περιεχόμενα του ευρετηρίου που δημιουργεί.

Κατά συνέπεια, όσο μεγαλύτερος είναι ο όγκος των δεδομένων που αποθηκεύονται στο ευρετήριο αυτό, τόσο μεγαλύτερος θα είναι ο όγκος των δεδομένων που θα ανακληθούν στην περίπτωση μιας αναζήτησης και τόσο περισσότερες οι πιθανότητες οι *ζητούμενοι όροι-λέξεις κλειδιά* ***(keywords)*** να βρίσκονται μέσα στα περιεχόμενα των σχετικών αποτελεσμάτων.

Η ανάγκη ύπαρξης εξελιγμένων εργαλείων αναζήτησης του διαδικτύου, είναι πλέον επιτακτική, καθώς το *Internet* εξακολουθεί να αναπτύσσεται ραγδαία και η ανάκτηση των πληροφοριών γίνεται όλο και πιο πολύπλοκη διαδικασία. Οι μηχανές αναζήτησης του διαδικτύου είναι ουσιαστικά, *Συστήματα Ανάκτησης Πληροφορίας* ***(Information Retrieval Systems, IR)***, που σκοπό έχουν μέσω διαφόρων μεθοδολογιών που υποστηρίζουν, να διευκολύνουν τους χρήστες στις αναζητήσεις τους. Βέβαια για να παρέχονται στον χρήστη μόνο εκείνες οι πληροφορίες που τον ενδιαφέρουν έχουν αναπτυχθεί τα συστήματα του ***Information Filtering***.

Στο παρόν κεφάλαιο, παρουσιάζεται αρχικά η δομή των συστημάτων Information Filtering και των συστημάτων ανάκτησης της πληροφορίας, και στη συνέχεια τα κυριότερα διαθέσιμα εργαλεία αναζήτησης του διαδικτύου, οι *Μηχανές Πολλαπλής Αναζήτησης* ***(Multi or Meta- Search Engines)***, οι *Θεματικοί Κατάλογοι* ***(Subject Directories)***, οι *Υβριδικές Μηχανές Αναζήτησης* ***(Hybrid Search Engines)***, *οι Πύλες και Εικονικές Βιβλιοθήκες* ***(Gateways and Virtual Libraries)***, αλλά δίνοντας μεγαλύτερη έμφαση στις *Μηχανές Αναζήτησης* ***(Search Engines)***. Θα αναπτυχθούν τα βασικά χαρακτηριστικά μιας μηχανής αναζήτησης αλλά κυρίως και το πώς δουλεύει μια μηχανή αναζήτησης. Αυτό σημαίνει ότι θα γίνει αναφορά γενικά για την *αρχιτεκτονική των μηχανών αναζήτησης* ***(Architecture of Search Engines)***, αφού στα επόμενα κεφάλαια θα αναπτυχθεί η αρχιτεκτονική για κάθε μηχανή αναζήτησης ξεχωριστά, για τα *ευρετήρια των μηχανών αναζήτησης* ***(Indexing of Search Engines)***, για την *ταξινόμηση* των σελίδων ***(Ranking)*** καθώς και για *το ειδικό λογισμικό των μηχανών αναζήτησης* ***(Robots, Spiders, Crawlers)***.

3.1. Δομή των Information Filtering Συστημάτων (Information Filtering Systems)

Τα διάφορα ηλεκτρονικά μέσα δημιουργούν και κοινοποιούν καθημερινά ένα τεράστιο όγκο πληροφορίας με αποτέλεσμα η ανάγκη χρήσης κατάλληλων εξειδικευμένων εργαλείων για την αντιμετώπιση του καθημερινού καταιγισμού πληροφοριών να έχει γίνει επιτακτική. Τα εργαλεία αυτά πρέπει να ανταποκρίνονται στην πρόκληση της εποχής για συλλογή, αξιολόγηση, επιλογή και απομόνωση ποιοτικής μόνο πληροφορίας.

Το ***Information Filtering*** *(Φιλτράρισμα Πληροφοριών)* είναι μια από τις μεθόδους που αναπτύσσονται ταχύτατα σήμερα προκειμένου να συμβάλλουν στη διαχείριση του μεγάλου όγκου πληροφοριών. Σκοπός του Information Filtering (IF) είναι να παρέχονται στο χρήστη μόνο εκείνες οι πληροφορίες που τον αφορούν. Έτσι τα τελευταία χρόνια έχουν δημιουργηθεί πολλά συστήματα IF για ποικίλα πεδία εφαρμογής. Μερικά παραδείγματα είναι: φίλτρα για τα email που βασίζονται στο ατομικό προφίλ του κάθε χρήστη, φίλτρα για newsgroups αναφερόμενα σε ομάδες χρηστών, φίλτρα σχεδιασμένα για παιδιά που τους επιτρέπουν την πρόσβαση μόνο σε αρμόζουσες σελίδες, φίλτρα για e-commerce εφαρμογές που αναλαμβάνουν την προώθηση προϊόντων απευθυνόμενα μόνο σε πιθανούς πελάτες, και πολλά άλλα. Τα διάφορα συστήματα καλύπτουν διαφορετικά πεδία, διαφοροποιούνται ως προς τη λειτουργικότητά τους και αναπτύσσονται σε διαφορετικές πλατφόρμες. Όλα όμως έχουν ένα κοινό σκοπό: την αυτόματη προώθηση των πιο αξιόλογων πληροφοριών στο χρήστη με βάση πάντα το προφίλ του. Τα IF συστήματα βρίσκουν εφαρμογή σε αδόμητη ή ημι-δομημένη πληροφορία (π.χ. έγγραφα, email κ.α.) μπορούν να διαχειριστούν μεγάλο πλήθος δεδομένων κειμένου, βασίζονται σε προφίλ χρήστη και έχουν σκοπό να αποτρέψουν την παρουσίαση μη σχετικών με το χρήστη πληροφοριών, τις οποίες αυτόματα απομακρύνουν.

Ο Hanani (Hanani et al., 2000) διακρίνει τέσσερα βασικά υποσυστήματα από τα οποία αποτελείται ένα σύστημα IF (Σχήμα 3.1.):

(a) το *υποσύστημα ανάλυσης δεδομένων* ***(data analyzer)***,

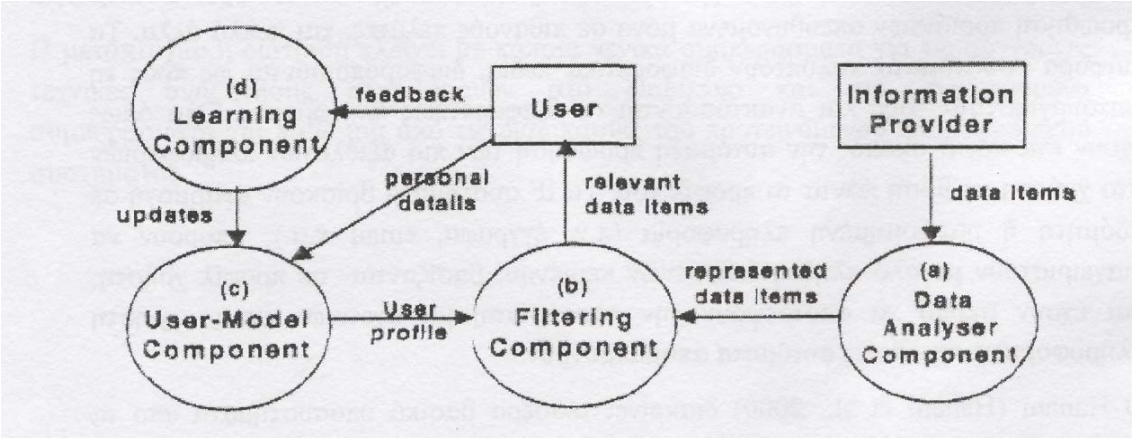
(b) το *υποσύστημα φιλτραρίσματος* ***(filtering)***,

(c) το *υποσύστημα μοντελοποίησης του χρήστη* ***(user model)***, και

(d) το *υποσύστημα μάθησης* **(learning)**.

Αναλυτικά:

- Το υποσύστημα ανάλυσης δεδομένων (a) προμηθεύεται ή συλλέγει δεδομένα (π.χ. έγγραφα, μηνύματα κ.α.) από τους παροχείς πληροφοριών. Τα δεδομένα αναλύονται και αναπαρίστανται με την κατάλληλη μορφή (π.χ. ως διανύσματα επιλεγμένων όρων). Αυτή η απεικόνιση αποτελεί την είσοδο στο υποσύστημα φιλτραρίσματος (b).



Σχήμα 3.1.: Υποσυστήματα ενός Information Filtering System (Πηγή: Hanani et al., 2000)

- Το υποσύστημα μοντελοποίησης χρήστη (c) συλλέγει άμεσα ή έμμεσα στοιχεία για τους χρήστες και τις ανάγκες τους για πληροφόρηση και δομεί μοντέλα χρηστών (π.χ. προφίλ χρήστη). Το μοντέλο χρήστη αποτελεί επίσης είσοδο στο υποσύστημα φιλτραρίσματος (b).
- Το υποσύστημα φιλτραρίσματος (b) συνδυάζει το προφίλ του χρήστη με τα δεδομένα και αποφασίζει αν αυτά είναι σχετικά με το χρήστη. Μερικές φορές η απόφαση είναι δυϊκής μορφής, σχετικά ή μη σχετικά, ενώ κάποιες άλλες είναι πιθανοτικής μορφής, τα δεδομένα δηλαδή κατατάσσονται ανάλογα με την πιθανή σχετικότητα τους ως προς τον χρήστη. Ο χρήστης που λαμβάνει τα πιθανώς σχετικά δεδομένα είναι ο τελικός κριτής. Η αξιολόγηση που κάνει ο χρήστης προσφέρει την δυνατότητα ανατροφοδότησης στο υποσύστημα μάθησης (d).
- Το υποσύστημα μάθησης (d) είναι απαραίτητο για την περαιτέρω βελτίωση της διαδικασίας φιλτραρίσματος. Εξαιτίας της δυσκολίας μοντελοποίησης των χρηστών και των συνεχώς αλλαγών των πληροφοριακών τους αναγκών, η ύπαρξη μιας διαδικασίας μάθησης θεωρείται απαραίτητη για τον εντοπισμό των εναλλαγών στα ενδιαφέροντα των χρηστών και την ανανέωση των μοντελοποιήσεων ενισχύοντας, επιβεβαιώνοντας ή

ακόμα και αναθεωρώντας υπάρχουσα γνώση για τους χρήστες. Η έλλειψη της διαδικασίας μάθησης προκαλεί ανακρίβειες και επηρεάζει σημαντικά τα αποτελέσματα του φιλτραρίσματος.

Πολλοί είναι οι ερευνητές (Belkin και Croft, 1992) που θεωρούν ότι τα συστήματα IF αποτελούν την εξέλιξη των Information Retrieval (IR) συστημάτων και έχουν κληρονομήσει πολλά από τα χαρακτηριστικά τους. Η άποψη αυτή είναι απόλυτα κατανοητή αν σκεφτεί κανείς ότι το υποσύστημα ανάλυσης δεδομένων λειτουργεί αποκλειστικά και μόνο βασιζόμενο σε τεχνικές από το χώρο του ***Information Retrieval*** *(Ανάκτηση Πληροφοριών)*, ενώ από τον ίδιο χώρο προέρχονται και οι τεχνικές που χρησιμοποιούνται για να αποφασιστεί εάν κάποια πληροφορία είναι σχετική με τις ανάγκες του χρήστη.

Υπάρχοντα Information Filtering Συστήματα

Τα τελευταία χρόνια έχουν αναπτυχθεί αρκετά συστήματα Information Filtering (IF). Τα συστήματα αυτά χρησιμοποιούν διάφορες μεθόδους και τεχνικές, ακολουθώντας διαφορετικές οπτικές γωνίες, και παρουσιάζοντας διαφορετικές λειτουργίες. Στη συνέχεια παρουσιάζεται ένα από τα πιο γνωστά IF συστήματα, το ***SIFT***.

- **Stanford Information Filtering Tool (Sift)**

Το Sift-Stanford Information Filtering Tool (Yan.T. W. and Garcia-Molina.H, 1995) είναι ένα εργαλείο φιλτραρίσματος ειδήσεων, οι οποίες προέρχονται από το Internet. Κάθε χρήστης μπορεί να εγγράφει στο server του SIFT για ένα ή παραπάνω τομείς ενδιαφέροντος. Κατά την εγγραφή ο χρήστης δημιουργεί άμεσα το προφίλ του προσδιορίζοντας λέξεις-κλειδιά, τις οποίες το σύστημα πρέπει να θεωρεί ως σχετικές ή να αποφεύγει, καθώς και άλλες παραμέτρους που έχουν να κάνουν με την συχνότητα ανανέωσης του προφίλ και τον όγκο των πληροφοριών που θέλει να λαμβάνει.

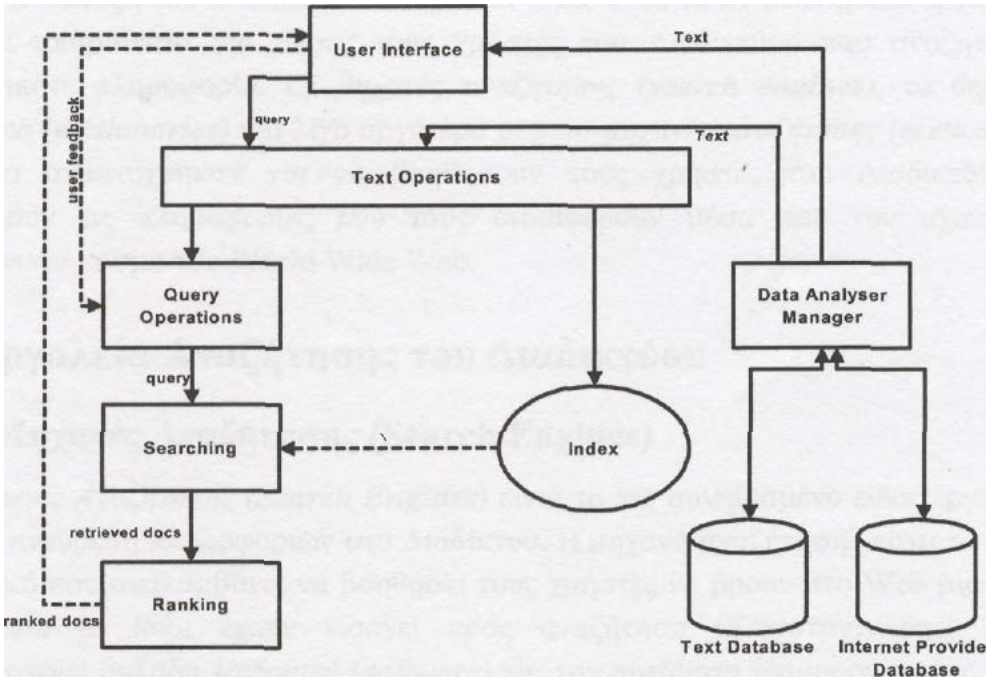
Μετά την δημιουργία του αρχικού προφίλ ο χρήστης μπορεί να το θέσει σε δοκιμαστική λειτουργία για να ελέγξει την αποτελεσματικότητα του χρησιμοποιώντας μια υπάρχουσα συλλογή άρθρων από το server του συστήματος. Ανάλογα με τα αποτελέσματα αυτής της δοκιμής μπορεί να οδηγηθεί στην μεταβολή του προφίλ προσθέτοντας ή αφαιρώντας λέξεις ή αυξομειώνοντας τα αντίστοιχα βάρη τους. Όταν ο χρήστης ικανοποιηθεί από την αποτελεσματικότητα του προφίλ του τότε το ανανεώνει

βάση των αλλαγών που προέκυψαν από την δοκιμαστική λειτουργία.

Για την πραγματοποίηση του φιλτραρίσματος τα προφίλ συγκρίνονται με κάθε άρθρο που φτάνει στο server του συστήματος. Για να βελτιωθεί η αποδοτικότητα του συστήματος τα προφίλ των χρηστών χωρίζονται σε ομάδες έτσι ώστε ένα μέρος της διαδικασίας φιλτραρίσματος να πραγματοποιείται βάση αυτών και όχι με το κάθε ένα ξεχωριστά. Κατά την ανάλυση των άρθρων το SIFT δε διακρίνει τις λέξεις ανάλογα με το αν βρίσκονται στον τίτλο ή στο σώμα του κειμένου. Όπου και να βρίσκεται μια λέξη έχει την ίδια βαρύτητα.

3.2. Δομή των Συστημάτων Ανάκτησης Πληροφορίας (Information Retrieval Systems)

Η Ανάκτηση Πληροφορίας είναι ένα πεδίο έρευνας που ασχολείται με τη δόμηση, ανάλυση, οργάνωση, αναζήτηση και ανάκτηση πληροφοριών. Ένα σύστημα ανάκτησης πληροφορίας λειτουργεί λαμβάνοντας υπόψη από τη μια μεριά μια συγκεντρωμένη συλλογή δεδομένων, και από την άλλη κάποιους χρήστες που θέλουν να αποκτήσουν πρόσβαση στα δεδομένα αυτά. Το *IR* σύστημα δε γνωρίζει τίποτα για το προφίλ του χρήστη του. Ο χρήστης εισέρχεται απλά στο σύστημα και αναζητά κάποιες πληροφορίες θέτοντας στο σύστημα το αντίστοιχο *ερώτημα* ***(query)***. Το *IR* σύστημα καλείται να του παρουσιάσει όλα εκείνα τα έγγραφα, που περιέχουν πληροφορίες σχετικά με το ερώτημα του, ανακτώντας όσο το δυνατό περισσότερα *σχετικά* ***(relevant)*** έγγραφα και προσπαθώντας να μην ανακτήσει *μη-σχετικά* ***(non- relevant)*** έγγραφα. Προκειμένου να αποφανθεί για τη σχετικότητα ενός εγγράφου, το *IR* σύστημα, «μεταφράζει» κάθε έγγραφο συλλέγοντας τις σημαντικότερες εννοιολογικές πληροφορίες του και στη συνέχεια ελέγχει κατά πόσο οι πληροφορίες αυτές καλύπτουν το ερώτημα του χρήστη.



Σχήμα 3.2.: Υποσυστήματα ενός Information Retrieval System (Πηγή: Baeza- Yates R. and Ribeiro- Neto B., 1999)

Το παραπάνω Σχήμα 3.2. παρουσιάζει τη δομή ενός IR συστήματος, όπως αυτή περιγράφεται στο βιβλίο "Modern Information Retrieval" (Baeza-Yates R. and Ribeiro-Neto B., 1999).

Όπως φαίνεται από το σχήμα, απαραίτητο για κάθε *IR* σύστημα είναι να διαθέτει μια συλλογή από έγγραφα. Η συλλογή αυτή μπορεί να είναι είτε *στατική* ***(Text Database)***, είτε να περιλαμβάνει έγγραφα τα οποία συλλέγονται περιοδικά από το web ***(Internet Provider Database)***. Σε κάθε *IR* σύστημα υπάρχει ένας μηχανισμός ***(Data Analyzer Manager)*** που καθορίζει εκείνες τις τεχνικές ***(text operations)*** που θα χρησιμοποιηθούν για την ανάλυση των εγγράφων (π.χ. αυτόματη ανάλυση ολόκληρου του κειμένου του εγγράφου, επεξεργασία μόνο των τίτλων ή μόνο της περίληψης κ.α.).

Με βάση όσα ορίζει ο μηχανισμός αυτός δημιουργείται το *ευρετήριο* ***(index)*** του συστήματος. Πρόκειται για ένα από τα βασικότερα τμήματα ενός *IR* συστήματος, αφού συγκεντρώνει όλη την πληροφορία που είναι διαθέσιμη στο σύστημα διαρθρωμένη με τέτοιο τρόπο ώστε να εξασφαλίζεται η εύκολη αναζήτηση και πρόσβασή της.

Με δεδομένο ότι έχει δημιουργηθεί το ευρετήριο, μπορεί να ξεκινήσει η διαδικασία αναζήτησης. Ο χρήστης θέτει το ερώτημα του, το οποίο υπόκειται σε επεξεργασία από το σύστημα με τον ίδιο τρόπο που αναλύονται και τα έγγραφα. Σε πολλές περιπτώσεις γίνεται ανάλυση του ερωτήματος του χρήστη και με βάση κάποιες περισσότερο *εξειδικευμένες τεχνικές* ***(query operations)*** έως ότου τελικά προκύπτει ένα ερώτημα σε μορφή κατανοητή για το σύστημα. Το ερώτημα αυτό χρησιμοποιείται για να γίνει η *αναζήτηση* ***(Searching)*** των σχετικών εγγράφων από το ευρετήριο.

Αφού ανακτηθούν τα σχετικά έγγραφα και πριν παρουσιαστούν στο χρήστη, τα περισσότερα συστήματα τα κατατάσσουν ***(ranking)*** σύμφωνα με την πιθανή σχετικότητα τους (ως προς το ερώτημα του χρήστη). Ο χρήστης εξετάζει τα έγγραφα που του παρουσιάζει το σύστημα και μπορεί αν θέλει να δηλώσει στο σύστημα κατά πόσο είναι ικανοποιημένος. Μπορεί δηλαδή να υποδείξει στο σύστημα εκείνα τα έγγραφα τα οποία θεωρεί περισσότερο σχετικά με τις απαιτήσεις του. Στο σημείο αυτό ξεκινά μια διαδικασία ανατροφοδότησης του συστήματος από το χρήστη ***(User Feedback)***, όπου μελετώντας τα έγγραφα που υποδεικνύονται, το σύστημα αναδιαμορφώνει την ερώτηση του χρήστη.

Τα *IR* συστήματα αναπτύχθηκαν αρχικά για να καλύψουν τις ανάγκες της βιβλιοθηκονομίας. Τεράστια ευρετήρια δημιουργούνταν είτε από ανθρώπους, είτε αυτόματα για να καταχωρηθεί το υλικό των βιβλιοθηκών έτσι ώστε να είναι εύκολη η διαχείριση του τόσο από τους βιβλιοθηκονόμους, όσο και από τους επισκέπτες των

βιβλιοθηκών.

Με την ανάπτυξη του Internet και του World Wide Web τα *IR* συστήματα άρχισαν να γίνονται απαραίτητα για όλους τους χρήστες του διαδικτύου που αναζητούσαν οποιαδήποτε πληροφορία. Οι *μηχανές αναζήτησης* ***(Search Engines),*** *τα θεματικά ευρετήρια* ***(Dictionaries)*** και λίγο αργότερα οι *μηχανές πολλαπλής αναζήτησης* ***(Meta- Search Engines)*** αναπτύχθηκαν για να βοηθήσουν τους χρήστες του διαδικτύου να εντοπίσουν τις πληροφορίες που τους ενδιαφέρουν μέσα από τον αχανή και ανοργάνωτο κόσμο του World Wide Web.

3.2.1. Διαφορές Ανάμεσα σε Information Filtering και Information Retrieval Systems

Είναι προφανές από τα παραπάνω ότι τόσο τα IF όσο και τα IR συστήματαέχουν ως κοινό σκοπό την επιλογή σχετικών πληροφοριών, διαφέρουν όμως στα ακόλουθα σημεία (Hanani et all., 2000):

- ✓ ***Συχνότητα χρήσης:*** Τα IR συστήματα είναι σχεδιασμένα για περιστασιακή χρήση ικανοποιώντας τις βραχυπρόθεσμες πληροφοριακές ανάγκες ενός χρήστη. Αντιθέτως τα IF συστήματα είναι σχεδιασμένα για μακροχρόνιους χρήστες με μακροχρόνιες πληροφοριακές ανάγκες και επαναλαμβανόμενη χρήση.
- ✓ ***Αναπαράσταση των πληροφοριακών αναγκών του χρήστη:*** *Στα* IR συστήματα, οι ανάγκες του χρήστη εκφράζονται με ερωτήματα (queries). Στα IF συστήματα, οι μακροχρόνιες ανάγκες του χρήστη περιγράφονται από το προφίλ του χρήστη.
- ✓ ***Στόχος:*** Τα IR συστήματα επιλέγουν δεδομένα από βάσεις δεδομένων (π.χ. κείμενα) τα οποία ταιριάζουν με τα ερωτήματα του χρήστη. Τα IF συστήματα αποκρύπτουν εισερχόμενα δεδομένα που δεν είναι σχετικά με το προφίλ του χρήστη (π.χ. e-mail δεν ενδιαφέρουν τον χρήστη) ή συλλέγουν και διανέμουν σχετικά δεδομένα από συγκεκριμένες πηγές βασιζόμενα πάντα στο προφίλ του χρήστη.
- ✓ ***Βάση δεδομένων:*** Τα IR συστήματα διαχειρίζονται σχετικά στατικές βάσεις δεδομένων (π.χ. περιοδικά ανανεώσιμες βάσεις δεδομένων), σε αντίθεση με τα IF συστήματα που διαχειρίζονται δυναμικά δεδομένα (π.χ. e-mails).
- ✓ ***Κατηγορία Χρηστών:*** Τα IR συστήματα υπηρετούν χρήστες που είναι "άγνωστοι" στο σύστημα, οποιοσδήποτε έχει πρόσβαση σε ένα IR σύστημα μπορεί να θέσει ένα ερώτημα (query). Από την άλλη, οι χρήστες ενός IF συστήματος πρέπει να είναι "γνωστοί" του συστήματος, αφού το σύστημα αντιστοιχεί ένα μοντέλο σε κάθε

χρήστη, το οποίο συνήθως έχει την μορφή ενός προφίλ.

- ✓ ***Πλαίσιο ανάπτυξης:*** Τα IF συστήματα κατά την ανάπτυξη τους λαμβάνουν υπόψη τους κοινωνικά θέματα όπως, την μοντελοποίηση χρήστη και την ιδιαίτερη προσωπικότητα του, θέματα που δεν λαμβάνονται υπόψη σε ένα IR σύστημα.

3.3. Οι Μηχανές Αναζήτησης

Η μηχανή αναζήτησης παρέχει στο χρήστη περισσότερες ευκολίες και έλεγχο σε ότι αφορά τον τρόπο με τον οποίο αναζητά τις πληροφορίες, ενώ επιτρέπει συνάμα το καλύτερο *φιλτράρισμα* ***(filtering)*** και τη *βελτιστοποίηση* ***(refining)*** των αποτελεσμάτων που θα προκύψουν. Κατά συνέπεια, οι μηχανές αναζήτησης αποτελούν το βασικό εργαλείο τόσο για τον αρχάριο όσο και για τον πιο εξοικειωμένο χρήστη του διαδικτύου και των εργαλείων που αυτό προσφέρει.

Όπως αναφέρθηκε, οι μηχανές αναζήτησης χρησιμοποιούν το ευρετήριο που έχει δημιουργηθεί από τα προγράμματα συλλογής και ευρετηρίασης ώστε να ανακαλέσει τους όρους που έχουν αναζητηθεί στο *ερώτημα* ***(query)***. Η ευρετηρίαση του συνόλου ή του μεγαλύτερου ποσοστού των λέξεων που θα βρεθούν σε κάθε σελίδα *web* αυξάνει τον όγκο των αποτελεσμάτων και κατά συνέπεια τις πιθανότητες μιας επιτυχημένης αναζήτησης. Φυσικά, εάν αναλογιστούμε τους ρυθμούς αύξησης των σελίδων του διαδικτύου, μπορούμε εύκολα να διαπιστώσουμε, γιατί η συνεχής ενημέρωση των ευρετηρίων αυτών αποτελεί καθοριστικό παράγοντα για μια επιτυχημένη αναζήτηση.

Οι μηχανές αναζήτησης στο *web* συνδυάζουν μια σειρά τεχνικών προκειμένου να βοηθούν το χρήστη τόσο στην ανάκληση των αποτελεσμάτων όσο και στην ακρίβεια των πληροφοριών που θα ανακτηθούν.

3.3.1. Ορισμός και Παραλλαγές

Οι *Μηχανές Αναζήτησης* ***(Search Engines)*** είναι το πιο συνηθισμένο είδος εργαλείου για την ανεύρεση πληροφοριών στο διαδίκτυο. Η μηχανή αναζήτησης είναι το ειδικό λογισμικό που αναλαμβάνει να βοηθήσει τους χρήστες να βρουν στο *Web* μια σειρά όρων που οι ίδιοι έχουν εισάγει προς αναζήτηση (Κωνσταντινίδης, 2000). Χρησιμοποιεί δηλαδή *λογισμικό* ***(Software)*** για την αυτόματη δημιουργία μίας βάσης δεδομένων από *δικτυακούς τόπους και σελίδες* ***(Websites and Pages)*** - το software επισκέπτεται συνεχώς σελίδες για τη δημιουργία της βάσης (Cooke, 1999). Σε γενικές γραμμές, ο ανωτέρω ορισμός ισχύει για τις περισσότερες μηχανές αναζήτησης που υπάρχουν αυτή τη στιγμή στο *web*, έστω κι αν παραλλαγές τους συναντούνται ολοένα συχνότερα.

Είναι σημαντικό να αναφερθεί ότι, όταν κάποιος χρησιμοποιεί μία μηχανή αναζήτησης, η αναζήτηση δεν γίνεται «ζωντανά», εκείνη ακριβώς τη στιγμή (Montebello and Ciappara, 2000). Αντίθετα, ο χρήστης διενεργεί την αναζήτηση σε μία

έτοιμη βάση δεδομένων, η οποία έχει δημιουργηθεί κάποιο χρόνο πριν την αναζήτηση.

Διαπιστώνουμε ότι η βασική αρχή της αναζήτησης σε οργανωμένη συλλογή παραμένει η ίδια είτε πρόκειται για έντυπες, είτε για ηλεκτρονικές πηγές και περιλαμβάνει τα εξής στάδια:

- *Συλλογή Δεδομένων* ***(Data Collection)***
- *Ευρετηρίαση Δεδομένων* ***(Data Indexing)***
- *Ανάκληση Δεδομένων* ***(Data Retrieval)***.

Η συλλογή δεδομένων αποτελεί την πλέον βασική αρχή. Τα δεδομένα συγκεντρώνονται και αποθηκεύονται με τυποποιημένο τρόπο και σε προκαθορισμένη μορφή, ώστε να είναι δυνατές η μετέπειτα αναζήτηση και η ανάκληση τους.

Η ευρετηρίαση των δεδομένων είναι απαραίτητη κυρίως για λόγους ταχύτητας, αφού έτσι αναζητιούνται οι όροι-θέματα που έχουν τεθεί με τυποποιημένο τρόπο.

Τέλος, η ανάκληση των δεδομένων έρχεται να συμπληρώσει τα παραπάνω βήματα, συνδυάζοντας τους όρους που έχουν ζητηθεί με αυτούς που βρίσκονται στην ευρετηριασμένη βάση δεδομένων και παρουσιάζοντας τα δεδομένα που ανακλήθηκαν. Ο τρόπος που παρουσιάζονται τα δεδομένα τείνει να τυποποιηθεί, αφού οι περισσότερες μηχανές αναζήτησης δίνουν πλέον μαζί με την παραπομπή στη συγκεκριμένη πληροφορία μια μικρή *περίληψη* ***(abstract)*** καθώς και ένα *ποσοστό επιτυχίας/ σχετικότητας* ***(success percentage)*** σε σχέση με το ζητούμενο όρο, όπως αυτός τέθηκε από το χρήστη.

Οι μηχανές αναζήτησης αποτελούν το βασικό εργαλείο τόσο για τον αρχάριο όσο και για τον πιο εξοικειωμένο χρήστη του διαδικτύου. Οι καταναλωτές τις χρησιμοποιούν για να εντοπίσουν και να αγοράσουν αγαθά ή για να συλλέξουν πληροφορίες σχετικά με τα θέματα που τους ενδιαφέρουν ή ακόμα και για να συμμετέχουν στην εκλογική διαδικασία. Από την άλλοι οι ερευνητές και οι επιστήμονες τις χρησιμοποιούν για να εντοπίσουν δημοσιευμένα άρθρα σε περιοδικά και συνέδρια, βιβλιογραφία και εν εξελίξει ερευνητικές εργασίες. Πλέον, το 85% των χρηστών του Web κάνουν χρήση των μηχανών αναζήτησης με αποτέλεσμα οι πρώτες δέκα (10) σε προτίμηση ιστοσελίδες του διαδικτύου να καταλαμβάνονται στην πλειοψηφία τους από κάποιες από αυτές. Ωστόσο, οι μηχανές αναζήτησης είναι προβληματικές επειδή δε διαχωρίζουν την ποιότητα του υλικού που βρίσκουν.

Στις γνωστότερες μηχανές αναζήτησης συγκαταλέγονται οι: *AltaVista, Excite, Google, HotBot, Infoseek, Lycos, NorthernLight* κ.α. Στις περισσότερες περιπτώσεις, οι μηχανές αναζήτησης χρησιμοποιούνται αποτελεσματικότερα για τον εντοπισμό μίας

συγκεκριμένης πληροφορίας, όπως ενός γνωστού εγγράφου, μίας εικόνας, ή ενός υπολογιστικού προγράμματος, ενώ δεν είναι τόσο αποτελεσματικές στην αναζήτηση γενικών θεμάτων.

3.3.2. Βασικά Χαρακτηριστικά

Η μηχανή αναζήτησης μπορεί να θεωρηθεί ότι αποτελείται γενικά από πέντε κύρια εσωτερικά στοιχεία (μέρη) (Randolph Hock, 2001), τα οποία έχουν αναπτυχθεί αναλυτικά στο πρώτο κεφάλαιο και ειδικότερα στην παράγραφο 1.3.3.:

- *Το ειδικό λογισμικό* ***(robot, spider, crawler κ.λ.π.)***
- *Η βάση δεδομένων ή αλλιώς ο κατάλογος* ***(database of information)***
- *Το πρόγραμμα ευρετηρίασης και το ευρετήριο* ***(the indexing program and the index)***
- *Η μηχανή ανάκτησης, μηχανή αναζήτησης, το ειδικό πρόγραμμα* ***(retrieval engine)*** και
- *Η γραφική διεπαφή* ***(the graphical HTML (Hyper-text Markup Language) interface)***

Πέρα από τα εσωτερικά στοιχεία που συνθέτουν τη μηχανή αναζήτησης, υπάρχουν μια σειρά άλλων εξωτερικών χαρακτηριστικών που προσδιορίζουν μια μηχανή αναζήτησης. Τα εξωτερικά αυτά χαρακτηριστικά είναι τα εξής:

- ✓ *Εμφάνιση*
- ✓ *Περιεκτικότητα*
- ✓ *Ευκολία χρήσης*
- ✓ *Ακρίβεια και ιεράρχηση των αποτελεσμάτων*
- ✓ *Παρουσίαση των αποτελεσμάτων*
- ✓ *Βελτίωση των αποτελεσμάτων*

ΕΜΦΑΝΙΣΗ

Σε γενικές γραμμές, η εμφάνιση μιας μηχανής αναζήτησης προδιαθέτει ευχάριστα ή δυσάρεστα το χρήστη από την αρχή. Η δομή της μηχανής, τα χρώματα, οι εικόνες -όπου αυτές εμφανίζονται- αποτελούν στοιχεία που προσθέτουν άλλοτε θετικά και άλλοτε αρνητικά στην εικόνα μιας μηχανής αναζήτησης. Σχετικές μελέτες δείχνουν ότι οι χρήστες προτιμούν τις μηχανές στις οποίες -εκτός των άλλων παραμέτρων- έχει δοθεί

ιδιαίτερη σημασία στο εύρος του πεδίου όπου ο χρήστης εισάγει τους όρους προς αναζήτηση, στις επεξηγήσεις που δίνονται τόσο στη σελίδα της αναζήτησης όσο και στο help file που συνοδεύει τη μηχανή καθώς και στα χρώματα και τη γενικότερη διάρθρωση της σελίδας όπου γίνεται η αναζήτηση.

ΠΕΡΙΕΚΤΙΚΟΤΗΤΑ

Ως βασική αρχή θα μπορούσαμε να αναφέρουμε πως το μέγεθος όντως κάνει τη διαφορά. Με τον όρο περιεκτικότητα εννοείται το ποσοστό κάλυψης του web που μπορεί να προσφέρει η κάθε μηχανή αναζήτησης και παράλληλα να είναι και ενημερωμένη πρόσφατα για τυχόν αλλαγές που έχουν συμβεί σε αυτό το ποσοστό. Το στοιχείο αυτό αποτελεί και την ουσιαστική διαφοροποίηση στις συγκρίσεις που γίνονται ανάμεσα στις μηχανές αναζήτησης. Σε γενικές γραμμές, όλα τα άλλα χαρακτηριστικά είναι περίπου ίδια. Οι μεγαλύτερες μηχανές αναζήτησης παρέχουν πληροφορίες σχετικά με το μέγεθος και τη συχνότητα ανανέωσης της βάσης τους, χωρίς όμως οι πληροφορίες αυτές να είναι πάντα έγκαιρες, μια και ο ιλιγγιώδης ρυθμός αύξησης των περιεχομένων του *internet* αποτελεί ανασταλτικό παράγοντα.

Πέρα από αυτό, βασικά χαρακτηριστικά που καθορίζουν την περιεκτικότητα μιας μηχανής αναζήτησης είναι:

- ⇨ Ο τρόπος με τον οποίο καταμετρούνται οι πηγές.
- ⇨ Το εύρος κάλυψης μιας μηχανής και
- ⇨ Ποια πληροφορία ευρετηριάζεται από μια web page

Ορισμένες μηχανές καταμετρούν τις *ηλεκτρονικές διευθύνσεις* ***(URL)*** των οποίων οι σελίδες έχουν ευρετηριαστεί στο σύνολο τους. Κάποιες άλλες καταμετρούν στα περιεχόμενά τους όλες τις συνδεόμενες σελίδες που παρουσιάζονται στις σελίδες τους, έστω και αν τα περιεχόμενα αυτά δε συγκεντρώνονται ούτε και ευρετηριάζονται από τη μηχανή αναζήτησης. Τέλος, υπάρχουν οι μηχανές εκείνες που καταμετρούν στα περιεχόμενά τους μια ηλεκτρονική διεύθυνση (URL) κάθε φορά που θα την επισκεφτούν, άσχετα με το αν είναι ήδη καταγεγραμμένη στο ευρετήριο της μηχανής.

Βασικό στοιχείο αποτελεί επίσης το εύρος της κάλυψης μιας μηχανής, για παράδειγμα το εάν καλύπτονται και πληροφοριακές πηγές εκτός του *web*, όπως *usegroups, Gopher, FTP,* κ.ά.

Τέλος, είναι ιδιαίτερα σημαντικό να γνωρίζουμε ποια πληροφορία ευρετηριάζεται από μια web page. Κάποιες μηχανές -η μειονότητα, στη συγκεκριμένη περίπτωση- ευρετηριάζουν το σύνολο των λέξεων και των περιεχομένων της σελίδας που

επισκέπτονται. Θεωρητικά οι μηχανές με αυτά τα ευρετήρια παρέχουν πρόσβαση στο σύνολο των περιεχομένων όλων των σελίδων που επισκέπτονται, ωστόσο τα ευρετήρια που δημιουργούνται, εξαιτίας του όγκου τους, απαιτούν τεράστιο κόστος προκειμένου να συντηρηθούν. Στον αντίποδα, άλλες μηχανές αναζήτησης ευρετηριάζουν τον τίτλο, την επικεφαλίδα, σημαντικές λέξεις και την ηλεκτρονική διεύθυνση από κάθε σελίδα που επισκέπτονται, με αποτέλεσμα τα ευρετήριο τους να υστερεί σημαντικά σε σχέση με τα ευρετήρια που συγκεντρώνουν το σύνολο των περιεχομένων κάθε σελίδας.

ΕΥΚΟΛΙΑ ΧΡΗΣΗΣ

Ιδιαίτερη βαρύτητα δίνεται τον τελευταίο καιρό στον τρόπο με τον οποίο ο χρήστης καθοδηγείται και τελικά χρησιμοποιεί τη μηχανή αναζήτησης. Στις περισσότερες περιπτώσεις, ο χρήστης θα εισαγάγει στο πεδίο αναζήτησης τον όρο για τον οποίο επιθυμεί να βρει πληροφορίες και, δίνοντας την εντολή search, η μηχανή θα αναλάβει τα υπόλοιπα. Ωστόσο, η ακρίβεια των αποτελεσμάτων και η σχετικότητα σε σχέση με το ζητούμενο όρο δίνεται πάντα -ή σχεδόν πάντα- από την επιλογή *Προχωρημένη Αναζήτηση* ***(Advanced Search)***, που διαθέτουν όλες ανεξαιρέτως οι μηχανές αναζήτησης. Στη σελίδα της προχωρημένης αναζήτησης ο χρήστης θα βρει οδηγίες σχετικά με τη σύνταξη της προχωρημένης αναζήτησης καθώς και επεξηγήσεις που αφορούν τα ειδικά σύμβολα που δύναται να χρησιμοποιήσει προκειμένου να πετύχει τα καλύτερα δυνατά αποτελέσματα.

ΑΚΡΙΒΕΙΑ (PRECISION) ΚΑΙ ΙΕΡΑΡΧΗΣΗ (RANKING) ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Η ιεράρχηση των αποτελεσμάτων αναφέρεται σε ένα χαρακτηριστικό που διαθέτουν οι περισσότερες από τις καθιερωμένες μηχανές και τα εργαλεία αναζήτησης. Σύμφωνα με αυτό, δίνεται στο χρήστη η δυνατότητα να διαμορφώσει τη σελίδα των αποτελεσμάτων, επεμβαίνοντας στον τρόπο με τον οποίο αυτά θα εμφανίζονται. Πιο συγκεκριμένα, του παρέχεται η δυνατότητα να ταξινομήσει τη λίστα των αποτελεσμάτων εμφανίζοντας τις εγγραφές που πληρούν συγκεκριμένες συνθήκες ή είναι οι πιο σχετικές σε σχέση με το αρχικό ερώτημα.

Η δυνατότητα αυτή υπάρχει σε αρκετές από τις γνωστές μηχανές αναζήτησης, ωστόσο ο τρόπος που επιτυγχάνεται σε κάθε μηχανή ή εργαλείο αναζήτησης είναι διαφορετικός. Κάποιες μηχανές καταμετρούν πόσες φορές εμφανίζεται ο αναζητήσιμος όρος μέσα στο έγγραφο που έχουν ανακτήσει, άλλες μηχανές διερευνούν εάν ο όρος βρίσκεται στην επικεφαλίδα της σελίδας, στην περίληψη ή στον τίτλο του εγγράφου-

web page, ενώ άλλες εξετάζουν πόσο κοντά βρίσκονται οι αναζητήσιμοι όροι μεταξύ τούς. Τέλος, οι νεότερες μηχανές και εργαλεία αναζήτησης, κυρίως όσες υποστηρίζουν τη δυνατότητα αναζήτησης με τη χρήση *φυσικής γλώσσας* ***(natural language)***, εξετάζουν τη θέση του αναζητήσιμου όρου μέσα στην πρόταση, προκειμένου να κατανοηθεί το νόημα του. Σε καμία περίπτωση τα αποτελέσματα μιας αναζήτησης δεν είναι απολύτως ταυτόσημα με τους όρους που αρχικά αναζητήθηκαν, εκτός βέβαια από τις περιπτώσεις που ο χρήστης έχει διαμορφώσει με μεγάλη ακρίβεια και λεπτομέρεια το ερώτημά του. Στις περισσότερες περιπτώσεις τα αποτελέσματα επιτρέπουν αρκετές βελτιώσεις, οι δε μηχανές που επιτρέπουν την εφαρμογή των τεχνικών αυτών θεωρούνται οι πιο επιτυχημένες.

Στη σελίδα των αποτελεσμάτων ο χρήστης θα βρει, ιεραρχημένα ή μη, τις καλύτερες -σύμφωνα με τη μηχανή αναζήτησης- *αναφορές* ***(links)*** σε σελίδες που περιέχουν έναν ή περισσότερους από τους όρους που είχε θέσει. Στο σημείο αυτό, ιδιαίτερη σημασία έχει ο τρόπος με τον οποίο ο χρήστης εντοπίζει τις καλύτερες δυνατές αναφορές μέσα από το σύνολο όσων παρουσιάζονται στη σελίδα και τις συγκρίνει με τους αναζητήσιμους όρους, προκειμένου να εξακριβώσει την *ακρίβεια* ***(precision)*** των αποτελεσμάτων. Η έγκαιρη πληροφορία, που υποδηλώνεται από την ένδειξη της ημερομηνίας δίπλα στην *ηλεκτρονική διεύθυνση* ***(URL)*** του αποτελέσματος, αποτελεί σημαντικό κριτήριο για την ποιότητα της μηχανής αναζήτησης.

ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Ο τρόπος με τον οποίο παρουσιάζονται τα αποτελέσματα δείχνει να απασχολεί σημαντικά τους κατασκευαστές των μηχανών αναζήτησης, αφού μια καλαίσθητη και σαφής παρουσίαση προσδίδει σημαντικά πλεονεκτήματα στη μηχανή αναζήτησης. Ειδικά στην περίπτωση που η μηχανή αναζήτησης παρέχει στο χρήστη τη δυνατότητα να *μορφοποιήσει ανάλογα με τις προτιμήσεις και τις ανάγκες του* ***(customization)*** τη σελίδα των αποτελεσμάτων, η μηχανή αναζήτησης «κερδίζει πόντους» απέναντι στον ανταγωνισμό. Επίσης, εάν η μηχανή αναζήτησης διαθέτει τη δυνατότητα *ιεράρχησης* ***(ranking)***, δίνει στο χρήστη τη δυνατότητα να εντοπίσει άμεσα, με τη χρήση δεικτών (80%, 75%), τις καλύτερες αναφορές με την πρώτη ματιά. Τέλος, στα θετικά στοιχεία μιας μηχανής αναζήτησης προσμετράται η δυνατότητα για απάλειψη *των διπλών εγγραφών* ***(dublicates ή dublicate entries)***. Φυσικά, η ακρίβεια των αποτελεσμάτων είναι το σημαντικότερο κριτήριο, ωστόσο μια σειρά πρόσθετων χαρακτηριστικών βελτιώνει σημαντικά τη συνολική εικόνα.

ΒΕΛΤΙΩΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Ένα ιδιαίτερο χαρακτηριστικό που διαθέτουν ορισμένες από τις μηχανές αναζήτησης, και το οποίο θα πρέπει να τονιστεί ιδιαίτερα, είναι η δυνατότητα που παρέχουν στο χρήστη να βελτιώσει τα αποτελέσματα της αναζήτησης εισάγοντας επιπλέον θεματικούς όρους ή βελτιώνοντας τους ήδη καταχωρισμένους. Η δυνατότητα αυτή διευκολύνει ιδιαίτερα το χρήστη, αφού τον απαλλάσσει από την επαναδιαμόρφωση του ερωτήματος και του επιτρέπει να επέμβει στην ήδη διατυπωμένη ερώτηση ή, ακόμη, και στα ίδια τα αποτελέσματα της αναζήτησης.

3.3.3. Πως Δουλεύουν οι Μηχανές Αναζήτησης

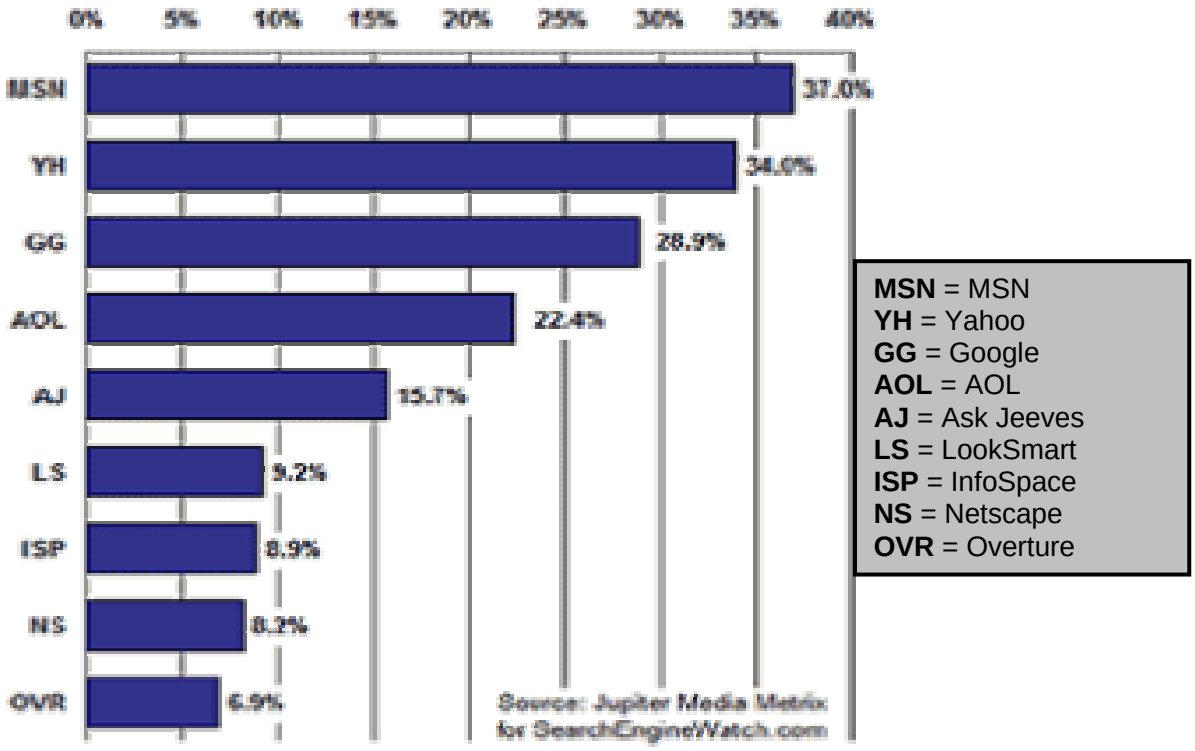
Όλο και περισσότερο σήμερα οι χρήστες του internet, χρησιμοποιούν τις μηχανές αναζήτησης. Ογδόντα ένα (81%) τοις εκατό των βρετανών χρηστών του διαδικτύου χρησιμοποιούν τις μηχανές αναζήτησης για την αναζήτηση ιστοσελίδων (Forrester Research, 2001), μια αύξηση από το εξήντα εφτά (67%) τοις εκατό του 1999. Μια μελέτη που πραγματοποιήθηκε από το Henley Centre (“On line Culture, 2001”), η οποία διερευνά πώς οι άνθρωποι αναζητούν τις ιστοσελίδες, κατέληξε στο συμπέρασμα ότι όλες σχεδόν οι αναζητήσεις ιστοσελίδων, έγιναν με τη χρήση των μηχανών αναζήτησης.

Η ίδια έρευνα επίσης δείχνει ότι οι άνθρωποι που ενδιαφέρονται για την αγορά ενός προϊόντος ή για μια υπηρεσία βρίσκουν την ιστοσελίδα ηλεκτρονικά μέσω των μηχανών αναζήτησης. Το εβδομήντα τοις εκατό (70%) όλων των συναλλαγών του ηλεκτρονικού εμπορίου προέρχονται από μια αναζήτηση (Jupiter MMXI, 2001).

Τα παραπάνω υποδηλώνουν ότι οι μηχανές αναζήτησης είναι ένα ζωτικής σημασίας μέρος οποιασδήποτε διαδικτυακής εμπορικής στρατηγικής.

Πραγματοποιούνται κάθε ημέρα πάνω από 300 εκατομμύρια αναζητήσεις από τις κορυφαίες έξι μηχανές αναζήτησης. Παρακάτω παρουσιάζουμε πόσες περίπου αναζητήσεις πραγματοποιούνται καθημερινά από τις κυριότερες μηχανές αναζήτησης:

- Google: 150 εκατομμύρια
- Inktomi: 100 εκατομμύρια
- Altavista: 50 εκατομμύρια
- Fast: 50 εκατομμύρια
- Overture: 6.5 εκατομμύρια
- AskJeeves: 4 εκατομμύρια



Σχήμα 3.3.: Αναζητήσεις των κυριότερων μηχανών αναζήτησης (Jupiter MMXI March 2002 - UK Home Users)

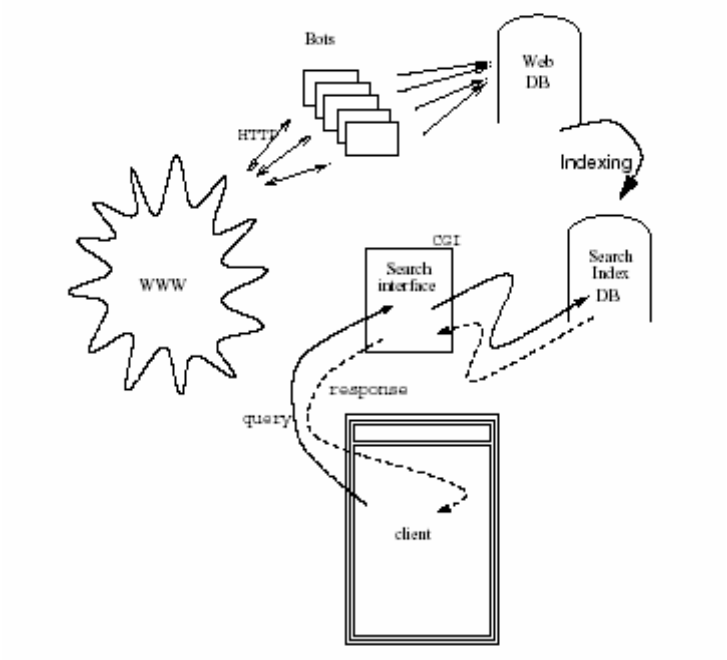
Αυτές οι αναζητήσεις πραγματοποιούνται μέσω ενός μεγάλου αριθμού ιστοσελίδων. Το διάγραμμα που παρουσιάζεται στο παραπάνω Σχήμα 3.3. αναφέρει τις σημαντικότερες (Jupiter MMXI March 2002 - UK Home Users).

3.3.3.1. Αρχιτεκτονική των Μηχανών Αναζήτησης (Architecture of Search Engines)

Οι μηχανές αναζήτησης βασίζονται σε μεγάλες βάσεις δεδομένων και εκτελούν αναζήτηση με βάση τις λέξεις-κλειδιά. Η ανάκτηση των πληροφοριών είναι πολύ μικρή σε σχέση με την ποσότητα και την ποιότητα των πληροφοριών που υπάρχουν στον παγκόσμιο ιστό. Αυτό έχει ως συνέπεια, οι μηχανές αναζήτησης να προσφέρουν μια λίστα από σελίδες, που κατά κύριο λόγο δεν περιλαμβάνουν την απάντηση που θα ικανοποιούσε το χρήστη, έτσι η ακρίβεια είναι πολύ μικρή.

Οι μηχανές αναζήτησης χρησιμοποιούν ειδικό λογισμικό (το οποίο ονομάζεται *robots*, *crawlers* ή *spiders* και για το οποίο θα γίνει εκτενής αναφορά για αυτό σε παρακάτω ενότητα) για την ανίχνευση του ιστού. Κάθε μηχανή αναζήτησης διαφέρει από την άλλη σε αρκετά χαρακτηριστικά (Martin E. Muller):

1. *Στη συμπεριφορά του Crawler* ***(Crawler Behavior)****:*
   - **a)** *Στόχος του Crawler* ***(Crawler Scope)****:* Όλες οι μηχανές αναζήτησης καλύπτουν διαφορετικά μέρη του ιστού και με διαφορετικό βαθμό (βάθος).
   - **b)** *Αναζήτηση Βάθους* ***(Crawler Depth)****:* Όταν ένας *crawler* μπει σε μια ιστοσελίδα, την ανιχνεύει με διαφορετικούς περιορισμούς αναζήτησης βάθους.
   - **c)** *Συχνότητα Αναζήτησης* ***(Crawler Frequency)****:* Για να εγγυηθεί ένα ενημερωμένο ευρετήριο, θα πρέπει να ανανεώνεται συνεχώς και έτσι ο ιστός ανιχνεύεται συνεχώς για τυχόν αλλαγές ή διαγραφές σελίδων.
2. *Στο ευρετήριο* ***(Indexing)****:* Η διαφορά έγκειται στη *μέθοδο ευρετηρίασης* ***(Indexing Method)***, που μπορεί να είναι είτε χειρωνακτική, είτε αυτόματη δημιουργία ευρετηρίου.
3. *Στις δυνατότητες αναζήτησης και διατύπωσης ερωτημάτων* ***(Search and Query Facilities)****:* Αυτή η διαφορά σχετίζεται με τη *γλώσσα του ερωτήματος* ***(Query Language)***, οι οποίες διαφέρουν από απλές *συμβολοσειρές* ***(Strings)***, τους λογικούς τελεστές μέχρι και τη φυσική γλώσσα.

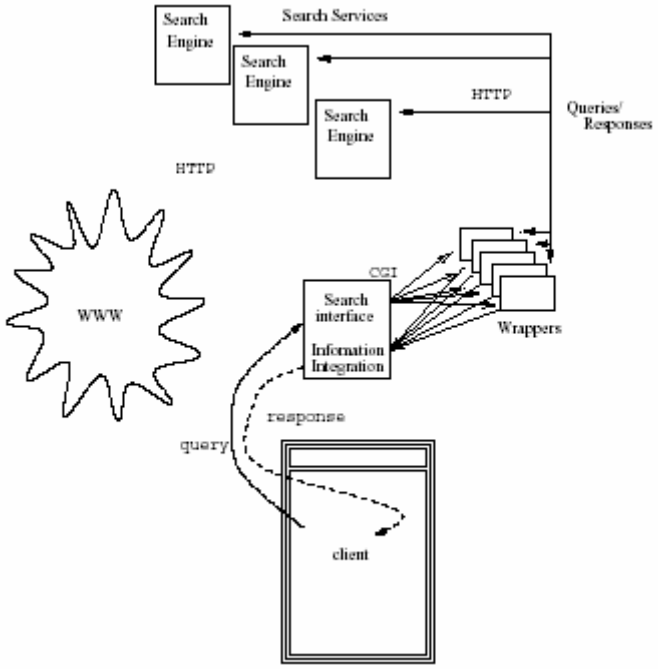


Σχήμα 3.4.: Αρχιτεκτονική μιας γενικής χρήσης μηχανής αναζήτησης (Martin E. Muller)

Το παραπάνω Σχήμα 3.4. δείχνει την αρχιτεκτονική μιας μηχανής αναζήτησης γενικού σκοπού, όπου ο *Crawler* των πληροφοριών ανιχνεύει τον ιστό και αποθηκεύει

τις πληροφορίες σε μια βάση δεδομένων, η οποία είναι προσβάσιμη μέσω ενός ***CGI Script***. Μια πιο προσεκτική ματιά αποκαλύπτει ότι τα αποτελέσματα αναζήτησης των μηχανών αναζήτησης διαφέρουν σημαντικά, λόγω των διαφορετικών εσωτερικών χαρακτηριστικών τους.

Τα μειονεκτήματα των απλών μηχανών αναζήτησης προσπαθούν να τα αντιμετωπίσουν οι μηχανές πολλαπλής αναζήτησης. Σε αυτού του είδους τις μηχανές αναζήτησης το ερώτημα του χρήστη μεταβιβάζεται σε πολλές διαφορετικές μηχανές αναζήτησης, όπως δείχνει και το παρακάτω Σχήμα 3.5., και τα αποτελέσματα εμφανίζονται σε μορφή ενιαίας λίστας. Συλλέγοντας αποτελέσματα από διαφορετικές μηχανές αναζήτησης, η ανάκληση των εγγράφων καθώς και η ακρίβειά τους αυξάνεται.



Σχήμα 3.5.: Αρχιτεκτονική μιας γενικής χρήσης μηχανής πολλαπλής αναζήτησης (Martin E. Muller)

3.3.3.2. Ευρετήρια των Μηχανών Αναζήτησης (Indexing of search Engines)

Αν και χρησιμοποιείται με την ίδια περίπου έννοια και στο χώρο της αναζήτησης και ταξινόμησης πληροφοριών, ο όρος *ευρετήριο* ***(index)*** έχει μια ιδιάζουσα σημασία. Συγκεκριμένα, κάποιοι ειδικοί έχουν δώσει τους ακόλουθους ορισμούς: *«Ένα από τα πλέον σημαντικά εργαλεία για την ανάκτηση πληροφοριών είναι το ευρετήριο - μια συλλογή από όρους με ενδείξεις για τα μέρη που υπάρχουν έγγραφα με σχετικές προς κάθε*

όρο πληροφορίες» (Manber U., 1999). «...δημιουργία ευρετηρίου είναι η διαδικασία κατάλληλης δόμησης των δεδομένων ώστε να εξασφαλίζεται η εύκολη και γρήγορη αναζήτησή τους» (Baeza-Yates, Ribeiro-Neto, 1999). «Οι όροι που καταχωρούνται σε ένα ευρετήριο είναι λέξεις, των οποίων η σημασιολογία βοηθά στην απομνημόνευση του κύριου θέματος ενός εγγράφου» (Baeza-Yates, Ribeiro-Neto, 1999).

Οι στρατηγικές της καταχώρησης σε *ευρετήρια* ***(indexing)*** αποτελούν εξέλιξη της χειρωνακτικής εργασίας των καταχωρήσεων σε καταλόγους βιβλιοθηκών, όπου οι βιβλιοθηκάριοι χειρωνακτικά καθόριζαν έναν αριθμό *λέξεων-κλειδιών* ***(keywords)*** ώστε να ταυτοποιούν κάθε αντικείμενο (βιβλίο, περιοδικό, κ.λ.π.). Η απόδοση κάθε μηχανής αναζήτησης όσον αφορά την ανάκληση και την ακρίβεια, εξαρτάται κυρίως από τη στρατηγική της καταχώρησης σε ευρετήριο. Καίρια θέματα αφορούν το είδος της πληροφορίας που αποσπάται από κάθε έγγραφο και την προσβασιμότητα αυτών των δεδομένων.

Η δημιουργία ευρετηρίου για τα έγγραφα που υπάρχουν στο διαδίκτυο, ώστε να διευκολύνεται η ανάκτησή τους, είναι μια διαδικασία ιδιαίτερα δύσκολη και απαιτητική. Ο τεράστιος αριθμός διαθέσιμων εγγράφων, η ραγδαία αύξηση και ανανέωσή τους έχει ως αποτέλεσμα, τη δημιουργία ενός αξιόπιστου ευρετηρίου, είτε γίνεται από άνθρωπο είτε από υπολογιστική μηχανή, να είναι φαινομενικά αδύνατη.

Μάλιστα αρκετοί ειδικοί στο χώρο της αναζήτησης πληροφοριών θεωρούν ότι ένας σημαντικός αριθμός εγγράφων δεν έχει καταγραφεί στα ευρετήρια καμίας μηχανής αναζήτησης. Οι Lawrence S. και Giles C. (1977) υπολόγισαν ως ελάχιστο αριθμό εγγράφων που θα μπορούσαν να καταχωρηθούν σε ευρετήριο, τον Απρίλιο του 1997, τα 320 εκατομμύρια, και την ίδια στιγμή κάθε μηχανή αναζήτησης μεμονωμένα καταχωρούσε στο ευρετήριο της μόλις το 3% - 34% αυτών. Εκτίμησαν επίσης ότι αν και η *αλληλοεπικάλυψη* ***(overlapping)*** των ευρετηρίων έξι σημαντικών μηχανών αναζήτησης (HotBot, AltaVista, Northern Light, Excite, Infoseek και Lycos) είναι πολύ μικρή, οι καταχωρήσεις τους συνολικά έφταναν μόλις στο 60%.

Οι μελέτες που ακολούθησαν δείχνουν ότι με το πέρασμα του χρόνου τα πράγματα γίνονται ακόμα δυσκολότερα. Αναλογικά το ποσοστό των εγγράφων που είναι σήμερα καταχωρημένες σε σχέση με αυτές που υπάρχουν στο Web είναι πολύ μικρότερο, ενώ καμία μηχανή αναζήτησης δεν έχει καταφέρει να καταχωρήσει στο ευρετήριο της ποσοστό μεγαλύτερο από το 16%.

Δεν είναι όμως μονάχα το πλήθος των εγγράφων που κάνει τη δημιουργία ευρετηρίων ιδιαίτερα δύσκολη. Υπάρχουν κι άλλα εξίσου πολύπλοκα προβλήματα,

όπως για παράδειγμα, το γεγονός ότι τα έγγραφα δεν ακολουθούν κάποια προκαθορισμένη μορφοποίηση. Στο κείμενο τους δηλαδή, συνήθως υπάρχει μεγάλο πλήθος εξειδικευμένων συμβόλων και όρων, ενώ πολλές φορές, εκτός από κείμενο εμπεριέχουν και πολλά στοιχεία πολυμέσων (εικόνες, video κ.λ.π.). Φυσικά, δε θα πρέπει να παραληφθεί και το γεγονός ότι η ανανέωση των ευρετηρίων δε γίνεται με βάση κάποιο συγκεκριμένο χρονοδιάγραμμα.

Ο Henziger (Henziger et al., 1999) πρότεινε μια μέθοδο που θα μπορούσε να λάβει υπόψη τα ευρετήρια των μηχανών αναζήτησης για την αξιολόγηση της ποιότητας των εγγράφων, αλλά και γενικότερα των ιστοσελίδων που καταχωρούν. Το γεγονός, αφενός ότι πλέον είναι δεδομένο ότι δεν είναι εφικτή η καταχώρηση όλων των διαθέσιμων ιστοσελίδων στο ευρετήριο μιας μεμονωμένης μηχανής αναζήτησης και αφετέρου, ότι ούτως ή άλλως η αλληλοεπικάλυψη μεταξύ των διαφόρων ευρετηρίων είναι μικρή, ο Henziger πρότεινε πέρα από το πλήθος των ιστοσελίδων που μπορούν να καταχωρηθούν να λαμβάνεται υπόψη και η ποιότητα τους. Ως κριτήρια για την ποιότητα μιας ιστοσελίδας πρότεινε *τον αριθμό των άλλων ιστοσελίδων που υποδεικνύουν τη συγκεκριμένη σελίδα* ***(indegree)*** αλλά και τον αριθμό των ιστοσελίδων τις οποίες αυτή υποδεικνύει.

Για τη δημιουργία ευρετηρίων ακολουθούνται δύο διαφορετικές τεχνικές:

1. ***Χειρωνακτική (manual)*** δημιουργία ευρετηρίου, που πραγματοποιείται από ανθρώπους
2. ***Αυτόματη (automatic)*** δημιουργία ευρετηρίου, που πραγματοποιείται με τη βοήθεια ειδικών προγραμμάτων ή *ευφυών πρακτόρων* ***(intelligent agents)***.

3.3.3.2.1. Χειρωνακτική Δημιουργία Ευρετηρίου

Η χειρωνακτική δημιουργία ευρετηρίου ακολουθείται σήμερα από ορισμένες μηχανές αναζήτησης όπως: Galaxy, GNN: Whole Internet Catalog, Infomine, LookSmart, Web Developer's Virtual Library, World – Wide Virtual Library Series Subject Catalog και Yahoo!. Τα ευρετήρια που δημιουργούνται χειρωνακτικά θεωρούνται αρκετά πλήρη. Ωστόσο, καθώς ο αριθμός των πληροφοριών που είναι διαθέσιμες στο Web αυξάνει με όλο και μεγαλύτερο ρυθμό είναι πολύ πιθανό μακροπρόθεσμα να σταματήσει η χειρωνακτική δημιουργία ευρετηρίων. Ένα άλλο μειονέκτημα της τεχνικής αυτής είναι η έλλειψη συνοχής ανάμεσα σε δύο διαφορετικούς συντάκτες ευρετηρίων: υπολογίζεται ότι μόλις το 20% των όρων που πρόκειται να καταχωρηθούν σε ένα ευρετήριο αντιμετωπίζονται με τον ίδιο τρόπο από

διαφορετικά άτομα (Korfhage R., 1997).

Αν και δεν είναι τέλεια, τα ευρετήρια που δημιουργούνται χειρωνακτικά, θεωρούνται περισσότερο ακριβή σε σχέση με εκείνα που δημιουργούνται αυτόματα, καθώς οργανώνονται από ειδικούς με βάση διάφορα δημοφιλή θεματικά αντικείμενα και συντάσσονται με τέτοιο τρόπο, ώστε να διευκολύνουν τη διαδικασία αναζήτησης. Εντούτοις, η τεχνολογική πρόοδος αναμένεται να περιορίσει το χάσμα στην ποιότητα των ευρετηρίων που δημιουργούνται αυτόματα από εκείνα που δημιουργούνται χειρωνακτικά. Συνεπώς, εκτιμάται ότι στο μέλλον η χειρωνακτική δημιουργία ευρετηρίων θα εφαρμόζεται μόνο για σχετικά μικρές και στατικές (ή σχεδόν στατικές) ή ιδιαίτερα εξειδικευμένες βάσεις δεδομένων.

3.3.3.2.2. Αυτόματη Δημιουργία Ευρετηρίου

Η πιο διαδεδομένη τεχνική για σύνταξη ευρετηρίων στο Web είναι σήμερα η αυτόματη δημιουργία τους με τη βοήθεια ειδικών προγραμμάτων ή ευφυών πρακτόρων. Γνωστά με την ονομασία ***robots, spiders, wanderers, Web walkers ή και Web agents***, τα προγράμματα αυτά κινούνται συνεχώς στο διαδίκτυο, επισκέπτονται τη μια ιστοσελίδα μετά την άλλη, συλλέγουν πληροφορίες και δημιουργούν τα ευρετήρια όπου καταχωρούν το περιεχόμενο του Web.

Οι περισσότερες μηχανές αναζήτησης ακολουθούν την τεχνική της αυτόματης δημιουργίας ευρετηρίων. Γνωστά παραδείγματα είναι οι: HotBot, AltaVista, Northern Light, Excite, Infoseek, Webcrawler, World Wide Web, Lycos και πολλές άλλες. Ακόμα και στο Yahoo, αν και οι περισσότερες καταχωρήσεις γίνονται χειρωνακτικά, χρησιμοποιείται σε περιορισμένη έκταση ένα ***robot*** που εντοπίζει τις πιθανές καινούριες ανακοινώσεις.

Τα robots κινούνται διαρκώς και απερίσπαστα στο *Web* και έτσι μπορούν να κάνουν σε λίγα λεπτά πράγματα για τα οποία ένας άνθρωπος θα χρειάζονταν μερικές ώρες. Η AltaVista για παράδειγμα υποστηρίζει ότι το *robot* που χρησιμοποιεί, γνωστό ως ***Scooter***, καταχωρεί στο ευρετήριο της καθημερινά περίπου 2,5 εκατομμύρια ιστοσελίδες.

Προκειμένου να γίνει κατανοητό πως λειτουργεί ένα *robot* είναι σημαντικό να καταλάβει κανείς πως λειτουργεί ένας ***browser***. Πρόκειται απλά για ένα πρόγραμμα που ανταποκρίνεται στην είσοδο που του δίνεται από τον χρήστη, στέλνοντας στο διαδίκτυο εντολές *Πρωτοκόλλου Μεταφοράς Υπερκειμένου* ***(Hypertext Transport Protocol)***, προκειμένου να ανακτήσει έγγραφα του διαδικτύου και να τα εμφανίσει στην οθόνη του

υπολογιστή. Ένα τέτοιο έγγραφο δεν είναι παρά ένα HTML αρχείο, που περιέχει κείμενο, πιθανότατα εικόνες ή γραφήματα, ***links*** προς άλλα HTML, αρχεία (παραπομπές) και διάφορες άλλες πληροφορίες.

Τις περισσότερες φορές ο χρήστης συναντά *links* προς άλλα έγγραφα, τα οποία μπορεί αν θέλει να επισκεφτεί. Στην πραγματικότητα κάθε φορά που ο χρήστης ζητά να επισκεφθεί ένα από αυτά τα link, ο browser ανακτά από την σελίδα στην οποία βρίσκεται ήδη ο χρήστης, την *Uniform Resource Locator* ***(URL)*** διεύθυνση του ζητούμενου προορισμού. Στη συνέχεια συνδέεται με τον *server* που φιλοξενεί το ζητούμενο έγγραφο και εκπέμπει μια ***http GET*** εντολή ώστε να ανακτήσει το ζητούμενο HTML αρχείο και το εμφανίζει στην οθόνη. Συνήθως μάλιστα πρώτα εμφανίζεται στην οθόνη το κείμενο. Οι εικόνες εναποθηκεύονταν σε διαφορετικά URL, έτσι ώστε να καλούνται μεμονωμένα. Μέσα στο κείμενο υπάρχουν *links* προς τις εικόνες αυτές έτσι ώστε ο *browser* να ξέρει πού μπορεί να τις βρει και σε ποιο σημείο μέσα στο κείμενο να τις τοποθετήσει.

Το *robot* είναι ένα πρόγραμμα αυτό-καθοδηγούμενο. Αντί να υπάρχει ένας χρήστης που να ακολουθεί *hypertext links*, το *robot* "κατεβάζει" μια σελίδα από το *Web* και αναζητεί σε αυτή *links* προς άλλες σελίδες. Διαλέγει ένα URL και μεταπηδά έτσι σε μια άλλη σελίδα και ξαναρχίζει την όλη διαδικασία από την αρχή. Όταν βρεθεί σε μια σελίδα που δεν περιέχει *links*, γυρίζει πίσω ανεβαίνοντας ένα ή δύο επίπεδα και μεταπηδά σε ένα από τα *links* που παρέλειψε πριν. Από τη στιγμή που ένα *robot* ξεκινά την πορεία του, μπορεί ακολουθώντας ένα απλό επαναλαμβανόμενο αλγόριθμο περιήγησης να καλύψει ένα τεράστιο τμήμα του κυβερνοχώρου και μάλιστα καθώς το *Web* αλλάζει καθημερινά, αλλάζει καθημερινά και η διαδρομή που ακολουθεί το *robot*. Κατά μια έννοια το *robot* κινείται στο *Web* όπως ακριβώς μια αράχνη. Το μόνο που απαιτείται είναι ένα μέρος να ξεκινήσει.

Τα *robots* που ασχολούνται με τη δημιουργία ευρετηρίων χρησιμοποιούν συνήθως κατάλληλους αλγορίθμους, με τη βοήθεια των οποίων αναλύουν το έγγραφο που συναντούν. Η ανάλυση του εγγράφου μπορεί να λαμβάνει υπόψη της ολόκληρο το έγγραφο, μονάχα τον τίτλο του ή μονάχα την περίληψή του. Σε κάθε περίπτωση από την ανάλυση του εγγράφου προκύπτει μια απεικόνιση του εγγράφου η οποία καταχωρείται στο ευρετήριο. Η πιο διαδεδομένη σήμερα τεχνική για την δημιουργία της απεικόνισης ενός εγγράφου είναι η αυτόματη ανάλυση κειμένου που περιγράφεται στην επόμενη ενότητα.

Τα *robots* που δημιουργούν ευρετήρια πολλές φορές συνεργάζονται και με άλλα

robots που υπάρχουν και κινούνται στο διαδίκτυο. Τέτοια *robots* είναι εκείνα που έχουν ως αρμοδιότητα τους να αναγνωρίζουν τα *links* που οδηγούν σε σελίδες που δεν είναι διαθέσιμες (έχουν αποσυρθεί). Άλλα *robots* χρησιμοποιούνται μόνο για συλλογή στατιστικών πληροφοριών που αφορούν τη χρήση του *Web*, όπως για παράδειγμα για την εκτίμηση εκείνων των ιστοσελίδων, οι οποίες είναι οι πλέον δημοφιλείς, μετρώντας τις παραπομπές που υπάρχουν για αυτές στις άλλες ιστοσελίδες ή για την καταγραφή του αριθμού των ιστοσελίδων, ώστε να μπορούν να εκτιμούν το ρυθμό ανάπτυξης του *Web*.

Αν και όπως είναι προφανές από τα παραπάνω, η συνεισφορά των *robots* είναι σημαντικότατη για τη δημιουργία ευρετηρίων, υπάρχουν και ***τρία σημαντικότατα προβλήματα*** που έχουν κάποια σχέση με τα *robots*:

1. Πολλοί φοβούνται ότι είναι ιδιαίτερα διεισδυτικά
2. Υπάρχει περίπτωση να υπερφορτώσουν τους *servers* και να δημιουργήσουν σημαντικό πρόβλημα, καθώς απαιτούν από αυτούς ένα μεγάλο αριθμό εγγράφων σε πολύ μικρό χρονικό διάστημα. Προκαλούν υπερφόρτωση αντίστοιχη με εκείνη που θα προκαλούσαν εκατό ή περισσότεροι χρήστες αν εισέρχονταν ταυτόχρονα στο σύστημα
3. Δεν μπορούν να αναγνωρίσουν αν η ιστοσελίδα που επισκέπτονται είναι σταθερή ή προσωρινή (υπάρχουν για παράδειγμα ιστοσελίδες που ανανεώνονται κάθε μερικές ώρες (π.χ. σελίδες εφημερίδων) ή ακόμα και κάθε 20 λεπτά (π.χ. η ιστοσελίδα του CNN).

❖ ΕΙΔΗ ΑΥΤΟΜΑΤΑ ΔΗΜΙΟΥΡΓΗΜΕΝΩΝ ΕΥΡΕΤΗΡΙΩΝ

Υπάρχουν γενικά δύο είδη αυτόματα δημιουργημένων ευρετηρίων: το *σταθμικό* ***(weighted)*** και *το μη-σταθμικό* ***(unweighted)*** (Kowalski, 1997). Σε ένα μη-σταθμικό ευρετήριο κάθε όρος αποθηκεύεται με μία τιμή, η οποία περιγράφει την τοποθεσία του και λίγες ή καθόλου περαιτέρω πληροφορίες. Αυτά τα ευρετήρια υποστηρίζουν με τον καλύτερο τρόπο Boolean αναζητήσεις, όπου το έγγραφο είναι είτε σχετικό είτε όχι. Καμία ένδειξη σε σχέση με το βαθμό σχετικότητας δεν μπορεί να αποκτηθεί εύκολα από αυτού του είδους τα ευρετήρια.

Λόγω του μεγέθους του διαδικτύου, είναι προφανής η ανάγκη για οργάνωση των αποτελεσμάτων που προκύπτουν από ένα *ερώτημα* ***(query)*** σε μία λίστα κατάταξης, όπου τα έγγραφα που πιθανώς είναι πιο σχετικά θα βρίσκονται στην κορυφή. Σε ένα σταθμικό ευρετήριο, αποδίδεται στους όρους ένα βάρος ανάλογα με τη συχνότητά τους

εντός του εγγράφου. Οι θεωρίες των Luhn, Brookstein, Klein και Raita (Kowalski, 1997) υποστηρίζουν την άποψη, ότι η σημαντικότητα μίας λέξης όσον αφορά τη δύναμή της να αποκαλύψει έννοιες εντός του εγγράφου, είναι άμεσα ανάλογη με τη συχνότητα με την οποία συναντάται εντός του εγγράφου. Οι τιμές των βαρών που αποδίδονται στους όρους του ευρετηρίου συνήθως κανονικοποιούνται σε ένα νούμερο μεταξύ 0 και 1, όπου το 1 δείχνει τη μέγιστη σημαντικότητα. Ο αριθμός συναντήσεων του όρου στη βάση δεδομένων, θεωρούμενης ως μιας ολότητας, χρησιμοποιείται συχνά προκειμένου να αποφευχθεί η απόδοση σημαντικών τιμών βάρους σε συνηθισμένες λέξεις ***('stop words')***. Αυτή η απόδοση βαρών επιτρέπει στο μηχανισμό ανάκτησης να βαθμολογήσει και να κατατάξει τα έγγραφα, σύμφωνα με την σχετικότητά τους με το ερώτημα του χρήστη. Συχνά, οι όροι της αναζήτησης είναι ίδια σταθμισμένοι, προκειμένου να προσδιορίσουν τις πιο σημαντικές λέξεις όσον αφορά τη δυνατότητα τους να ανακτήσουν σχετικά έγγραφα. Αυτό επιτυγχάνεται με την απόδοση βαρών σύμφωνα με τη συχνότητα των λέξεων εντός της βάσης δεδομένων.

Το ***Vector Space Model*** (Kowalski, 1997) είναι μία συνηθισμένη προσέγγιση *IR* του σταθμικού ευρετηρίου και της επακόλουθης ανάκτησης. Τα έγγραφα αναπαριστώνται ως διανύσματα, το κάθε ένα από τα οποία έχει μία θέση διανύσματος για κάθε γνωστό όρο (λέξη) στη βάση δεδομένων. Ο *indexing* μηχανισμός αποδίδει ένα βάρος στη θέση κάθε ευρισκομένου όρου ανάλογα με τη συχνότητα του. Οι όροι που δε βρίσκονται έχουν τιμή 0. Στη συνέχεια τα ερωτήματα μεταφράζονται σε διανύσματα ώστε να μπορεί να αποκτηθεί ένα μέτρο ομοιότητας μεταξύ του διανύσματος των ερωτημάτων και του διανύσματος των εγγράφων. Μία εκδοχή αυτής της προσέγγισης χρησιμοποιείται από τη μηχανή αναζήτησης Excite, ως μέρος της διαδικασίας ***Intelligent Concept Extraction (ICE)***. Στην επόμενη ενότητα γίνεται πιο αναλυτική αναφορά στο μοντέλο.

Μία άλλη συνηθισμένη προσέγγιση IR βασίζεται σε ένα *πιθανοθεωρητικό μοντέλο* ***(probabilistic model),*** το πιο συνηθισμένο από το οποίο είναι το ***Bayesian Model*** (Kowalski, 1997, Oddy et. Al.,1981), όπου η πιθανότητα ανάκτησης ενός εγγράφου που περιέχει μία συγκεκριμένη έννοια υπολογίζεται βάσει του γεγονότος ότι περιέχει συγκεκριμένες λέξεις. Στην επόμενη ενότητα γίνεται πιο αναλυτική αναφορά για το πιθανοθεωρητικό μοντέλο.

Οι πιο προχωρημένες τεχνικές *indexing* προσπαθούν να ορίσουν τις *έννοιες* ***(concepts),*** οι οποίες χρησιμοποιούνται σε ένα έγγραφο, χρησιμοποιώντας στατιστικές μεθόδους, οι οποίες συσχετίζουν την εμφάνιση λέξης και έννοιας. Συχνά διατηρείται

τόσο η συχνότητα εμφάνισης των λέξεων (φράσεων ή εννοιών) όσο και η τοποθεσία τους.

Μερικές μηχανές αναζήτησης αξιώνουν τη χρήση επεξεργασίας της φυσικής γλώσσας. Δηλαδή, προσδιορίζονται οι δομές εντός της γλώσσας -η εννοιολογική πληροφορία συνδυάζεται με την στατιστική πληροφορία για την αναγνώριση φράσεων και πρότυπα λέξεων. Η συχνή συνύπαρξη όρων σε ένα εύρος εγγράφων χρησιμοποιείται επίσης για την αναγνώριση φράσεων και εννοιών.

Είναι σημαντικό το γεγονός ότι, από τη στιγμή που οι όροι του ευρετηρίου έχουν εξακριβωθεί, αποθηκεύονται με τρόπο που να επιτρέπει γρήγορη πρόσβαση από το μηχανισμό ανάκτησης. Μία συνηθισμένη μέθοδος γρήγορης σύνδεσης των όρων της αναζήτησης με τους *αύξοντες αριθμούς* ***(accession numbers)*** των εγγράφων είναι με χρησιμοποίηση ενός *αντεστραμμένου ευρετηρίου αρχείων* ***(inverted file index)***. Δηλαδή, κάθε πιθανός όρος έχει μία καταχώρηση σε ένα αρχείο ευρετηρίου μαζί με μία λίστα συνδεομένων αυξόντων αριθμών των εγγράφων. Αυτοί οι αύξοντες αριθμοί μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτίωση περαιτέρω *μετα-δεδομένων* ***(metadata)*** και συχνά για ένα τοπικό αντίγραφο του *πλήρους κειμένου* ***(full text)*** του εγγράφου.

3.3.3.3. Ανάκτηση Πληροφορίας

Η *βασισμένη στο Δίκτυο* ***(web-based)*** ανάκτηση πληροφορίας διαφέρει σε διάφορα σημεία από την παραδοσιακή ανάκτηση. Για παράδειγμα, μερικές μηχανές αναζήτησης επιτρέπουν στο χρήστη να περιορίσει την ανάκτηση σε συγκεκριμένα πεδία (μόνο .com sites) ή συγκεκριμένα ονόματα πεδίων (όπως ibm.com) ή ακόμα να καθορίσει, για παράδειγμα, ότι όλες οι σελίδες που θέλει να δει θα πρέπει να έχουν ένα *link* στο ibm.com. Επιπλέον, μερικές μηχανές αναζήτησης επιτρέπουν στο χρήστη να καθορίσει, για παράδειγμα, ότι οι σελίδες που τον ενδιαφέρουν θα πρέπει να περιέχουν ένα αρχείο ***plug-in, Java applet*** ή ***real audio*** (Gordon & Pathak, 1999). Επίσης, οι *αλγόριθμοι ταιριάσματος* ***(matching algorithms)*** που χρησιμοποιούν οι μηχανές αναζήτησης συχνά βασίζονται σε αρχές που εφαρμόζονται στην *web-based* αναζήτηση, αλλά όχι στην παραδοσιακή *IR*.

Οι αλγόριθμοι ανάκτησης που χρησιμοποιούνται από το στοιχείο της αναζήτησης των μηχανών αναζήτησης, γενικά ταξινομούνται σε 3 κατηγορίες:

- ✓ *Set Theoretic Models:* που ακολουθούν τη λογική Boolean και κάποιες

προεκτάσεις της

- ✓ *Vector Space Model:* που χρησιμοποιούν κυρίως αλγεβρικές μεθόδους και
- ✓ *Πιθανοθεωρητικό Μοντέλο (Probabilistic Models):* που χρησιμοποιούν μεθόδους από το χώρο της στατιστικής και των πιθανοτήτων.

3.3.3.3.1. Set Theoretic Models

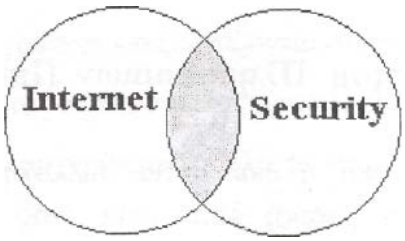
- ❑ **BOOLEAN LOGIC**

Η αναζήτηση με χρήση τελεστών Boolean είναι βασισμένη σε ένα σύστημα λογικής που αναπτύχθηκε στις αρχές του 19ου αιώνα από τον μαθηματικό George Boole. Πρόκειται για ένα ιδιαίτερα ισχυρό εργαλείο αναζήτησης καθώς δίνει σχετικά ακριβή αποτελέσματα και περιορίζει την ανάκτηση μη σχετικών αποτελεσμάτων.

Όταν διεξάγεται μια αναζήτηση με χρήση τελεστών Boolean, αναζητούνται στα ευρετήρια της μηχανής αναζήτησης οι όροι, οι *λέξεις-κλειδιά* ***(keywords)***, που περιγράφουν καλύτερα το θέμα που ενδιαφέρει το χρήστη (όπως αυτό εκφράζεται μέσω της ερώτησής του). Η χρήση των τελεστών Boolean επιτρέπει το συνδυασμό των όρων με τη χρήση τριών *τελεστών* ***(operators)***. Πρόκειται για τους τελεστές ***AND***, ***OR*** και ***NOT***.

- ➢ **Ο ΤΕΛΕΣΤΗΣ AND**

Ο τελεστής AND περιορίζει τα αποτελέσματα μιας αναζήτησης συνδυάζοντας δύο λέξεις-κλειδιά και ανακτά έτσι κάθε έγγραφο που περιέχει απαραίτητα και τους δύο όρους που έχουν προσδιοριστεί. Για παράδειγμα αν κάποιος αναζητά έγγραφα που πραγματεύονται το θέμα της *ασφάλειας (security)* στο *Internet* χρησιμοποιεί το:



Σχήμα 3.6.: Χρήση του τελεστή AND (Πηγή: LSCC, 2001)

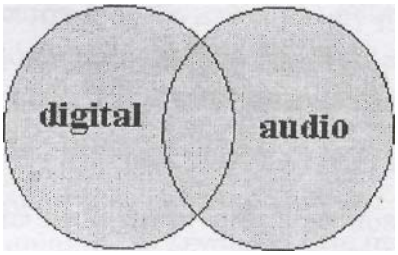
Το παραπάνω διάγραμμα απεικονίζει τη χρήση του τελεστή AND. Στον αριστερό κύκλο περιλαμβάνονται όλα τα έγγραφα που περιέχουν τον όρο *Internet*. Ο δεξιός κύκλος περιλαμβάνει όλα τα έγγραφα που περιέχουν τον όρο security. Το αποτέλεσμα

θα είναι τελικά η μηχανή αναζήτησης να επιστρέψει τα έγγραφα εκείνα στα οποία βρέθηκαν και οι δύο όροι (εντονότερη σκίαση).

Φυσικά μπορεί να περιοριστεί ακόμα περισσότερο μια αναζήτηση αν συνδυαστούν περισσότερες από δύο λέξεις-κλειδιά με τη βοήθεια πολλαπλών τελεστών AND.

- **Ο ΤΕΛΕΣΤΗΣ OR**

Ο τελεστής OR διευρύνει μια αναζήτηση ώστε να ανακτηθούν όλα τα έγγραφα που περιέχουν τουλάχιστον μία από τις λέξεις-κλειδιά που έχουν εισαχθεί. Η αναζήτηση με χρήση του τελεστή OR είναι ιδιαίτερα χρήσιμη όταν υπάρχουν διάφορα συνώνυμα για τον ίδιο όρο ή όταν είναι γενικά δύσκολο να βρεθούν πληροφορίες για κάποιο θέμα και με τον τρόπο αυτό καλύπτεται όσο το δυνατό μεγαλύτερο εύρος.



Σχήμα 3.7.: Χρήση του τελεστή OR (Πηγή: LSCC, 2001)

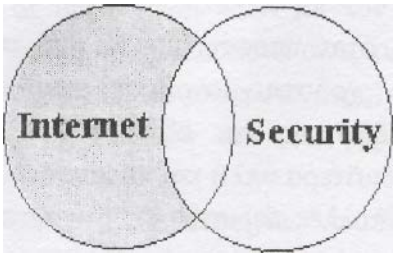
Όπως φαίνεται στο παραπάνω διάγραμμα γίνεται αναζήτηση για όλα τα έγγραφα που περιέχουν τον όρο *digital* (αριστερός κύκλος) και για εκείνα που περιέχουν τον όρο *audio* (δεξιός κύκλος). Τα έγγραφα που θα ανακτηθούν θα είναι όλα εκείνα που περιέχονται στους δύο κύκλους.

Βέβαια μια αναζήτηση τέτοιας μορφής είναι αναπόφευκτο ότι θα επιστρέψει αρκετά μεγάλο όγκο αποτελεσμάτων, μεγαλύτερο ίσως από εκείνον που ο χρήστης θέλει και μπορεί να επεξεργαστεί. Και μάλιστα αν η αναζήτηση αναφέρεται σε περισσότερες από δύο λέξεις-κλειδιά συνδεδεμένες με πολλαπλά OR, ο όγκος των αποτελεσμάτων θα είναι ακόμα μεγαλύτερος.

- **Ο ΤΕΛΕΣΤΗΣ NOT**

Ο τελεστής NOT χρησιμοποιείται για να περιορίσει τα αποτελέσματα σε μια αναζήτηση αποκλείοντας τις ανεπιθύμητες λέξεις-κλειδιά. Για παράδειγμα αν κάποιος αναζητά έγγραφα που έχουν να κάνουν με το *Internet* αλλά δεν τον ενδιαφέρει το θέμα

της ασφάλειας στον κυβερνοχώρο χρησιμοποιεί:



Σχήμα 3.8.: Χρήση του τελεστή NOT (Πηγή: LSCC, 2001)

Στην αναζήτηση που απεικονίζεται στο παραπάνω διάγραμμα ανακτώνται όλα εκείνα τα έγγραφα που περιλαμβάνουν τον όρο *Internet* (αριστερός κύκλος), ενώ αποκλείονται εκείνα που περιέχουν τον όρο *Security* (δεξιός κύκλος). Τα έγγραφα δηλαδή που ανακτώνται είναι εκείνα που απεικονίζονται στο διάγραμμα με εντονότερη σκίαση.

- **ΣΥΝΔΥΑΣΜΟΣ ΤΩΝ ΤΕΛΕΣΤΩΝ BOOLEAN**

Η λογική Boolean επιτρέπει και συνδυασμούς των παραπάνω τελεστών σε μια αναζήτηση. Για μια τέτοια πιο 'σύνθετη' αναζήτηση χρησιμοποιούνται συνήθως παρενθέσεις. Αρχικά γίνεται αναζήτηση για τις λέξεις-κλειδιά και τους τελεστές που εσωκλείονται στις παρενθέσεις και στη συνέχεια για εκείνους που βρίσκονται εκτός παρενθέσεων. Για παράδειγμα στην αναζήτηση για:

(K1 AND K2) OR (K3 NOT K4)

αρχικά αναζητούνται όλα τα έγγραφα που περιέχουν ταυτόχρονα και τους δύο όρους KI και K2. Μετά αναζητούνται τα έγγραφα που περιέχουν τον όρο K3, αλλά στα οποία δεν εμφανίζεται καθόλου ο όρος K4, και τελικά ανακτώνται όλα τα έγγραφα που έχουν ανεβρεθεί στις δύο υπο-αναζητήσεις.

- **ΠΑΡΑΛΛΑΓΕΣ ΤΩΝ ΤΕΛΕΣΤΩΝ BOOLEAN**

- **ΤΕΛΕΣΤΕΣ ΠΡΟΣΘΗΚΗΣ ΚΑΙ ΑΠΟΚΟΠΗΣ, + ΚΑΙ -**

Σε πολλές περιπτώσεις οι μηχανές αναζήτησης που υποστηρίζουν τους λογικούς τελεστές Boolean, προκειμένου να διευκολύνουν τους χρήστες τους, υποκαθιστούν τους τελεστές AND και NOT με τα σύμβολα + και - αντίστοιχα. Η πρακτική αυτή δεν ακολουθείται από όλες τις μηχανές αναζήτησης, καθώς τα σύμβολα αυτά σε ορισμένες

περιπτώσεις επιτελούν συγκεκριμένη λειτουργία κατά τη διάρκεια της αναζήτησης. Έτσι ο τελεστής + θα χρησιμοποιηθεί στις περιπτώσεις που είναι απαραίτητη η παρουσία των όρων που συνδέει μέσα στην ίδια ιστοσελίδα ή έγγραφο. Αντίστοιχη χρήση με τον τελεστή + έχει και ο τελεστής αφαίρεσης -, ο οποίος χρησιμοποιείται στις περιπτώσεις που ο χρήστης επιθυμεί να αναζητήσει συγκεκριμένους όρους, αλλά χωρίς τους όρους που συνήθως τους συνοδεύουν. Προφανώς ο χρήστης μπορεί να εισάγει περισσότερες από μια φορές τον τελεστή, προκειμένου να εξειδικεύσει ακόμα περισσότερο την ερώτηση του (και βέβαια το ίδιο ισχύει και για τον τελεστή +).

Οι συγκεκριμένοι δύο τελεστές, όταν υποστηρίζονται από τη μηχανή ή το εργαλείο αναζήτησης, προσφέρουν στο χρήστη μοναδική ευκολία, καθώς περιορίζουν σημαντικά το εύρος της αναζήτησης και εξειδικεύουν το ερώτημα του, ενώ παράλληλα η χρήση τους είναι ιδιαίτερα απλή και κατανοητή.

Θα πρέπει να σημειωθεί ότι σε πολλές περιπτώσεις ο χρήστης έχει την εναλλακτική δυνατότητα να χρησιμοποιήσει αντί των τελεστών αυτών τα *ανωφερή εισαγωγικά " "* ***(quotation marks)***, προκειμένου να περιορίσει το εύρος της αναζήτησής του. Αλλά θα πρέπει να είναι ιδιαίτερα προσεκτικός, αφού η χρήση τους δεν υποστηρίζεται από όλες τις μηχανές και τα εργαλεία αναζήτησης. Η χρήση των συγκεκριμένων τελεστών στην αρχή και το τέλος της φράσης ή των όρων που θέλουμε να αναζητήσουμε ως φράση καθοδηγεί τη μηχανή στο να τους αναζητήσει ακριβώς όπως έχουν διατυπωθεί μέσα στα εισαγωγικά. Το θετικό και στις δύο περιπτώσεις είναι η απλότητα της σύνταξης και οι δυνατότητες που προσφέρουν ως τελεστές περιορισμού των αποτελεσμάτων.

- **ΤΕΛΕΣΤΕΣ ΕΓΓΥΤΗΤΑΣ - ΧΡΗΣΗ ΤΟΥ ΤΕΛΕΣΤΗ NEAR**

Η *εγγύτητα* ***(proximity)*** αφορά το πόσο κοντά βρίσκονται οι αναζητήσιμοι όροι μεταξύ τους. Οι περισσότερες μηχανές αναζήτησης, τουλάχιστον όσες υποστηρίζουν την *πλήρη ευρετηρίαση του περιεχομένου κειμένου* ***(full-text indexing)***, επιτρέπουν την αναζήτηση *ολόκληρων φράσεων* ***(phrase search)***, όπου οι αναζητήσιμοι όροι βρίσκονται ο ένας ακριβώς δίπλα στον άλλο. Επιπλέον παρέχεται η δυνατότητα αναζήτησης των όρων αυτών ακόμη κι όταν βρίσκονται στην ίδια πρόταση.

Πολλές μηχανές αναζήτησης υποστηρίζουν μια παρεμφερή λειτουργία με εκείνη που περιγράφηκε προηγούμενα, τη λεγόμενη αναζήτηση *εγγύτητας* ***(proximity searching)***. Η αναζήτηση αυτή υλοποιείται κυρίως με τη χρήση του τελεστή εγγύτητας ***Near***, ο οποίος αναλαμβάνει να ανακτήσει δύο ή περισσότερους όρους έστω κι αν αυτοί

δε βρίσκονται ο ένας μετά τον άλλο, αλλά απλά στην ίδια πρόταση ή παράγραφο.

Οι αναζητήσεις εγγύτητας βασίζονται συνήθως στην εγγύτητα μεταξύ μόνο δύο όρων. Αν ένας χρήστης ήθελε να κάνει αναζήτηση μίας ολόκληρης φράσης, είναι συχνά δυνατό να εσωκλείσει μία φράση, όπως για παράδειγμα *'Multiobjective Linear Programming'*, σε εισαγωγικά, Σ' αυτή την περίπτωση, μόνο εκείνα τα έγγραφα που περιέχουν την ακριβή φράση θα ανακτηθούν.

- **FUZZY BOOLEAN LOGIC**

Αν και ιδιαίτερα διαδεδομένη στο χώρο της αναζήτησης πληροφοριών, η λογική Boolean παρουσιάζει και κάποια μειονεκτήματα, τα οποία οφείλονται κυρίως στην αυστηρότητά της. Τα σημαντικότερα μειονεκτήματα της είναι:

- ✓ Δεν υπάρχει άμεσος έλεγχος του μεγέθους του συνόλου των ανακτημένων εγγράφων. Έτσι μπορεί σε κάποιο ερώτημα του χρήστη η μηχανή αναζήτησης να επιστρέφει ένα υπερβολικά μεγάλο και δύσχρηστο σύνολο εγγράφων, ενώ για κάποιο άλλο ερώτημα να μην επιστρέφεται κανένα έγγραφο.
- ✓ Καθώς δεν υπάρχει δυνατότητα να αξιολογούνται τα έγγραφα που ανακτώνται, μπορεί τα αποτελέσματα μιας μηχανής αναζήτησης να μην ικανοποιούν τις προσδοκίες του χρήστη. Για παράδειγμα σε μια ερώτηση που όλοι οι όροι συνδέονται με τον τελεστή OR (KI OR K2 OR....OR K3) τα έγγραφα που ικανοποιούν μονάχα μια συνθήκη (δηλ. εκείνα που περιέχουν μονάχα έναν από τους ζητούμενους όρους) θεωρούνται ίσης χρησιμότητας με εκείνα που ικανοποιούν όλες τις συνθήκες. Ή σε μια ερώτηση που όλοι οι όροι συνδέονται με τον τελεστή AND (KI AND K2 AND... AND K3) ένα έγγραφο που ικανοποιεί όλες τις συνθήκες εκτός από μία (δηλ. περιέχει όλους τους ζητούμενους όρους πλην ενός) δεν ανακτάται από τη μηχανή αναζήτησης, αντιμετωπίζεται δηλαδή σαν να μην σχετίζεται καθόλου με την ερώτηση.
- ✓ Δε δίνεται στο χρήστη η δυνατότητα να δηλώσει τη σημασία κάποιου συγκεκριμένου όρου της ερώτησής του αποδίδοντας σε αυτόν μεγαλύτερο βάρος.
- ✓ Εμπειρικά αποτελέσματα δείχνουν ότι ο αυστηρός τρόπος με τον οποίο λειτουργούν οι τελεστές Boolean δε συμφωνεί με τον ανθρώπινο τρόπο σκέψης, αξιολόγησης και λήψης αποφάσεων. Για παράδειγμα στην περίπτωση ερώτησης που οι όροι συνδέονται με τον τελεστή OR, για τον ανθρώπινο τρόπο σκέψης ένα έγγραφο το οποίο καλύπτει τους περισσότερους δυνατούς όρους είναι το πλέον σημαντικό. Και βέβαια στην περίπτωση ερώτησης που οι όροι συνδέονται με τον

τελεστή AND έγγραφα τα οποία ικανοποιούν μερικώς την ερώτηση δεν απορρίπτονται πλήρως.

Τα παραπάνω μειονεκτήματα έρχεται να αντιμετωπίσει η ***Fuzzy Boolean Logic*** παραμετροποιώντας τους τελεστές Boolean και αποδίδοντας διαφορετικά βάρη στους όρους κάθε ερώτησης. Έτσι κάθε φορά που ο χρήστης υποβάλει μια ερώτηση, η μηχανή αναζήτησης αποδίδει σε κάθε έγγραφο ένα *βαθμό συσχέτισης* ***(degree of match)*** που αποτελεί μέτρο ικανοποίησης της ερώτησης. Με άλλα λόγια δημιουργείται ένα σύνολο εγγράφων τα οποία ανάλογα με τη συσχέτιση τους με την ερώτηση κατατάσσονται με κάποιο βαθμό στο σύνολο αυτό.

Όταν ένα *query string* περιέχει περισσότερους από έναν όρους, εάν δε χρησιμοποιηθούν τελεστές Boolean, χρησιμοποιείται συνήθως ***Fuzzy Boolean***. Τα έγγραφα κατατάσσονται σύμφωνα με τον αριθμό των όρων που ταιριάζουν. Αυτό τείνει να βελτιώσει την ακρίβεια στην κορυφή της λίστας..

Ο όρος ***Fuzzy Searching*** (Kowalski, 1997) περιγράφει έναν μηχανισμό, που συχνά χρησιμοποιείται ως αποτέλεσμα φτωχής ανάκλησης. Αναζητώνται όροι με παρόμοια ορθογραφία με τους όρους της αναζήτησης, στην περίπτωση που οι όροι της αναζήτησης δεν έχουν γραφεί σωστά.

❑ ΑΝΑΖΗΤΗΣΕΙΣ ΣΥΝΩΝΥΜΩΝ ΚΑΙ ΕΠΕΚΤΑΣΗ ΕΡΩΤΗΜΑΤΟΣ (THESAURUS SEARCHES AND QUERY EXPANSION)

Μία μέθοδος βελτίωσης της ανάκλησης είναι η ανάκτηση εγγράφων που δεν περιέχουν μόνο τους όρους της αναζήτησης, αλλά και συνώνυμα αυτών των όρων. Τη δυνατότητα αυτή την παρέχουν τα *ηλεκτρονικά λεξικά συνωνύμων* ***(electronic thesauri).*** Το πρόβλημα αυτής της προσέγγισης είναι ότι συχνά η εστίαση των ερωτημάτων μπορεί να αποπροσανατολιστεί από ακατάλληλα συνώνυμα, με αποτέλεσμα τη βελτίωση της ανάκλησης αλλά και την καταστροφικά χαμηλή ακρίβεια. Για την αποφυγή αυτού του αρνητικού αποτελέσματος οι μηχανές αναζήτησης που υποστηρίζουν αυτό το χαρακτηριστικό, συχνά εμφανίζουν στο χρήστη μία λίστα με συνώνυμα, που σχετίζονται με τους αρχικούς του όρους, ούτως ώστε να μπορεί να επιλέξει αυτά που θεωρεί σχετικά. Το ***Alta Vista's Live Topics*** είναι ένα παράδειγμα αυτής της προσέγγισης.

Τα *στατιστικά λεξικά συνωνύμων* ***(statistical Thesauri)*** παρέχουν μία εναλλακτική μέθοδο. Αντί να αναζητούν εννοιολογικά συνώνυμα, προσθέτουν στο *query* όρους, οι

οποίοι έχουν *στατιστικά υψηλή σύμπτωση* ***(statistically high coincidence)*** με τους όρους αναζήτησης του χρήστη. Αυτή η προσέγγιση είναι γνωστή ως αυτόματη *επέκταση ερωτήματος* ***(query expansion).*** Συχνά, το αρχικό ερώτημα μεταβάλλεται ώστε να αποκαλυφθεί ποιοι όροι είναι πιο πιθανό να εστιάζουν στο ερώτημα - εκείνοι που είναι λιγότερο συνηθισμένοι εντός της βάσης δεδομένων - και στη συνέχεια το ερώτημα επεκτείνεται με στατιστικά συνώνυμα εκείνων των όρων. Η μηχανή αναζήτησης ***Muscat EuroFerret*** χρησιμοποιεί πιθανοθεωρητική ανάκτηση σε συνδυασμό με *«ανατροφοδότηση σχετικότητας»* ***('Relevance Feedback')***. Με αυτόν τον τρόπο δίνεται η δυνατότητα στο χρήστη να δηλώσει ποια αποτελέσματα είναι πιο σχετικά με το ερώτημα του, ούτως ώστε στη συνέχεια να αναζητηθούν παρόμοια έγγραφα με σημαντικούς όρους υψηλής σύμπτωσης.

- **STEMMING AND TERM MASKING, TRUNCATION, WILDCARDS**

Ο μηχανισμός ανάκτησης μπορεί επίσης να βελτιώσει την ανάκληση εκτελώντας *«ξεγύμνωμα»* ***(stripping)*** καταλήξεων ή αλλιώς ***'stemming'*** στην ακολουθία του ερωτήματος. Αυτό σημαίνει ότι οποιοσδήποτε όρος που καταλήγει σε 's', 'ed', 'ing', 'ology', 'ologist', 'ological' κ.λ.π. θα αποκοπεί, έτσι ώστε, για παράδειγμα, μία έρευνα για *'Psychological Conferences'* θα βρει πολύ σχετικό ένα έγγραφο που περιέχει τις λέξεις *'Psychological Conferences'*.

Το *stemming* χρησιμοποιείται συχνά για να βελτιώσει την ανάκληση, αλλά μπορεί να έχει αρνητική επίδραση στην ακρίβεια. Ο *stemming* αλγόριθμος του Porter (Porter, 1980) αναγνωρίζει λέξεις με συγκεκριμένες καταλήξεις και τις αντικαθιστά με *stemmed* εκδοχές. Αυτό μπορεί να οδηγήσει σε μειωμένη ακρίβεια, όπως αναφέρει ο Kowalski, 1997). Το 'memorial' και το 'memorise' έχουν πολύ διαφορετικές σημασίες, αλλά θα γίνονταν και τα δύο 'memory' σύμφωνα με τον αλγόριθμο του Porter. Μία εναλλακτική μέθοδος είναι η χρησιμοποίηση μίας προσέγγισης, *βασισμένης σε λεξικό* ***(dictionary based),*** όπως το Kstem (Kowalski, 1997), όπου με αντικατάσταση της λέξης με το πιο ακριβές stem από ένα λεξικό, λαμβάνονται πιο ακριβή stems. Η αξιολόγηση του Frakes (Frakes et al., 1992) των πειραμάτων *stemming* επιβεβαίωσαν ότι οι *stemming* αλγόριθμοι έχουν θετική επίδραση μόνο στην ανάκληση και όχι στην ακρίβεια.

Μία εντελώς διαφορετική προσέγγιση είναι η χρησιμοποίηση ***term masking*** *(κάλυψη όρου)* (Kowalski, 1997) ή ***truncation*** *(αποκοπή)* ή ***wildcard*** στο ερώτημα. Οι καταλήξεις των λέξεων *καλύπτονται* ***(masked)*** και κάθε συνδυασμός χαρακτήρων μετά τους *μη καλυπτόμενους* ***(unmasked)*** χαρακτήρες μπορεί να γίνει αποδεκτός ως

ταίριασμα/ αντιστοίχιση ***(match)***. Για παράδειγμα, ο καλυπτόμενος όρος psycho* θα μπορούσε να ταιριάξει με τους όρους psycho, psychology, psychologist, psychological, κ.λ.π.

3.3.3.3.2. Vector Space Models

Μια δεύτερη κατηγορία τεχνικών αναζήτησης που μελετάται τα τελευταία χρόνια είναι τα ***Vector Space Models***. Πρόκειται για μεθόδους που βασίζονται σε ένα τυπικό θεωρητικό μαθηματικό μοντέλο για την αναζήτηση: μοντελοποιούν τα έγγραφα ως ένα σύνολο όρων, καθένας εκ των οποίων μπορεί να αξιολογηθεί και να χρησιμοποιηθεί και μεμονωμένα, εκτελούν τις ερωτήσεις συγκρίνοντας την απεικόνιση της ερώτησης με την απεικόνιση κάθε εγγράφου στο χώρο και μπορούν να ανακτήσουν και έγγραφα τα οποία δεν περιέχουν κάποιους από τους όρους αναζήτησης. Αν και παρουσιάζουν ομοιότητες με κάποιες άλλες τεχνικές αναζήτησης, οι *vector space* τεχνικές έχουν κάποια ιδιαίτερα γνωρίσματα που τις διαφοροποιούν.

Βασική προϋπόθεση σε κάθε *vector space model* είναι ότι η ερμηνεία κάθε εγγράφου είναι απόρροια των όρων που το συνθέτουν. Τα έγγραφα απεικονίζονται ως διανύσματα των όρων $d=\{d_1,d_2, ...d_n\}$ όπου d_i με $(1<i<\eta)$ είναι μη αρνητικός αριθμός που δηλώνει την μεμονωμένη ή πολλαπλή εμφάνιση του όρου *i* στο έγγραφο d.

Πρόκειται προφανώς για το αποτέλεσμα της αυτόματης ανάλυσης του κειμένου του εγγράφου με την αναγνώριση των βασικών όρων του κειμένου και τη δημιουργία της απεικόνισης του κειμένου στο η-διάστατο χώρο. Αντίστοιχα κάθε ερώτηση απεικονίζεται με ένα διάνυσμα της μορφής $q=\{q_1,q_2, ...q_m\}$, όπου q_i με $(1<i<m)$ είναι ένας μη-αρνητικός αριθμός που δηλώνει τη συχνότητα εμφάνισης του όρου *i* στην ερώτηση q (συνήθως μάλιστα είναι ο αριθμός 1 και δηλώνει απλά την ύπαρξη του όρου *i* στην ερώτηση q). Τα διανύσματα που απεικονίζουν τα έγγραφα, όπως και το διάνυσμα που απεικονίζει την ερώτηση ορίζουν τη θέση των αντίστοιχων αντικειμένων στο χώρο. Η ανάκτηση των εγγράφων γίνεται με βάση τον υπολογισμό της "απόστασης" του διανύσματος της ερώτησης από τα διανύσματα των υπόλοιπων αντικειμένων στο χώρο.

Τα *vector space models* διαχειρίζονται κάθε όρο μεμονωμένα, δημιουργώντας έτσι ένα είδος ανάστροφου πίνακα. Επιπλέον όμως σε κάθε όρο μπορεί και να αποδοθεί και κάποιο βάρος κάνοντας τον έτσι περισσότερο ή λιγότερο σημαντικό για το συγκεκριμένο έγγραφο ή και για το σύνολο των εγγράφων.

Υπάρχουν διάφορα μέτρα προσδιορισμού της συσχέτισης ανάμεσα σε μια

ερώτηση και ένα σύνολο όρων ή ένα έγγραφο που χρησιμοποιούνται από τα *vector space models*. Έτσι για παράδειγμα το εσωτερικό γινόμενο είναι ένας τρόπος μέτρησης της συσχέτισης ανάμεσα σε μια ερώτηση και κάποιους όρους ή ένα έγγραφο, όπου υπολογίζεται η Ευκλείδεια απόσταση των αντίστοιχων διανυσμάτων στο χώρο. Άλλο μέτρο σύγκρισης μπορεί να είναι η γωνία που σχηματίζουν στο χώρο τα διανύσματα ή ακόμα και η κατεύθυνση των διανυσμάτων.

Τα *vector space models* αναπτύχθηκαν για να αντιμετωπιστούν πολλά από τα προβλήματα που δημιουργεί η χρήση των αυστηρών, λεξικογραφικών τεχνικών αναζήτησης. Συγκεκριμένα υπάρχουν λέξεις με πολλαπλές σημασίες πράγμα που καθιστά δύσκολο για μια λεξικογραφική τεχνική σύγκρισης να εντοπίσει τη διαφορά δύο εγγράφων που χρησιμοποιούν την ίδια λέξη αλλά με διαφορετικό τρόπο, αφού δεν αντιλαμβάνεται το γενικότερο πλαίσιο μέσα στο οποίο χρησιμοποιείται η λέξη αυτή. Επίσης, καθώς υπάρχουν πολλοί διαφορετικοί τρόποι για να αποδοθεί μια έννοια, έγγραφα που έχουν κοινό θέμα μπορεί να μη χρησιμοποιούν κοινή ορολογία. Έτσι υπάρχει περίπτωση με τη χρήση λεξικογραφικών τεχνικών αναζήτησης να μην ανακτώνται όλα τα έγγραφα που είναι σχετικά με μια ερώτηση (ή στη χειρότερη περίπτωση να μην ανακτάται κανένα έγγραφο) γιατί χρησιμοποιείται διαφορετική ορολογία.

Τοποθετώντας όρους, έγγραφα και ερωτήσεις στο διανυσματικό χώρο και υπολογίζοντας τις συσχετίσεις ανάμεσα σε ερωτήσεις και έγγραφα, τα *vector space models* μπορούν και να κατατάξουν τα έγγραφα που ανακτώνται για κάθε ερώτηση. Σε αντίθεση με τις λεξικογραφικές τεχνικές σύγκρισης που δεν κάνουν κατάταξη των εγγράφων ή κάνουν μια πολύ πρόχειρη κατάταξη (βασιζόμενες για παράδειγμα μόνο στη συχνότητα εμφάνισης ενός όρου), τα vector space models βασιζόμενα στην Ευκλείδεια απόσταση ή τη γωνία ανάμεσα σε ερώτηση και όρους ή έγγραφο, μπορούν με μεγαλύτερη αξιοπιστία να ανακτήσουν και να κατατάξουν τα καταλληλότερα κάθε φορά έγγραφα.

3.3.3.3.3. Πιθανοθεωρητικά Μοντέλα (Probabilistic Models)

Η χρήση πιθανοθεωρητικών μοντέλων στο χώρο της ανάκτησης πληροφοριών ξεκίνησε από τις αρχές της δεκαετίας του ’60 (Φωστιέρη, 2001), όμως μόλις τα τελευταία χρόνια άρχισε να μελετάται συστηματικά και να βρίσκει διάφορες εφαρμογές, μια από αυτές είναι και οι μηχανές αναζήτησης.

Όπως ακριβώς και στα δύο άλλα μοντέλα αναζήτησης που παρουσιάστηκαν

παραπάνω *(set theoretic και vector space models)* όταν πραγματοποιείται αναζήτηση σε ένα σύνολο εγγράφων, ο στόχος είναι να συλλεχθούν τα σχετικά με τις απαιτήσεις του χρήστη έγγραφα, χωρίς να συλλεχθούν τα μη σχετικά. Η σχετικότητα ενός εγγράφου ως προς τις απαιτήσεις του χρήστη δεν μπορεί βέβαια να προσδιοριστεί με ακρίβεια, καθώς κάτι τέτοιο θα ήταν εφικτό μόνο εάν ο χρήστης προχωρούσε σε πλήρη ανάγνωση του εγγράφου. Τα πιθανοθεωρητικά μοντέλα αναζήτησης προσπαθούν να εκτιμήσουν τη σχετικότητα κάθε εγγράφου, υπολογίζοντας για κάθε έγγραφο την πιθανότητα να είναι σχετικό:

☞ ***PQ (relevance/ document),*** όπου το Q δηλώνει ότι εκτιμάται η συσχέτιση του εγγράφου ως προς ένα συγκεκριμένο *ερώτημα* ***(query)*** του χρήστη.

Όπως ακριβώς και οι συναρτήσεις συσχέτισης που εφαρμόζονται στα *Vector space Models* έτσι και τα πιθανοθεωρητικά μοντέλα αποδίδουν μια βαθμολογία σε κάθε έγγραφο, που απεικονίζει τη σχετικότητα του εγγράφου ως προς την ερώτηση του χρήστη. Το βασικό θεώρημα που χρησιμοποιείται από τα πιθανοθεωρητικά μοντέλα αναζήτησης είναι το ***θεώρημα του Bayes***.

3.3.3.4. Ταξινόμηση των Σελίδων (Ranking)

Μόλις ο χρήστης εισαγάγει μια ερώτηση, αυτή εξυπηρετείται από το πρόγραμμα που ψάχνει τη βάση δεδομένων της μηχανής αναζήτησης για να καθορίσει:

- ποια αρχεία ταιριάζουν στην ερώτηση, και
- με ποια σειρά αυτά τα αρχεία πρέπει να ταξινομηθούν και να εμφανιστούν στη συνέχεια στο χρήστη.

Αυτές οι δύο λειτουργίες μπορούν να λειτουργήσουν μάλλον ανεξάρτητα ή μπορούν να αποτελούν ουσιαστικά μια ενιαία λειτουργία.

Η πρώτη λειτουργία, ο προσδιορισμός των σχετικών με την ερώτηση του χρήστη αρχείων, είναι βασισμένη είτε (α) στη χρήση μιας προεπιλεγμένης προσέγγισης στην οποία ο χρήστης έχει εισαγάγει τους όρους, τις φράσεις, ή τις προτάσεις χωρίς σύνταξη, είτε (β) χρησιμοποιώντας ο χρήστης για τη δημιουργία του ερωτήματος μια καθορισμένη σύνταξη που περιλαμβάνει κριτήρια όπως τελεστές Boolean, τελεστές εγγύτητας, κ.λ.π.

Όταν ο χρήστης δεν χρησιμοποιεί μια δομημένη σύνταξη, η απλοϊκότερη προσέγγιση για τον προσδιορισμό των αρχείων είναι για το πρόγραμμα της ανάκτησης να λάβει όλες ή μερικές από τις λέξεις που ο χρήστης έχει εισαγάγει, να τις συνδέσει, με

έναν τελεστή Boolean (AND, OR), και να αναζητήσει τη βάση δεδομένων χρησιμοποιώντας αυτή τη Boolean έκφραση. Αυτό μπορεί να αναφερθεί ως "αναζήτηση με βάση τη *φυσική γλώσσα* ***(natural languange)***". Στις περισσότερες από τις κυριότερες μηχανές αναζήτησης, η σύγκριση των Boolean τελεστών είναι ένα αναπόσπαστο τμήμα της όλης διαδικασίας. Υπάρχουν εναλλακτικές λύσεις που παρακάμπτουν τους Boolean τελεστές και προσδιορίζουν τα ανακτημένα αρχεία βάσει κάποιων δημοφιλών παραγόντων και της περίπλοκης γλωσσικής ανάλυσης που περιλαμβάνει τέτοιους παράγοντες.

Όταν ο χρήστης χρησιμοποιεί συγκεκριμένη σύνταξη, όπως Boolean, μπορεί ακόμη και να αγνοήσει τον αλγόριθμο προεπιλογής μιας μηχανής. Μια ενιαία μηχανή αναζήτησης μπορεί να παρέχει όλες αυτές τις εναλλακτικές λύσεις: έναν προεπιλεγμένο αλγόριθμο βασισμένο στους Boolean τελεστές και σε άλλα κριτήρια σύνταξης που ο ίδιος ο χρήστης εφαρμόζει, όπως επίσης και στη γλωσσική ανάλυση.

Με την πρώτη λειτουργία του προγράμματος πραγματοποιείται ο προσδιορισμός των κατάλληλων "αρχείων", η δεύτερη σημαντική λειτουργία του προγράμματος ανάκτησης/ ταξινόμησης της μηχανής αναζήτησης είναι ο καθορισμός της σχετικότητας κάθε αρχείου. Αυτό εκφράζεται συχνά ως "αποτέλεσμα" ***(score)*** ή "ταξινόμηση" ***(ranking)***- δηλ., η εκτίμηση του προγράμματος ως προς το πόσο καλά ένα συγκεκριμένο αρχείο σχετίζεται με την ερώτηση. Όπως αναφέρθηκε προηγουμένως, αυτό μπορεί να επιτευχθεί με την πρώτη λειτουργία, με την "ταξινόμηση" επιτυγχάνεται εάν το αρχείο ανακτάται ή όχι (μόνο εκείνα που ικανοποιούν τα κατώτατα όρια θα εμφανισθούν στα αποτελέσματα).

Λόγω της ανταγωνιστικής φύσης της βιομηχανίας των μηχανών αναζήτησης, οι λεπτομέρειες των αλγορίθμων ανάκτησης και ταξινόμησης φυλάσσονται καλά. Για την αποτελεσματική χρήση των μηχανών αναζήτησης, είναι χρήσιμο να αναφερθούμε με περισσότερη λεπτομέρεια στους παράγοντες που επηρεάζουν- τους παράγοντες που η μηχανή αναζήτησης ψάχνει σε ένα αρχείο για να καθορίσει εάν ανακτηθεί και πώς πρέπει να ταξινομηθεί από πλευρά σχετικότητας, έπειτα καθορίζεται και η σειρά εμφάνισης των αρχείων στο χρήστη.

Μία ***crawler-based*** μηχανή αναζήτησης επιλέγει μέσω εκατομμυρίων σελίδων και παρουσιάζει στο χρήστη τα αποτελέσματα εκείνα που ταιριάζουν με το ερώτημά του. Τα αποτελέσματα μάλιστα κατατάσσονται έτσι ώστε τα πιο σχετικά να βρίσκονται στην κορυφή της λίστας.

Προκειμένου οι μηχανές αναζήτησης να καθορίσουν την σχετικότητα της κάθε

σελίδας με το ερώτημα του χρήστη ακολουθούν μία ομάδα κανόνων, έναν αλγόριθμο. Η ακριβής λειτουργία του αλγορίθμου για κάθε μία μηχανή αναζήτησης είναι άκρως απόρρητη για εμπορικούς λόγους. Ωστόσο, όλες οι κύριες μηχανές αναζήτησης ακολουθούν τους παρακάτω γενικούς κανόνες:

- ✓ Ένας από τους κυριότερους κανόνες σε έναν αλγόριθμο κατάταξης αφορά την *τοποθεσία* ***(location)*** και τη *συχνότητα* ***(frequency)***. Για παράδειγμα, οι σελίδες των οποίων οι όροι της αναζήτησης εμφανίζονται στην HTML *ετικέτα τίτλου* ***(title tag)*** συχνά θεωρούνται πιο σχετικές.
- ✓ Οι μηχανές αναζήτησης ελέγχουν επίσης αν οι λέξεις-κλειδιά της αναζήτησης εμφανίζονται κοντά στην κορυφή μίας ιστοσελίδας, όπως για παράδειγμα στην επικεφαλίδα ή στις πρώτες παραγράφους του κειμένου. Υποθέτουν, δηλαδή ότι κάθε σχετική με το θέμα σελίδα θα αναφέρει αυτές τις λέξεις από την αρχή.
- ✓ Η συχνότητα είναι ο άλλος σημαντικός παράγοντας σχετικά με τον τρόπο που οι μηχανές αναζήτησης καθορίζουν τη σχετικότητα. Μία μηχανή αναζήτησης θα αναλύσει πόσο συχνά εμφανίζονται οι λέξεις-κλειδιά σε σχέση με άλλες λέξεις σε μία ιστοσελίδα. Αυτές με την υψηλότερη συχνότητα θεωρούνται συχνά πιο σχετικές από άλλες ιστοσελίδες.
- ✓ Καμία βέβαια μηχανή αναζήτησης δε χρησιμοποιεί τη *μέθοδο τοποθεσίας/ συχνότητας* ***(location/frequency method)*** με ακριβώς τον ίδιο τρόπο. Αυτός είναι και ο λόγος που η ίδια αναζήτηση σε διαφορετικές μηχανές αναζήτησης παράγει διαφορετικά αποτελέσματα.
- ✓ Ο αριθμός των όρων σε μια ιστοσελίδα που ταιριάζουν με την ερώτηση του χρήστη- εάν η ερώτηση αποτελείται από τρεις λέξεις, εκείνο το αρχείο που έχει και τις τρεις λέξεις θα ταξινομηθεί σε καλύτερη σειρά από εκείνο το αρχείο που έχει μόνο έναν ή δύο από τους όρους.
- ✓ Το πόσο σπάνιος είναι ο όρος μιας ερώτησης- εάν η ερώτηση έχει έναν όρο πολύ κοινό και έναν δεύτερο που εμφανίζεται λίγες φορές στη βάση δεδομένων της μηχανής αναζήτησης, το αρχείο που περιέχει τον σπάνιο όρο μπορεί να ταξινομηθεί υψηλότερα από το άλλο αρχείο.
- ✓ Η εγγύτητα των όρων- εάν δύο από τους όρους μιας ερώτησης είναι μαζί, ο ένας δίπλα στον άλλον, αυτό μετράει περισσότερο από το να είναι μακριά.
- ✓ Η ημερομηνία των εγγράφων- τα πιο πρόσφατα αρχεία ταξινομούνται υψηλότερα από τα παλαιότερα αρχεία.
- ✓ Κάποιες μηχανές αναζήτησης καταχωρούν σε ευρετήριο πιο πολλές ιστοσελίδες

από άλλες. Μερικές μηχανές αναζήτησης επίσης, καταχωρούν σε ευρετήριο τις ιστοσελίδες πιο συχνά από άλλες. Αυτό έχει ως αποτέλεσμα, ότι καμία μηχανή αναζήτησης δεν έχει ακριβώς την ίδια συλλογή ιστοσελίδων με κάποια άλλη, όταν διενεργεί την αναζήτηση. Φυσικά κάτι τέτοιο δημιουργεί διαφορές, όταν συγκρίνονται τα αποτελέσματα.

- ✓ Ακόμα, μερικές μηχανές αναζήτησης αυξάνουν τη *βαθμολογία* ***(score)*** σχετικότητας των σελίδων, στις οποίες καταλήγουν πολλά *links*, με τη λογική ότι αυτές είναι δημοφιλείς σελίδες και άρα ο κόσμος θα θέλει να τις ανακτήσει. Ομοίως, αν μία μηχανή αναζήτησης υποστηρίζει μία λίστα *«αναθεωρημένων sites»*, τότε αυτά τα *sites* μπορεί να λάβουν υψηλότερο βαθμό σχετικότητας, επειδή ένα αναθεωρημένο *site* δείχνει ότι είναι πάνω από το μέσο όρο σε ποιότητα και συνεπώς, είναι πολύ πιθανό οι σελίδες του να είναι εκείνες που κάποιος θέλει να ανακτήσει.
- ✓ Μερικές μηχανές αναζήτησης μπορεί ακόμα, να αποκλείσουν κάποιες σελίδες από το ευρετήριο, αν εντοπίσουν ‘***spamming’***. Για παράδειγμα, μπορεί μία λέξη να επαναλαμβάνεται εκατοντάδες φορές σε μία σελίδα, προκειμένου να αυξηθεί η συχνότητα της και να προωθηθεί υψηλότερα στη λίστα. Οι μηχανές αναζήτησης ελέγχουν για συνηθισμένες μεθόδους *spamming* με διάφορους τρόπους, όπως για παράδειγμα, λαμβάνοντας υπόψη τα παράπονα των χρηστών τους.

Βέβαια εκτός από τους παραπάνω παράγοντες που λαμβάνουν υπ’ όψιν οι μηχανές αναζήτησης για την ταξινόμηση των εγγράφων, θα πρέπει να αναφερθεί και να συμπεριληφθεί ότι και το ***Meta Tag*** παίζει σημαντικό ρόλο στην ταξινόμηση των εγγράφων.

✓ META TAG

Το HTML *META tag (Miller*, 1997), παρέχει ένα μηχανισμό για τη βελτίωση της ακρίβειας των υπαρχόντων αυτοματοποιημένων εργαλείων, δίνοντας στους συντάκτες του Δικτύου ***(Web authors)*** τη δυνατότητα να καθορίσουν τα δικά τους *μεταδεδομένα* ***(metadata).***

Τα metadata είναι πληροφορίες σχετικά με μία ιστοσελίδα, οι οποίες είναι γραμμένες στην HTML μορφή της ίδιας της σελίδας, αλλά δεν εμφανίζονται αυτόματα στο χρήστη. Οι συντάκτες μπορούν να περιλάβουν περιγραφικές πληροφορίες σχετικά με τις σελίδες τους ως *metadata*, οι οποίες στη συνέχεια εντοπίζονται από τις μηχανές

αναζήτησης και εμφανίζονται στα αποτελέσματα της αναζήτησης.

Οι περισσότερες από τις μεγάλες μηχανές αναζήτησης υποστηρίζουν πλέον το META tag, εκτός από την Excite. Τα *keywords* και τα *description tags* είναι τα δύο πιο σημαντικά META tags, που αφορούν τις μηχανές αναζήτησης:

```
<HEAD>
<TITLE> Search Engines - An Evaluation </ TITLE>
<META NAME = "DESCRIPTION" CONTENT = "World Wide Web Search Engines: an Evaluation of tools and Methodologies">
<META NAME = "KEYWORDS" CONTENT = "Search Engines, Classified Directories, Meta Search Engines, Subject Specific gateways, indexing, Boolean Syntax">
</HEAD>
```

Αυτά τα *tags*, που βρίσκονται στο στοιχείο HEAD μίας HTML ιστοσελίδας, επιτρέπουν στο συντάκτη να καθορίσει όρους ευρετηρίου και περιγραφές κειμένου των εγγράφων του. Τα αυτοματοποιημένα εργαλεία στη συνέχεια αναγνωρίζουν αυτά τα *tags* και χρησιμοποιούν την πληροφορία που παρέχεται από τον χρήστη, προκειμένου να δημιουργήσουν πιο ακριβή *metadata*. Αυτή η διαδικασία συνδυάζει αυτοματισμό με *metadata* που καθορίζονται από τον άνθρωπο και γενικά θεωρείται ότι βελτιώνει την ακρίβεια των αυτοματοποιημένων εργαλείων. Η Excite, ωστόσο, αντιτίθεται στο META tag με τη λογική ότι μπορεί να έχει λάθος εφαρμογή. Οι συντάκτες θα μπορούσαν να καθορίσουν ανακριβή *metadata*, άσχετα με το περιεχόμενο των σελίδων τους, για να δελεάσουν τους χρήστες και να έχουν περισσότερες επισκέψεις των σελίδων τους.

Οι *crawler-based* μηχανές αναζήτησης έχουν πλέον αποκτήσει αρκετή εμπειρία με τους ***webmasters***, οι οποίοι συνεχώς ανασυντάσσουν τις ιστοσελίδες τους, σε μία προσπάθεια να αποκομίσουν καλύτερη κατάταξη. Κάποιοι έμπειροι *Webmasters* μπορούν να επηρεάσουν τη μέθοδο τοποθεσίας/συχνότητας μίας συγκεκριμένης μηχανής αναζήτησης. Για αυτό το λόγο, όλες οι μεγάλες μηχανές αναζήτησης πλέον χρησιμοποιούν και κριτήρια κατάταξης, τα οποία δεν εξαρτώνται από την ιστοσελίδα ***("off the page")***.

Αυτοί οι παράγοντες, που δεν εξαρτώνται από τη σελίδα, είναι εκείνοι που δεν μπορούν οι *Webmasters* να επηρεάσουν εύκολα. Ο πιο σημαντικός παράγοντας είναι η

ανάλυση των συνδέσμων ***(link analysis)***. Αναλύοντας τον τρόπο με τον οποίο οι σελίδες συνδέονται μεταξύ τους, μία μηχανή αναζήτησης μπορεί, αφενός να καθορίσει με τι σχετίζεται μία σελίδα, αφετέρου το αν αυτή η σελίδα θεωρείται «σημαντική» και επομένως πρέπει να καταταχθεί υψηλά. Επίσης, χρησιμοποιούνται προχωρημένες τεχνικές προκειμένου να εντοπιστούν προσπάθειες που κάνουν οι Webmasters ώστε να δημιουργήσουν «τεχνητούς» συνδέσμους και άρα να βελτιώσουν την κατάταξη των σελίδων τους.

Ένας άλλος παράγοντας, ο οποίος δεν εξαρτάται από τη σελίδα είναι η μέτρηση ***clickthrough***. Συνοπτικά, αυτό σημαίνει ότι μία μηχανή αναζήτησης μπορεί να ελέγχει ποια αποτελέσματα κάποιος επιλέγει για μία συγκεκριμένη αναζήτηση, οπότε τελικά να «ρίχνει» σε κατάταξη κάποιες υψηλά- κατατασσόμενες σελίδες που δεν είναι δημοφιλείς, ενώ ταυτόχρονα προωθεί χαμηλότερα- κατατασσόμενες σελίδες οι οποίες έχουν πολλούς επισκέπτες. Όπως συμβαίνει και με την ανάλυση συνδέσμων, χρησιμοποιούνται συστήματα, ώστε να εντοπιστούν τεχνητοί σύνδεσμοι, οι οποίοι δημιουργούνται από επιτήδειους Webmasters.

- ***Αλγόριθμοι Ranking***

Ποικίλες τεχνικές έχουν αναπτυχθεί για την *ταξινόμηση* ***(ranking)*** των ανακτημένων εγγράφων μιας μηχανής αναζήτησης για μια δεδομένη ερώτηση του χρήστη. Θα αναφερθούμε σε μερικές κλασσικές τεχνικές που μπορούν να τροποποιηθούν ώστε να χρησιμοποιηθούν από τις μηχανές αναζήτησης του ιστού (Baeza-Yates, Ribeiro-Neto, 1999, Berry, Browne 1999, Frakes, Baeza-Yates, 1992). Παρουσιάζονται επίσης και τεχνικές που αναπτύσσονται συγκεκριμένα για το Web.

Οι αναλυτικές πληροφορίες σχετικά με τους αλγορίθμους ταξινόμησης που χρησιμοποιούνται από τις κυριότερες μηχανές αναζήτησης δεν είναι δημόσια διαθέσιμες, ωστόσο φαίνεται ότι οι περισσότεροι χρησιμοποιούν την τεχνική *απόδοσης βάρους σε κάθε όρο* ***(Indexing Term Weighting)*** ή παραλλαγές αυτής καθώς και το ***Vector Space Model*** (Baeza-Yates, Ribeiro-Neto, 1999) που έχει αναφερθεί παραπάνω.

Στα *Vector Space Models*, κάθε έγγραφο (στη βάση δεδομένων της μηχανής αναζήτησης) απεικονίζεται από ένα διάνυσμα, κάθε συντεταγμένη αντιπροσωπεύει μια ιδιότητα του εγγράφου (Salton, 1971). Στην ιδανική περίπτωση, μόνο εκείνα τα χαρακτηριστικά του εγγράφου που μπορούν να βοηθήσουν στη διάκρισή του ενσωματώνονται. Σε ένα πρότυπο Boolean, κάθε συντεταγμένη του διανύσματος είναι μηδέν (όταν η αντίστοιχη ιδιότητα είναι απούσα) ή μονάδα (όταν η αντίστοιχη ιδιότητα

είναι παρούσα). Πολλές βελτιώσεις έχουν γίνει στο πρότυπο Boolean. Η πιο ευρέως χρησιμοποιούμενη τεχνική είναι αυτή της *απόδοσης βάρους σε κάθε όρο*, η οποία λαμβάνει υπ' όψιν τη συχνότητα εμφάνισης μιας ιδιότητας (π.χ., keyword) ή της θέσης εμφάνισής της (π.χ., keyword στον τίτλο, στην κεφαλίδα ενότητας, ή στην περίληψη του εγγράφου). Στα απλούστερα συστήματα ανάκτησης και ταξινόμησης, κάθε ερώτημα μοντελοποιείται επίσης από ένα διάνυσμα κατά τον ίδιο τρόπο με τα έγγραφα. Η ταξινόμηση ενός εγγράφου με βάση το ερώτημα του χρήστη καθορίζεται από *'την απόστασή του'* από το διάνυσμα του ερωτήματος. Ένα συχνά χρησιμοποιημένο κριτήριο είναι η γωνία που καθορίζεται από την ερώτηση και το διάνυσμα του εγγράφου. Η ταξινόμηση ενός εγγράφου βασίζεται στον υπολογισμό της παραπάνω γωνίας. Δεν είναι πρακτικό για τις πολύ μεγάλες βάσεις δεδομένων.

Ένας από τους ευρύτερα διαδεδομένους αλγόριθμους που βασίζεται στο *Vector Space Model* για τη μείωση του μεγέθους του εγγράφου που αφορά το πρόβλημα της ταξινόμησης είναι ο ***Latent Semantic Indexing (LSI)*** (Deerwester et al., 1990). Το μοντέλο αυτό μικραίνει το πρόβλημα της ανάκτησης και ταξινόμησης σε μια από τις σημαντικά μικρότερη διάσταση, έτσι ώστε η ανάκτηση από τις πολύ μεγάλες βάσεις δεδομένων να μπορεί να πραγματοποιείται σε πραγματικό χρόνο. Αν και ποικίλοι αλγόριθμοι, βασισμένοι στο *Vector Space Model*, έχουν προταθεί για την *ομαδοποίηση* ***(clustering)*** των εγγράφων ώστε να διευκολύνουν την ανάκτηση και την ταξινόμηση τους, το *LSI* είναι ένα από τους λίγους που λαμβάνει υπ' όψιν επιτυχώς *τη συνωνυμία* ***(synonymy)*** και *την πολυσημία* ***(polysemy)***. Η συνωνυμία αναφέρεται στην ύπαρξη ισοδύναμων ή παρόμοιων όρων, οι οποίοι μπορούν να χρησιμοποιηθούν για να εκφράσουν μια ιδέα ή ένα αντικείμενο στις περισσότερες από τις γλώσσες, και η πολυσημία αναφέρεται στην ιδιότητα μερικών λέξεων να έχουν πολλαπλές, ανεξάρτητες έννοιες. Εάν δε λαμβάνονται υπ' όψιν τα συνώνυμα οδηγούμαστε σε πολύ μικρές, όχι εύκολα διακεκριμένες ομαδοποιήσεις, μερικές από τις οποίες θα έπρεπε να ομαδοποιηθούν μαζί, ενώ η απουσία ελέγχου της πολυσημίας μπορεί να οδηγήσει στη συγκέντρωση ανεξάρτητων εγγράφων.

Στον ***LSI*** αλγόριθμο, τα έγγραφα απεικονίζονται ως διανύσματα. Αντιπροσωπεύεται η σχέση μεταξύ των ιδιοτήτων και των εγγράφων από έναν δισδιάστατο πίνακα Α, διαστάσεων mxn. Οι στήλες του Α αντιπροσωπεύουν τα έγγραφα στη βάση δεδομένων. Έπειτα, υπολογίζουμε τη *μοναδική τιμή αποσύνθεσης* ***(Singular Value Decomposition- SVD)*** του Α, κατόπιν παράγεται ένας τροποποιημένος

πίνακας $A_κ$, από τις κ μεγαλύτερες τιμές $σ_i$, i = 1, 2, 3…κ και τα αντίστοιχα διανύσματά τους:

$$A_κ = Uκ\ Σ_κ\ V^T_κ$$

όπου:

- $Σ_κ$ είναι ένας διαγώνιος πίνακας με μονοτονικά μειωμένα διαγώνια στοιχεία $σ_i$.
- Uκ και $V_κ$ είναι πίνακες των οποίων στήλες είναι τα αριστερά και δεξιά μοναδικά διανύσματα των μεγαλύτερων κ μοναδιαίων τιμών του Α.

Αναλυτικές περιγραφές των τεχνικών γραμμικής άλγεβρας, συμπεριλαμβανομένου και του LSI και των εφαρμογών της στην ανάκτηση πληροφοριών αναφέρονται από τους Berry et al. (1995a) και από τους Letsche και Berry (1997).

Οι στατιστικές προσεγγίσεις που χρησιμοποιούνται στη φυσική γλώσσα και το IR μπορούν πιθανώς να επεκταθούν προς χρήση από τις μηχανές αναζήτησης Ιστού (Crestani et το Al, 1998, Schutze1999).

Διάφοροι επιστήμονες έχουν προτείνει αλγορίθμους ανάκτησης πληροφοριών βασισμένους στην ανάλυση των *δομών αναφοράς σε σελίδες* ***(hyperlink structures)*** (Botafogo et al., 1992, Carriere, Kazman 1997, Chakrabarti et al., 1988a, Chakrabarti et al., 1998b, Frisse, 1988, Kleinberg 1998, Pirolli et al., 1996a και Rivlin et al., 1994).

Ένας απλός τρόπος μέτρησης της ποιότητας της ιστοσελίδας (Carriere και Kazman 1997), είναι ο υπολογισμός του αριθμού των σελίδων οι οποίες έχουν δείκτες στη σελίδα που χρησιμοποιείται στο ***WebQuery*** σύστημα και στη μηχανή αναζήτησης ***Randex***. Μια άλλη μηχανή αναζήτησης που χρησιμοποιεί τις *πληροφορίες των συνδέσεων* ***(link information)*** είναι το Google. Αυτή η πολιτική εφαρμόζεται περισσότερο σε εκπαιδευτικού περιεχομένου και κρατικών sites παρά στα εμπορικά sites. Τον Νοέμβριο 1999, η NorthernLight εισήγαγε ένα νέο σύστημα ταξινόμησης, το οποίο βασιζόταν κατά ένα μέρος στο link data.

Οι δομές αναφοράς σελίδων ***(hyperlink structures)*** χρησιμοποιούνται για την ταξινόμηση των ανακτημένων σελίδων καθώς επίσης μπορούν να χρησιμοποιηθούν και για την ομαδοποίηση των σχετικών σελίδων σε διαφορετικά θέματα- κατηγορίες.

3.3.3.5. Ειδικό Λογισμικό (Robots, Spiders, Crawlers, Agents)

Το *ειδικό λογισμικό* ***(crawler ή spider ή robot ή agent)*** είναι ένα πρόγραμμα που

σκοπό έχει την ανάκτηση ιστοσελίδων, τις περισσότερες φορές για μια μηχανή αναζήτησης. Στις περισσότερες περιπτώσεις το ειδικό λογισμικό αρχίζει με την ηλεκτρονική διεύθυνση ***(URL)*** μιας αρχικής σελίδας P_0. Ανακτά αυτή τη σελίδα P_0, εξάγει οποιεσδήποτε *URLs* σε αυτή, και τις προσθέτει σε μια ουρά από *URLs* που θα ανιχνευθούν. Έπειτα το ειδικό λογισμικό παίρνει *URLs* από την ουρά αναμονής (με κάποια σειρά), και επαναλαμβάνει τη διαδικασία. Κάθε σελίδα που ανιχνεύεται μεταβιβάζεται σε έναν *πελάτη* ***(client)*** ο οποίος σώζει αυτές τις σελίδες, δημιουργεί ένα ευρετήριο, συνοψίζει και αναλύει το περιεχόμενο για αυτές τις σελίδες.

Το ειδικό λογισμικό χρησιμοποιείται ευρέως σήμερα. Το ειδικό λογισμικό των κυριότερων μηχανών αναζήτησης (Altavista, InfoSeek, Excite και Lycos) προσπαθεί να επισκεφτεί τις περισσότερες ιστοσελίδες κειμένων, προκειμένου να δημιουργηθούν τα ευρετήρια περιεχομένων. Άλλο ειδικό λογισμικό επισκέπτεται πολλές σελίδες με σκοπό την εύρεση-αναζήτηση ορισμένων τύπων πληροφοριών (π.χ., *διευθύνσεις ηλεκτρονικού ταχυδρομείου* ***(email addresses)***). Επίσης, υπάρχουν και *προσωπικοί crawlers* ***(personal crawlers)*** που ανιχνεύουν σελίδες που ενδιαφέρουν έναν συγκεκριμένο χρήστη, προκειμένου να δημιουργήσει ένα ευρετήριο γρήγορης πρόσβασης (π.χ. NetAttche, WebSnake).

Το να σχεδιάσει κάποιος ένα καλό ειδικό λογισμικό αντιμετωπίζει πολλές δυσκολίες. Εξωτερικά, το ειδικό λογισμικό πρέπει να αποφεύγει πολυσύχναστες ιστοσελίδες και απασχολημένα δίκτυα καθώς αναζητά ιστοσελίδες. Εσωτερικά, το ειδικό λογισμικό πρέπει να καταφέρνει να αντεπεξέρχεται στις δυσκολίες που δημιουργούν οι τεράστιοι όγκοι των δεδομένων. Εκτός αν έχει απεριόριστους υπολογιστικούς πόρους και απεριόριστο χρόνο, θα πρέπει προσεκτικά να αποφασιστεί ποιες *URLs* θα ανιχνεύσει και με ποια σειρά. Το ειδικό λογισμικό πρέπει επίσης να αποφασίσει πόσο συχνά θα επισκέπτεται σελίδες που έχει ήδη δει, προκειμένου να είναι ενήμερος ο πελάτης για τις αλλαγές στον ιστό. Παρόλα αυτά τα προβλήματα για το σχεδιασμό του ειδικού λογισμικού και τη σημασία αυτών για το διαδίκτυο, δεν έχει πραγματοποιηθεί μεγάλη έρευνα.

Εντούτοις, το ειδικό λογισμικό δεν είναι σε θέση να επισκεφθεί κάθε πιθανή σελίδα του ιστού για δύο κύριους λόγους:

- ✓ Οι πελάτες τους μπορεί να έχουν περιορισμένη αποθηκευτική ικανότητα, και μπορεί να μην είναι σε θέση να συντάξουν ευρετήριο ή να αναλύσουν όλες τις σελίδες. Σήμερα ο ιστός περιέχει πάνω από 1.5TB και αυξάνεται όλο και

γρηγορότερα, έτσι είναι λογικό οι περισσότεροι πελάτες να μην θέλουν ή να μην είναι σε θέση να αντιμετωπίσουν όλα αυτά τα δεδομένα.

- ✓ Το ***crawling*** χρειάζεται χρόνο, έτσι κάποια στιγμή το ειδικό λογισμικό μπορεί να αρχίσει να επισκέπτεται σελίδες που έχουν ανιχνευθεί προηγουμένως, ελέγχοντας για τυχόν αλλαγές. Αυτό σημαίνει ότι δεν μπορεί ποτέ να ανιχνεύσει όλες τις σελίδες. Υπολογίζεται ότι άνω των 600GB του ιστού αλλάζει κάθε μήνα (1997), σήμερα αυτό το μέγεθος έχει αυξηθεί πάρα πολύ.

Σε κάθε περίπτωση, είναι σημαντικό για το ειδικό λογισμικό να επισκέπτεται τις "σημαντικές" σελίδες πρώτα, έτσι ώστε το ποσοστό του ιστού που επισκέπτεται (και είναι ενήμερος) να είναι το σημαντικότερο.

Διαφορές των Robots- Crawlers- Spiders- Agents

Οι μηχανές αναζήτησης έχουν διαφορετικούς τρόπους συλλογής των ιστοσελίδων που περιλαμβάνονται στις βάσεις δεδομένων τους. Οι περισσότερες χρησιμοποιούν το ειδικό λογισμικό που ονομάζεται "spiders-crawlers- robots- Agents" που περιπλανώνται στον ιστό, από *σύνδεση σε σύνδεση* ***(link to link)***, προς αναζήτηση πρόσφατα προστιθέμενων ιστοσελίδων. Επιπλέον, υπάρχουν διάφοροι τρόποι για τους συντάκτες ιστού να καταχωρούν τις σελίδες τους, και στη συνέχεια κάθε μηχανή αναζήτησης υιοθετεί διαφορετικούς αλγορίθμους αναζήτησης για την ανάκτησή τους.

Δεδομένου ότι τα δημοφιλή ευρετήρια των μηχανών αναζήτησης διαφέρουν στο μέγεθος, υπάρχουν ευρετήρια με μόνο 5 εκατομμύρια ή περισσότερα από 50 εκατομμύρια σελίδες. Αλλά καμία μηχανή αναζήτησης δεν ευρετηριάζει ολόκληρο τον ιστό, ακόμη και οι μεγαλύτερες μηχανές αναζήτησης αποκλείουν πάνω από το μισό αυτού.

Εκτός από το πόσο περιεκτικό είναι το ευρετήριό τους (από την άποψη του συνόλου των εγγράφων, το οποίο κάποιος μπορεί να ανακτήσει), οι μηχανές αναζήτησης διαφέρουν και στο πόσο ενημερωμένο είναι αυτό. Μερικές από αυτές τις διαφορές εμφανίζονται λόγω των πολιτικών υποβολής των νέων ιστοσελίδων. Οι περισσότερες μηχανές αναζήτησης επιτρέπουν μια φορά να υποβληθεί η αρχική σελίδα μιας ιστοσελίδας. Μια υποβληθείσα σελίδα μπορεί έπειτα να προστεθεί στο ευρετήριο μιας μηχανής αναζήτησης μετά από μερικά λεπτά ή στην ακραία περίπτωση, αρκετές εβδομάδες αργότερα. Οι σελίδες που συνδέονται με τις υποβληθείσες σελίδες συνήθως δεν προστίθενται αμέσως στο ευρετήριο της μηχανής αναζήτησης. Αντιθέτως,

εξαρτάται από τον δρομολογητή (ειδικό πρόγραμμα) του spider της κάθε μηχανής αναζήτησης, που θα ψάξει για τις ήδη ευρετηριασμένες σελίδες, οι συνδεμένες σελίδες μπορεί να ανιχνευθούν μέσα σε μερικές εβδομάδες ή μερικούς μήνες. Οι πολιτικές πάλι διαφέρουν από μηχανή αναζήτησης σε μηχανή αναζήτησης σχετικά με το πότε μια πρόσφατα ανιχνευμένη σελίδα θα προστεθεί στο ευρετήριο της μηχανής αναζήτησης. (Αυτό μπορεί να εμφανίζεται καθημερινά για μερικές μηχανές ή με μια καθυστέρηση μέχρι ενός μήνα για άλλες.) Οι διαφορές υπάρχουν, επίσης, σχετικά με το βάθος ***(depth)*** στο οποίο οι spiders θα ανατρέξουν για νέες σελίδες. Κάποιες μηχανές αναζήτησης προσπαθούν να αναζητήσουν όλες τις συνδεόμενες σελίδες με ***links,*** άλλες επιλέγουν όλες τις συνδεμένες σελίδες όλων των δημοφιλών sites (μετριούνται με βάση τον αριθμό των εισερχόμενων συνδέσεων), αλλά αποκλείει ορισμένες σελίδες που συνδέονται με λιγότερο δημοφιλή sites (είτε με τον αποκλεισμό των σελίδων που συνδέονται με πάρα πολλά "jumps" από την αρχική σελίδα, είτε περιλαμβάνοντας συνδεμένες σελίδες σε μια δειγματοληπτική βάση).

Δεδομένου ότι ο ιστός είναι τόσο δυναμικός από την άποψη των νέων σελίδων που προστίθενται σε αυτόν (και διαγράφονται, επίσης), οι spiders ξαναεπισκέπτονται τις ήδη ευρετηριασμένες σελίδες με σκοπό να έχουν το ευρετήριό τους όσο το δυνατό περισσότερο ενημερωμένο. Στην πραγματικότητα, μια μηχανή αναζήτησης μπορεί να θεωρηθεί ότι έχει ένα *spider ενημέρωσης* ***(freshness spider)*** και ένα *spider πληρότητας* ***(completeness spider)***. Η πρώτη κατηγορία των spiders μπορεί να θεωρηθεί ότι επισκέπτεται τα δύο πρώτα εκατομμύρια σελίδων της μηχανής αναζήτησης μία φορά την εβδομάδα, και η άλλη κατηγορία ότι επισκέπτεται το υπόλοιπο των ευρετηριασμένων σελίδων περίπου μία φορά το μήνα. Οι συχνότερα επισκεπτόμενες περιοχές περιλαμβάνουν εκείνες που έχουν τις πιο πολλές εισερχόμενες συνδέσεις καθώς επίσης και εκείνες που αλλάζουν το περιεχόμενό τους και οι συνδέσεις τους συχνά (χαρακτηριστικά που οι μηχανές αναζήτησης είναι σε θέση να τα μάθουν).

Γενικά, το πόσο ενημερωμένη είναι η κάθε μηχανή αναζήτησης ποικίλλει αρκετά και όπως αναφέρθηκε παραπάνω εξαρτάται από το ειδικό λογισμικό της. Μια μηχανή αναζήτησης που αφιερώνει τους δικούς της πόρους για την ενημέρωση του ευρετηρίου της και εξερευνά δέκα εκατομμύρια σελίδες ημερησίως μπορεί να χρειαστεί σχεδόν μια εβδομάδα για να επισκεφτεί όλα τα ευρετηριασμένα sites της. Ακόμη και για την ίδια μηχανή αναζήτησης, η πλήρης ενημέρωσή της μπορεί να ποικίλει από λεπτά μέχρι και αρκετούς μήνες, συνήθως είναι συνάρτηση του πόσο δημοφιλής μια ιστοσελίδα θεωρείται.

Επειδή οι μηχανές αναζήτησης διαφέρουν και στην περιεκτικότητά τους και στην ενημέρωσή τους, είναι δύσκολο να προβλεφθεί το ποσοστό της επικάλυψης ανάμεσα στα αποτελέσματα που επιστρέφονται από τις διάφορες μηχανές αναζήτησης που επεξεργάζονται την ίδια πληροφορία.

3.3.3.5.1. Ανασκόπηση της Βιβλιογραφίας

Αρχικά ως ***ARPANET***, το διαδίκτυο χρησιμοποιήθηκε πρώτιστα ως μέσο σύνδεσης από απόσταση και ήταν ένας πειραματισμός για τις τηλεπικοινωνίες. Εντούτοις, κυρίαρχη χρήση του έγινε γρήγορα η επικοινωνία μέσω *ηλεκτρονικού ταχυδρομείου* ***(email)***. Αυτή η τάση συνεχίζεται και στην παρούσα μορφή του διαδικτύου, αλλά με όλο και περισσότερη υποστήριξη στη συνεργάσιμη διανομή δεδομένων και στη διανεμημένη πρόσβαση πληροφοριών των πολυμέσων, ιδιαίτερα με τη χρήση του *World Wide Web* ***(WWW)***. Πολλοί θεωρούν το διαδίκτυο και το *WWW* ως το σημαντικότερο μέσο ταχείας διάδοσης πληροφοριών και ως το παράθυρο στον κυβερνοχώρο.

Ο παγκόσμιος ιστός αναπτύχθηκε αρχικά για την υποστήριξη των επιστημόνων της φυσικής και των μηχανικών στο *Ευρωπαϊκό Εργαστήριο Πυρηνικών Μελετών και Ερευνών* ***(CERN)****, στη Γενεύη της Ελβετίας*. Το 1993, όταν διάφορα προγράμματα *φυλλομετρητών* ***(browser)*** ήταν διαθέσιμα για πρόσβαση σε πληροφορίες, το διαδίκτυο έγινε ο προάγγελος ενός κυβερνοχώρου πλούσιου και με διαφορετικούς τύπους πληροφοριών. Εντούτοις, καθώς οι υπηρεσίες του διαδικτύου βασιζόταν στον παγκόσμιο ιστό γινόταν όλο και περισσότερο πιο δημοφιλείς, ο μεγάλος όμως όγκος των πληροφοριών έχει δημιουργήσει ένα σημαντικό ερευνητικό πρόβλημα. Το παράδειγμα της αναζήτησης πληροφοριών στο διαδίκτυο επιβεβαιώνει αυτό το πρόβλημα, αφού πλέον έχει καταργηθεί η αναζήτηση με βάση το απλό υπερκείμενο όπως *η περιήγηση* ***(browsing)*** (αναζήτηση καθοδηγημένη από τον άνθρωπο για την εξερεύνηση ενός χώρου πληροφοριών) και έχει επικρατήσει η *αναζήτηση με βάση τα περιεχόμενα* (***content-based searching*** -μια διαδικασία στην οποία ο χρήστης περιγράφει μια ερώτηση και ένα σύστημα εντοπίζει τις πληροφορίες που ταιριάζουν στο ερώτημά του). Πολλοί ερευνητές και επαγγελματίες θεωρούν ότι η αναζήτηση στο *internet/intranet* θα είναι ένας από τους σημαντικότερους μελλοντικούς τομείς έρευνας.

Η αναζήτηση στο διαδίκτυο έχει γίνει ένα από τα σημαντικότερα θέματα συζήτησης και έρευνας. Δύο σημαντικές προσεγγίσεις έχουν αναπτυχθεί:

- ✓ *Προγράμματα αναζήτησης βασισμένα στον πελάτη* ***(client-based search spider)***

- ✓ *Άμεση ευρετηρίαση και αναζήτηση σε μια βάση δεδομένων* ***(on line database indexing and searching)***

Ωστόσο μερικά συστήματα περιέχουν και τις δύο προσεγγίσεις.

- ***Προγράμματα αναζήτησης βασισμένα στον πελάτη (client-based search spider)***

Ο ευρέως χρησιμοποιούμενος όρος του *crawler-spider* είναι ένα πρόγραμμα που μπορεί να λειτουργεί αυτόνομα και να ολοκληρώνει κάποιες εργασίες χωρίς άμεση ανθρώπινη επίβλεψη. Η βασική ιδέα της έρευνας των *crawler* είναι να αναπτυχθούν συστήματα λογισμικού που δεσμεύουν και βοηθούν όλους τους τύπους των χρηστών. Τέτοιοι πράκτορες μπορούν να ενεργούν ως *spiders* στο διαδίκτυο και να αναζητούν σχετικές πληροφορίες. Πολλοί ερευνητές έχουν επικεντρωθεί στην ανάπτυξη *κινητών πρακτόρων* ***(mobile agents)*** από μόνες τους. Μερικοί ερευνητές προσπαθούν να εξετάσουν την ερώτηση: "πώς θα ήταν δυνατό να αλληλεπιδρούν οι πράκτορες μεταξύ τους για τη δημιουργία ψηφιακής ομάδας". Άλλοι ερευνητές ενδιαφέρονται περισσότερο για τη σχεδίαση *ευφυών πρακτόρων* ***(intelligent agents)***.

Διάφορα προγράμματα λογισμικού που έχουν αναπτυχθεί είναι βασισμένα στην έννοια των spiders, robots. Τα ***TueMosaic*** και ***WebCrawler*** είναι δύο πρόσφατα παραδείγματα. Και τα δύο χρησιμοποιούν παραλλαγές *της καλύτερης πρώτης τοπικής αναζήτησης στρατηγικής* ***(best first search strategies)***.

Χρησιμοποιώντας τον TueMosaic, οι χρήστες μπορεί να εισαγάγουν λέξεις κλειδιά, να διευκρινίσουν το βάθος ***(depth)*** και το πλάτος ***(width)*** της αναζήτησης των αναφορών σε σελίδες που περιέχουν οι τρέχουσες αρχικές σελίδες που επέδειξαν και ζητούν από τον *spider* να προσκομίσει τις αρχικές σελίδες που συνδέονται με την τρέχουσα αρχική σελίδα. Ο αλγόριθμος αναζήτησης ***Fish*** είναι μια τροποποίηση *της καλύτερης πρώτης μεθόδου αναζήτησης*. Ωστόσο, ενδεχομένως σχετικές αρχικές σελίδες που δεν συνδέουν τις τρέχων ενεργές αρχικές σελίδες δεν θα μπορούν να ανακτηθούν και, όταν το βάθος και το εύρος της αναζήτησης γίνονται μεγάλα (μια εκθετική αναζήτηση), το διάστημα αναζήτησης γίνεται τεράστιο. Το WebCrawler επεκτείνει την έννοια του TueMosaic αρχίζοντας την αναζήτηση χρησιμοποιώντας το δείκτη συνδέσεών του και ακολουθώντας τις συνδέσεις μέσω μιας ευφυούς σειράς. Το Webcrawler αξιολογεί τη σχετικότητα μια αναφοράς σε σελίδα βασισμένο στην ομοιότητά της με το κείμενο της ερώτησης του χρήστη.

Λόγω του πολλαπλασιασμού των σελίδων του www, πολλά νεώτερα προγράμματα *spiders* με διαφορετικές λειτουργίες έχουν αναπτυχθεί. Το ρομπότ

TkWWW αναπτύχθηκε από τη Spetka και χρηματοδοτήθηκε από το Εργαστήριο Πολεμικής Αεροπορίας της Ρώμης (Air Force Rome Laboratory). Τα TkWWW ρομπότ αποστέλλονται από τον TkWWW φυλλομετρητή και σχεδιάστηκαν για την αναζήτηση γειτονικών χώρων του ιστού που βρίσκουν λογικά σχετικές αρχικές σελίδες και επιστρέφουν έναν κατάλογο σημαντικών "συνδέσεων". Εντούτοις, η διαδικασία αναζήτησής τους περιορίζεται σε μια ή δύο τοπικές συνδέσεις από τις αρχικές σελίδες. Τα ρομπότ TkWWW μπορούν επίσης να λειτουργούν στο παρασκήνιο και να δημιουργούν ευρετήριο HTML, να κρατούν στατιστικά στοιχεία του WWW, να δημιουργούν ένα χαρτοφυλάκιο εικόνων. Τα ρομπότ ***WebAnts***, που αναπτύχθηκαν από τη Leavitt στο πανεπιστήμιο του Carnegie Mellon, ερευνούν τη διανομή της συγκεντρωμένης πληροφορίας από ένα σύστημα συνεργαζόμενων *crawlers*. Ο στόχος των WebAnts είναι η δημιουργία συνεργαζόμενων *crawlers* που μοιράζονται τα αποτελέσματα των ερευνών και συντάσσουν ευρετήριο χωρίς επανάληψη της ίδιας διαδικασίας. Το ***RBSE (Repository-Based Software Engineering) spider*** αναπτύχθηκε από τον Eichmann και χρηματοδοτήθηκε από τη NASA. Το RBSE Spider ήταν το πρώτο ρομπότ ευρετηρίασης εγγράφων με βάση το περιεχόμενό του. Έχει τέσσερις τοπικούς μηχανισμούς αναζήτησης:

- ✓ *Πρώτη αναζήτηση εύρους* ***(Breadth first search)*** γνωστής URL
- ✓ *Πρώτη αναζήτηση περιορισμένου βάθους* ***(limited depth first search)*** γνωστής URL
- ✓ *Πρώτη αναζήτηση εύρους* ***(Breadth first search)*** από άγνωστες URLs στη βάση δεδομένων και
- ✓ *Πρώτη αναζήτηση περιορισμένου βάθους* ***(limited depth first search)*** από άγνωστες URLs στη βάση δεδομένων.

➢ ***Άμεση ευρετηρίαση και αναζήτηση σε μια βάση δεδομένων (on line database indexing and searching)***

Μια εναλλακτική προσέγγιση της εύρεσης πόρων του διαδικτύου είναι βασισμένη στην έννοια των βάσεων δεδομένων του ευρετηρίου και της αναζήτησης με βάση τις λέξεις-κλειδιά. Τέτοια συστήματα συλλέγουν πλήρη ή μερικά έγγραφα ιστού και τα αποθηκεύουν στους κεντρικούς υπολογιστές. Αυτά τα έγγραφα αναζητούνται με βάση τις λέξεις-κλειδιά. Οι περισσότερες δημοφιλείς βάσεις δεδομένων του διαδικτύου όπως Lycos, AltaVista, και Yahoo βασίζονται σε μια τέτοια σχεδίαση.

Η Lycos χρησιμοποιεί έναν συνδυασμό από *spiders* για τη συλλογή αρχείων και

για απλή εγγραφή ιδιοκτησίας. Οι κεντρικοί υπολογιστές του διαδικτύου μπορούν να έχουν πρόσβαση στον κεντρικό υπολογιστή της Lycos και η πλήρης εγγραφή εκτελείται με απλά βήματα. Επιπλέον, η Lycos χρησιμοποιεί *spiders* βασισμένους στις συνδέσεις των καταχωρημένων *αρχικών σελίδων* ***(homepages)*** για τον προσδιορισμό άλλων μη καταχωρημένων αρχικών σελίδων. Εφαρμόζοντας αυτές τις τεχνικές, η Lycos έχει αποκτήσει έναν εντυπωσιακό κατάλογο *URLs* του διαδικτύου. Η Lycos υιοθέτησε μια τεχνική ευρετηρίασης με βάση τον τίτλο, τους υποτίτλους, τις 100 πιο σημαντικές λέξεις, τις 20 πρώτες γραμμές, το μέγεθος σε *bytes* και τον αριθμό των λέξεων. Ωστόσο, η επιτυχία της Lycos επίσης επεξηγεί την αδυναμία της προσέγγισης καθώς και το γεγονός της μη δημιουργίας ευφυών και αποδοτικών μηχανών αναζήτησης διαδικτύου. Η δημοτικότητά της έχει προκαλέσει σοβαρή υποβάθμιση της αποδοτικής πρόσβασης πληροφοριών, λόγω της δυσχέρειας της επικοινωνίας και της αναζήτησης επιλεγμένων εγγράφων σε μια βάση δεδομένων των *homepages* του διαδικτύου.

Η AltaVista, αναπτύχθηκε στα ψηφιακά ερευνητικά εργαστήρια (Digital's Research Laboratories) του Palo Alto, συνδυάζει έναν γρήγορο *crawler* ιστού δημιουργώντας ένα μεγάλο ευρετήριο ιστού. Δημοσιοποιήθηκε στις 15 Δεκεμβρίου του 1995 και έχει αναπτυχθεί πολύ γρήγορα σε μια από τις πιο περιεκτικές, αναζητήσιμες βάσεις δεδομένων του διαδικτύου. Παρέχει επίσης ένα ευρετήριο *full-text* που ενημερώνεται σε πραγματικό χρόνο για πάνω από 13.000 ομάδες πληροφοριών. Αν και βασίζεται σε παρόμοιους spider αλγορίθμους τοπικής αναζήτησης, ο *server* AltaVista έγινε δημοφιλής λόγω των ανώτερων *πλατφόρμων υλικού* ***(hardware platforms)*** και του υψηλού εύρους των ζωνών επικοινωνίας.

Αντί να υιοθετήσει την προσέγγιση μιας *βάσης δεδομένων* ***(all-in one database)***, των Lycos και AltaVista, ο κεντρικός υπολογιστής του Yahoo αντιπροσωπεύει μια προσπάθεια να χωριστούν οι πληροφορίες του διαδικτύου σε σημαντικές θεματικές κατηγορίες (π.χ. επιστήμη, ψυχαγωγία, εφαρμοσμένη μηχανική κ.λ.π.). Εντούτοις, οι κατηγορίες αυτές που δημιουργούνται από ανθρώπους είναι περιορισμένες στην τροποποίησή τους καθώς και η διαδικασία δημιουργίας τους είναι χρονοβόρα. Η απαίτηση να είναι οι κατηγορίες ενημερωμένες καθώς επίσης και η απαίτηση ενός ιδιοκτήτη να τοποθετηθεί μια αρχική σελίδα σε μια κατάλληλη κατηγορία έχει μειώσει σημαντικά την επιτυχία και τη δημοτικότητα του Yahoo.

3.3.3.5.2. Αλγόριθμοι Spidering

Ο ιστός έχει πληθώρα χρήσιμων πόρων, αλλά η δυναμική του και η μη-δομημένη

φύση του κάνει δύσκολο τον εντοπισμό τους. Οι μηχανές αναζήτησης μας βοηθούν, αλλά ο αριθμός των ιστοσελίδων υπερβαίνει τα δύο δισεκατομμύρια, καθιστώντας δύσκολο για μια γενική μηχανή αναζήτησης να διατηρεί περιεκτικό και ενημερωμένο ευρετήριο. Επιπλέον, καθώς ο ιστός μεγαλώνει, τόσο περισσότερες απαντήσεις λαμβάνει ο χρήστης στο ερώτημά του. Μια γενικής χρήσης μηχανή αναζήτησης, όπως η Google ή η AltaVista, παράγει συνήθως χιλιάδες *αναφορές* ***(hits)***, πολλές από τις οποίες είναι άσχετες με την ερώτηση του χρήστη.

Οι περισσότεροι *spiders* χρησιμοποιούν απλούς αλγορίθμους αναζήτησης γραφικών παραστάσεων, όπως η *πρώτη αναζήτηση εύρους* ***(breadth-first search)***, για να συλλέξει ιστοσελίδες. Χωρίς έλεγχο, οι *spiders* θα προσκομίσουν σελίδες οποιουδήποτε θέματος. Υπάρχουν δύο δημοφιλείς τρόποι που ελέγχουν τη σχετικότητα και την ποιότητα των σελίδων που θα συλλέξει το *spider*:

• ***Περιορισμός των spider σε συγκεκριμένες περιοχές ιστού:*** Παραδείγματος χάριν, οι περισσότερες ιστοσελίδες που περιέχουν www.toyota.com θα ήταν σχετικές με σελίδες που αναφέρονται σε αυτοκίνητα.

• ***Φιλτράρισμα των σελίδων με βάση το περιεχόμενό τους:*** Παραδείγματος χάριν, ένα πρόγραμμα θα μπορούσε να αφαιρεί τις σελίδες που έχουν μικρότερο αριθμό σχετικών λέξεων-κλειδιά από ένα κατώτατο όριο.

Και οι δύο προσεγγίσεις έχουν μειονεκτήματα. Περιορίζοντας τις περιοχές πρόσβασης ενδεχομένως σχετικές ιστοσελίδες να μην συλλέγονται που είναι έξω από αυτή την περιοχή. Επίσης δεν λειτουργεί για ιστοσελίδες που έχουν διαφορετικό περιεχόμενο. Από την άλλη πλευρά φιλτράροντας τις σελίδες που συλλέχθηκαν, για μια πλήρης αναζήτηση ιστού αυτή η προσέγγιση είναι ανεπαρκής.

Οι καλοί *spidering* αλγόριθμοι μπορούν να βελτιώσουν την ακρίβεια των αποτελεσμάτων αναζήτησης, με την πρόβλεψη ότι μια *URL* δείχνει σχετική ιστοσελίδα πριν γίνει ***download*** στην τοπική βάση δεδομένων της μηχανής αναζήτησης. Τέτοιες προβλέψεις εξαρτώνται από το περιεχόμενο του ιστού και της δομής του, με τρόπο κατανοητό στις μηχανές αναζήτησης. Θα αναφερθούμε σε δύο κατηγορίες:

- *Βασισμένες στο περιεχόμενο* ***(Content-Based Web Analysis)***
- *Βασισμένες στις αναφορές* ***(Link-Based Web Analysis)***

- ***Βασισμένες στο περιεχόμενο (Content-Based Web Analysis)***

Οι *spiders* μπορούν να εφαρμόσουν τεχνικές ευρετηρίασης για ανάλυση κειμένων

και εξαγωγής λέξεων-κλειδιών για τον καθορισμό της σχετικότητας του περιεχομένου μιας σελίδας. Μπορούν να ενσωματώσουν τη γνώση των περιοχών ιστού στην ανάλυσή τους για βελτίωση των αποτελεσμάτων. Παραδείγματος χάριν, μπορούν να ελέγξουν τις λέξεις της ιστοσελίδας σε σχέση ενάντια με έναν κατάλογο ορολογίας και δίνουν μεγαλύτερο βάρος στις σελίδες που περιέχουν τις λέξεις του καταλόγου. Δίνουν μεγαλύτερο βάρος σε λέξεις και φράσεις που βρίσκονται στον τίτλο ή στις κεφαλίδες που είναι μια συνηθισμένη πρακτική ανάκτησης πληροφοριών στην οποία οι spiders μπορούν να εφαρμόσουν βασισμένοι στα κατάλληλα HTML tags.

Η διεύθυνση *URL* περιέχει συχνά χρήσιμες πληροφορίες για μια σελίδα. Παραδείγματος χάριν, http://ourworld.compuserve.com/homepages/LungCancer/ δείχνει ότι η σελίδα προέρχεται από τον compuserve.com τομέα και ότι πιθανόν σχετίζεται με πληροφορίες για τον καρκίνο των πνευμόνων.

Ένα *ευφυές spider* θα εξετάσει τις σελίδες από έναν τομέα .gov που είναι πιο επιτακτικό από τις σελίδες ενός τομέα .com. Μερικές τεχνικές θεωρούν *URLs* με λιγότερες καθέτους πιο χρήσιμες από εκείνες με περισσότερες καθέτους.

- ***Βασισμένες στις αναφορές (Link-Based Web Analysis)***

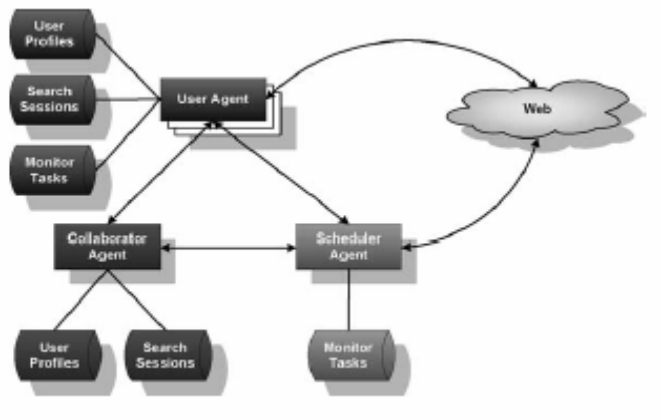
Η πρόσφατη έρευνα έχει χρησιμοποιήσει τη *δομή των συνδέσεων ιστού* ***(web link structure)*** για την εξαγωγή σημαντικών πληροφοριών για τις σελίδες. Διαισθητικά, ο συντάκτης μιας ιστοσελίδας Α, ο οποίος τοποθετεί μια αναφορά σε μια ιστοσελίδας Β, θεωρεί ότι η Β είναι σχετική με την Α. Ο όρος ***in-links*** αναφέρεται στις συνδέσεις υπερ-κειμένου που καταλήγουν στη σελίδα. Συνήθως, όσο μεγαλύτερος είναι ο αριθμός των *in-links*, τόσο υψηλότερα ένα *spider* θα εκτιμήσει τη σελίδα. Η λογική είναι παρόμοια με την ανάλυση παραπομπής, στο οποίο ένα συχνά-αναφερόμενο άρθρο εξετάζεται καλύτερα από ένα που δεν αναφέρεται ποτέ.

Το ***anchor text*** είναι η λέξη ή η φράση που υπερ-συνδέεται με μια επιθυμητή σελίδα. Το ***anchor text*** μπορεί να παρέχει μια καλή πηγή πληροφοριών για μια επιθυμητή σελίδα επειδή αντιπροσωπεύει πως οι άνθρωποι συνδέονται με αυτή τη σελίδα, στην πραγματικότητα την περιγράφει. Διάφορες μελέτες έχουν προσπαθήσει να χρησιμοποιήσουν είτε το *anchor text* ή το κείμενο δίπλα σε αυτό για να προβλέψει το περιεχόμενο της σελίδας. Στο παράρτημα Δ υπάρχουν μερικοί αλγόριθμοι του ειδικού λογισμικού.

3.3.3.5.3. Παραδείγματα Αρχιτεκτονικής Ειδικού Λογισμικού

Στη βιβλιογραφία υπάρχουν πολλοί διαφορετικοί τύποι ειδικού λογισμικού (crawlers, spiders και agents). Γίνεται αναφορά για την αρχιτεκτονική δύο τύπων ειδικού λογισμικού, για το ***collaborative spider*** και για τους ***InfoSpiders***.

Collaborative Spider



Σχήμα 3.9.: Αρχιτεκτονική του Collaborative Spider (Michael Chau et all., 2002)

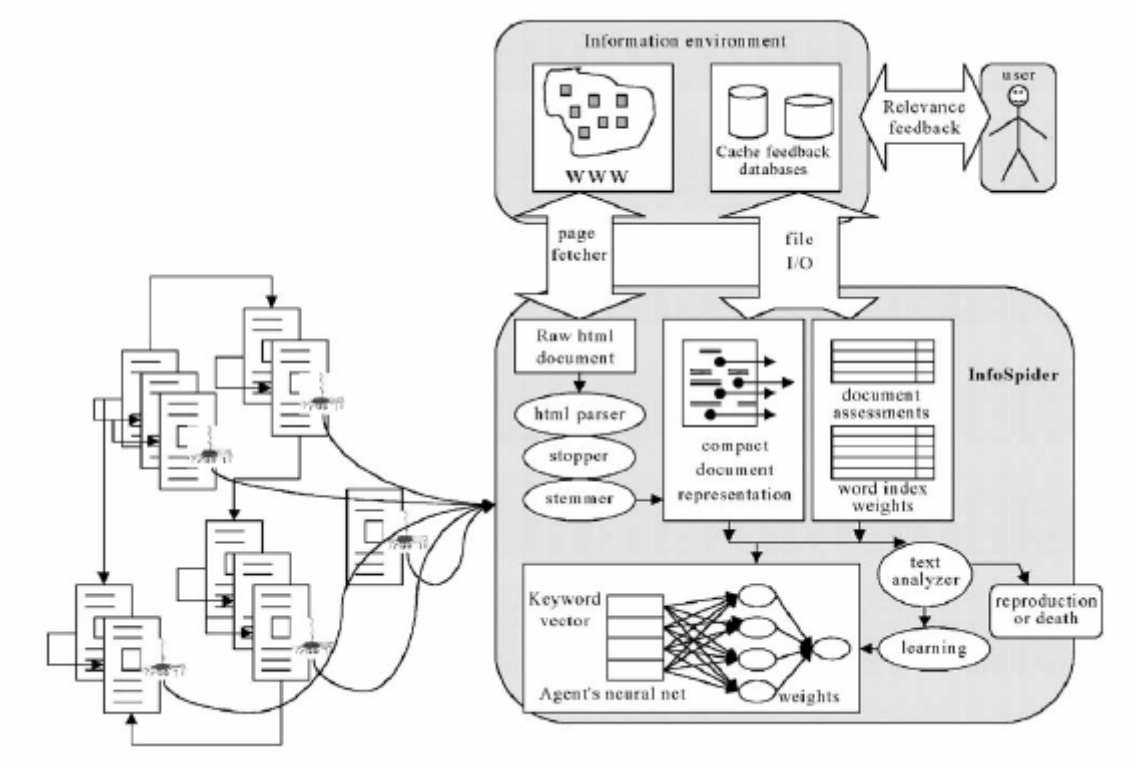
Το collaborative spider (Συνεργάσιμο spider) (Σχήμα 3.9.) αποτελείται από τρεις τύπους *spider* λογισμικού, οι οποίοι είναι:

- ✓ *Πράκτορας χρηστών* ***(User Agent)***,
- ✓ *Συνεργάσιμος πράκτορας* ***(Collaborator Agent)***, και
- ✓ *Πράκτορας δρομολογητής* ***(Scheduler Agent)***.

Ο *User Agent* είναι υπεύθυνος για την ανάκτηση σελίδων από τον ιστό, και για την αλληλεπίδρασή του με τους χρήστες. Ο *Collaborator Agent* διευκολύνει τη διανομή της πληροφορίας μεταξύ των διαφορετικών *User Agents*. Ο *Scheduler Agent* κρατά μια λίστα διεργασιών και είναι υπεύθυνος για την εκτέλεση αυτών των διεργασιών οι οποίες είναι βασισμένες σε διεργασίες χρηστών.

InfoSpiders

Το InfoSpiders είναι ένα εξελικτικό *πολυπρακτορικό σύστημα* ***(multiagent system)*** στο οποίο κάθε πράκτορας προσαρμόζεται σε ένα τοπικό περιβάλλον πληροφοριών. Το Σχήμα 3.10. παρουσιάζει κάθε πράκτορα *InfoSpider*. Ο πράκτορας αλληλεπιδρά με το περιβάλλον πληροφοριών που αποτελείται από την πραγματική συλλογή δικτύου (ιστός) συν τις πληροφορίες που κρατιούνται στις τοπικές δομές δεδομένων.



Σχήμα 3.10.: Αρχιτεκτονική του InfoSpiders (Filippo Menczer, 2002)

Η προσαρμοσμένη αναπαράσταση των *InfoSpiders* κατά προσέγγιση αποτελείται από έναν κατάλογο λέξεων-κλειδιών, που αρχικοποιείται με τους όρους της ερώτησης, και ενός feed-forward νευρωνικού δικτύου. Οι λέξεις-κλειδιά αντιπροσωπεύουν την άποψη ενός πράκτορα για το ποιοι όροι καλύτερα διακρίνουν τα σχετικά έγγραφα με του χρήστη το ερώτημα. Το νευρωνικό δίκτυο χρησιμοποιείται για να υπολογίζει τις συνδέσεις.

Τα *InfoSpiders* στηρίζονται είτε στις παραδοσιακές μηχανές αναζήτησης είτε σε ένα σύνολο *προσωπικών σελιδοδεικτών* ***(personal bookmarks)*** προκειμένου να ληφθεί ένα σύνολο URLs που αναφέρονται σε σελίδες υποθετικά σχετικές με την ερώτηση που διατυπώνεται από τον χρήστη.

Σε κάθε βήμα, κάθε πράκτορας αναλύει το κείμενο του εγγράφου στο οποίο βρίσκεται, για να υπολογίσει τη σχετικότητα της γειτονικής πληροφορίας, που δίνεται από τους εξερχόμενους συνδέσμους της τρέχουσας σελίδας. Πιο απλά, ένας πράκτορας υπολογίζει κάθε εξερχόμενη σύνδεση κοιτάζοντας την ύπαρξη των όρων της ερώτησης στην περιοχή κοντά στη σύνδεση. Ο πράκτορας χρησιμοποιεί έπειτα εκτιμήσεις της σχετικότητας των συνδέσεων για την επιλογή της επίσκεψης του επόμενου εγγράφου.

3.3.4. Πλεονεκτήματα και Μειονεκτήματα των Μηχανών Αναζήτησης

❖ ***Όγκος του υλικού που καλύπτουν***

Το κυριότερο πλεονέκτημα των μηχανών αναζήτησης είναι η εκτεταμένη κάλυψη του δικτύου. Οι *crawlers* ή οι *spiders* αναζητούν και δημιουργούν ευρετήρια από καταχωρήσεις του δικτύου τακτικά και ενημερώνουν τις βάσεις δεδομένων με τυχόντα νέα *sites* καθώς και με τυχόν αλλαγές σε υπάρχουσες ιστοσελίδες. Ένα άλλο θετικό στοιχείο έχει να κάνει με τη χρήση της κατάταξης βάσει της *σχετικότητας των αποτελεσμάτων με τους όρους της αναζήτησης* ***(relevance ranking)*** και διαφόρων άλλων χαρακτηριστικών διευκόλυνσης της ανάκτησης. Για αυτούς τους λόγους οι μηχανές αναζήτησης είναι πολύτιμες αν κάποιος ενδιαφέρεται για εκτεταμένη κάλυψη του δικτύου.

Ωστόσο, το κυριότερο πλεονέκτημα των μηχανών αναζήτησης μπορεί να είναι και μειονέκτημα. Ο τεράστιος αριθμός των *αποτελεσμάτων* ***(hits)*** για οποιαδήποτε αναζήτηση απαιτεί χρόνο προκειμένου να εξεταστούν όλα τα αποτελέσματα και αυτό το πρόβλημα αυξάνεται με την έλλειψη επεξηγηματικής πληροφορίας, σχετικής με το ανακτώμενο υλικό.

❖ ***Έλλειψη διάκρισης της ποιότητας της πληροφορίας***

Ένα από τα πιο σημαντικά μειονεκτήματα των μηχανών αναζήτησης είναι ότι δεν διαχωρίζουν το υλικό που υπάρχει στη βάση δεδομένων τους με κριτήριο την ποιότητα, δεδομένου ότι οποιοσδήποτε, οπουδήποτε μπορεί να διαχύσει πληροφορία μέσω του δικτύου, με την προϋπόθεση ότι έχει πρόσβαση στο απαιτούμενο *software* και *hardware*. Επειδή οι βάσεις δεδομένων των μηχανών αναζήτησης δημιουργούνται αυτόματα, τα αποτελέσματα οποιασδήποτε αναζήτησης μπορούν να προέρχονται από οποιαδήποτε πηγή, είτε είναι επίσημη και αξιόπιστη είτε είναι προκατειλημμένη και παραπλανητική. Συνεπώς, ένα *site* μπορεί να περιλαμβάνει θεματικούς όρους που ενδιαφέρουν το χρήστη, αλλά η πληροφορία μπορεί να είναι ανεπίκαιρη, ανακριβής ή ανεπίσημη, ή ένα *site* μπορεί να παίζει σχεδόν το ρόλο ενός εξωφύλλου, χωρίς να παρέχει λεπτομερή πληροφορία.

Ένα άλλο σχετικό με τις μηχανές αναζήτησης πρόβλημα είναι ότι δεν υπάρχει διάκριση όσον αφορά τα θέματα που καλύπτονται και ότι αναζητούν κάθε λέξη κάθε

σελίδας που ανταποκρίνεται στο *query*. Οπότε, για παράδειγμα μία αναζήτηση για τη λέξη *'virus'* θα ανακτούσε υλικό τόσο για ιούς σχετικούς με υπολογιστές όσο και για βιολογικούς ιούς. Ο μηχανισμός κατάταξης της σχετικότητας των *sites* απλά ιεραρχεί σε μία λίστα εκείνα τα αποτελέσματα των οποίων οι όροι αναζήτησης εμφανίζονται συχνότερα, άσχετα με το θέμα. Αν και διάφορες μηχανές αναζήτησης προσφέρουν ένα αυξανόμενο εύρος επιλογών *βελτιωμένων αναζητήσεων* ***(refining search)***, όπως π.χ. μέσω 'advanced' ή 'power' σελίδων αναζήτησης, παραμένει το πρόβλημα ότι τα αποτελέσματα δημιουργούνται αυτόματα και επομένως για άλλη μία φορά δεν υπάρχει εγγύηση της ποιότητας ή της σχετικότητας τους.

❖ ***Αποκλεισμός ορισμένου τύπου υλικού***

Περαιτέρω προβλήματα της χρήσης των μηχανών αναζήτησης σχετίζονται με το αποκλεισμό του υλικού από τη βάση δεδομένων τους. Αν και οι μηχανές αναζήτησης αξιώνουν κάλυψη τεράστιων ποσοτήτων ιστοσελίδων, στην πραγματικότητα καλύπτουν μόνο ένα μικρό ποσοστό του πιθανολογούμενου χρήσιμου υλικού που είναι πλέον διαθέσιμο μέσω του διαδικτύου.

Οι μηχανές αναζήτησης συνήθως περιορίζονται σε βασισμένο στο δίκτυο ***(web-based)*** υλικό - δημιουργείται αυτόματα χρησιμοποιώντας software που εξερευνά το δίκτυο - έτσι ώστε αν το υλικό δεν περιλαμβάνεται σε μία ιστοσελίδα, το *software* δεν μπορεί να αποκτήσει πρόσβαση και να το καταχωρήσει στο ευρετήριο. Προκειμένου για παράδειγμα να εντοπιστούν μηνύματα από newsgroups, απαιτείται γενικά ένα ξεχωριστό μέσον αναζήτησης. Κατά τον ίδιο τρόπο, αν μία *full-text* αναφορά ή ένα άρθρο δεν είναι σε HTML. μορφή, οι μηχανές αναζήτησης έχουν λιγότερες πιθανότητες να το συμπεριλάβουν. Για παράδειγμα, πολλά άρθρα εφημερίδων δημοσιεύονται σε ***Portable Document Format (PDF)***. Αυτό επιτρέπει στα έγγραφα να τυπώνονται σε μία μορφή η οποία είναι πανομοιότυπη με την πρωτότυπη έκδοση και η οποία επιτρέπει τη χρήση περισσότερων δυνατοτήτων επεξεργασίας κειμένου και γραφικών από αυτές που υπάρχουν χρησιμοποιώντας HTML. Ένα ξεχωριστό *software*, το *Adobe Acrobat*, απαιτείται προκειμένου να υπάρχει πρόσβαση σ' αυτή τη μορφή. Συνεπώς, οι περισσότερες μηχανές αναζήτησης δεν μπορούν να καταχωρήσουν στα ευρετήρια την πληροφορία και απαιτείται ένα ξεχωριστό εργαλείο. Εξαιρείται το Google, το οποίο περιλαμβάνει PDF έγγραφα στα αποτελέσματα του.

Επίσης, οι μηχανές αναζήτησης περιορίζονται στο λεγόμενο *«Ορατό Δίκτυο»*

'Publicly Indexable Web'. Το *«Αόρατο Δίκτυο»* ***'Invisible Web'*** αναφέρεται σε πληροφορία που οι μηχανές αναζήτησης δεν μπορούν να καταχωρήσουν στο ευρετήριο επειδή είτε είναι κρυμμένη πίσω από την επιφάνεια αναζήτησης, όπως σε μία βάση δεδομένων, είτε είναι πίσω από μία login screen, όπως τα άρθρα σε ηλεκτρονικά περιοδικά. Οι εκτιμήσεις αναφέρουν ότι το «Αόρατο Δίκτυο» είναι 500 φορές πιο μεγάλο από το «ορατό» (Cooke, 1999). Για τους ακαδημαϊκούς χρήστες του διαδικτύου ιδιαίτερα, ο αποκλεισμός πληροφορίας από βάσεις δεδομένων και ηλεκτρονικά περιοδικά σημαίνει πιθανό αποκλεισμό ακριβούς και υψηλής ποιότητας πληροφορίας, που μάλλον τους ενδιαφέρει. Από την άλλη, ξεχωριστά εργαλεία αναπτύσσονται για την δυνατότητα πρόσβασης σ' αυτού του είδους το υλικό.

3.4. Μηχανές Πολλαπλής Αναζήτηση (Multi or Meta-Search Engine)

Οι *Μηχανές Πολλαπλής Αναζήτησης* ***(Meta-Search Engines),*** οι οποίες ονομάζονται και meta-crawlers ή multi-search engines, δίνουν τη δυνατότητα ταυτόχρονης αναζήτησης σε διάφορα εργαλεία αναζήτησης και όχι σε μία απλή βάση δεδομένων. Στη συνέχεια τα αποτελέσματα παρουσιάζονται συνήθως μαζί σε μία σελίδα.

Ως παραλλαγή στις μηχανές αναζήτησης, που περιγράφτηκαν στην προηγούμενη ενότητα, οι πολλαπλές ή συνδυαστικές μηχανές αναζήτησης αποτελούν ένα συνδυαστικό εργαλείο πολλών μηχανών αναζήτησης, όπου η έρευνα διεξάγεται παράλληλα, δίνοντας έτσι τη δυνατότητα στο χρήστη με μια ερώτηση να διεξάγει την αναζήτηση του ταυτόχρονα σε περισσότερες από μία μηχανές αναζήτησης. Στις γνωστότερες meta-search μηχανές αναζήτησης συγκαταλέγονται οι: All-In-One, Highway61, Mamma, Metacrawler, Metasearch, OneSeek, Ixquick, SurfWax, Dogpile, ProFusion κ.α.

Τα περισσότερα εργαλεία αναζήτησης προκειμένου να διευκολύνουν το χρήστη, παρέχουν τη δυνατότητα για εκμετάλλευση τόσο των θεματικών καταλόγων όσο και των μηχανών αναζήτησης, καθώς και συνδυασμού τους.

3.4.1. Βασικά Χαρακτηριστικά

Ο μηχανισμός μιας μηχανής meta-search στέλνει το ερώτημα του χρήστη σε πολλές μηχανές ταυτόχρονα. Κατόπιν ανακτά τα αποτελέσματα που βρέθηκαν και, αφού αφαιρέσει τις διπλές εγγραφές, παρουσιάζει στο χρήστη τα αποτελέσματα που βρέθηκαν σε όλα τα ευρετήρια που χρησιμοποιήθηκαν. Τα πλεονεκτήματα είναι σημαντικά. Ο χρήστης κερδίζει χρόνο εισάγοντας μια φορά το ερώτημα του σε μια μηχανή πολλαπλής αναζήτησης. Ακόμη, εξοικονομεί το χρόνο που θα απαιτούσε μια στοιχειώδης αξιολόγηση των μηχανών αυτών, στην περίπτωση που ήθελε να επιλέξει μια από αυτές. Τέλος, επιτυγχάνει καλύτερη κάλυψη, αφού τα ευρετήρια που ερευνούνται δεν ανήκουν σε μία αλλά σε δύο ή περισσότερες βάσεις δεδομένων.

Στον αντίποδα, οι εν λόγω μηχανές είναι τις περισσότερες φορές αργές στην αναζήτηση και στην παρουσίαση των αποτελεσμάτων -κάτι δικαιολογημένο, αφού ερευνούν σε πολλαπλάσιο αριθμό σελίδων και όρων που έχουν καταχωρηθεί σε ευρετήρια- ενώ δεν προσφέρουν τις ευκολίες και την παραμετροποίηση ***(customization)***

της έρευνας που μπορεί να βρει ο χρήστης σε μια μεμονωμένη μηχανή αναζήτησης.

Είναι προφανές ότι όπως και στις «απλές» μηχανές αναζήτησης έτσι και στις meta-search engines είναι δύσκολο να υπάρξει ομοιογένεια σε ότι αφορά τα διατιθέμενα χαρακτηριστικά. Γενικά, υπάρχουν ορισμένα χαρακτηριστικά που θα συναντήσει κάποιος στις περισσότερες από αυτές τις μηχανές και που τείνουν να καθιερωθούν και στις υπόλοιπες. Τα πιο σημαντικά χαρακτηριστικά που διαθέτουν ή θα έπρεπε να διαθέτουν οι *meta-search engines* είναι (Κωνσταντινίδης, 2000):

- *Εύρος αναζήτησης* ***(full-text indexing, phrase search)***
- *Τελεστές Boolean*
- *Τελεστές αποκοπής και εγγύτητας* ***(truncation and proximity operators)***
- *Παρουσίαση αποτελεσμάτων*
- *Δυνατότητα εντοπισμού και απάλειψης των διπλών εγγραφών*
- *Δυνατότητα ενοποίησης των αποτελεσμάτων από τις διαφορετικές μηχανές*
- *Δυνατότητα βελτίωσης αποτελεσμάτων* ***(refine search)***
- *Δυνατότητες περιορισμού*
- *Οδηγίες χρήσης και βοήθειας*

Σε γενικές γραμμές οι συνδυαστικές μηχανές αναζήτησης έχουν τρεις παραλλαγές στον τρόπο με τον οποίο λειτουργούν. Μια meta-search engine μπορεί να αναζητά τους ζητούμενους όρους ενεργοποιώντας τη μία μηχανή μετά την άλλη, ενεργοποιώντας όλες τις μηχανές ταυτόχρονα ή, τέλος, παραθέτοντας απλά τη λίστα με όλες τις διαθέσιμες μηχανές και τα εργαλεία αναζήτησης, προκειμένου ο χρήστης να διαμορφώσει το ερώτημα του σε όποια από αυτές τελικά επιλέξει. Καθεμία από τις εκδοχές αυτές, όπως είναι φυσικό, έχει τα πλεονεκτήματα και τα μειονεκτήματα της.

ΣΕΙΡΙΑΚΗ ΑΝΑΖΗΤΗΣΗ (SERIAL SEARCH)

Η προσέγγιση αυτή είναι χαρακτηριστική των περισσότερων συνδυαστικών μηχανών αναζήτησης. Ο χρήστης εισάγει στο μοναδικό πεδίο αναζήτησης που υπάρχει το ερώτημα του και κατόπιν, ενεργοποιεί είτε με pull down menus είτε με *«κουτιά ελέγχου»* ***(check boxes)*** τις μηχανές εκείνες, στις οποίες θέλει να απευθύνει το ερώτημα και να ξεκινήσει την αναζήτηση.

Τα αποτελέσματα παρουσιάζονται συνήθως ανά μηχανή αναζήτησης, δίνοντας έτσι την ευκαιρία στο χρήστη να εξακριβώσει με μια ματιά τη μηχανή με τα καλύτερα και περισσότερα αποτελέσματα και πολύ εύκολα να κάνει τις συγκρίσεις του. Το

βασικό μειονέκτημα σε αυτού του είδους τα συστήματα είναι, ότι καθώς η αναζήτηση είναι σειριακή, η μία μηχανή ενεργοποιείται μετά την άλλη και μόνο όταν και η τελευταία επιστρέψει τα αποτελέσματα της παρουσιάζεται το σύνολο των αποτελεσμάτων. Έτσι, μπορεί ο χρήστης να εισάγει μονάχα μια φορά το ερώτημα του, ωστόσο χάνει πολύτιμο χρόνο περιμένοντας όλες τις μηχανές να παρουσιάσουν τα αποτελέσματα τους.

ΤΑΥΤΟΧΡΟΝΗ ΑΝΑΖΗΤΗΣΗ (SIMULTANEOUS SEARCH)

Η δεύτερη προσέγγιση αφορά την ταυτόχρονη αναζήτηση του ερωτήματος που έθεσε ο χρήστης από όλες τις μηχανές. Η προσέγγιση αυτή ουσιαστικά δε διαφέρει από την προηγούμενη παρά μόνο στο ότι τα αποτελέσματα παρουσιάζονται τη στιγμή που ανακτώνται από κάθε μηχανή. Έτσι ο χρήστης εξοικονομεί πολύτιμο χρόνο, αφού δε χρειάζεται να περιμένει να ολοκληρωθούν η αναζήτηση και η παρουσίαση των αποτελεσμάτων από το σύνολο των μηχανών.

ΣΥΓΚΕΝΤΡΩΣΗ ΣΥΣΤΗΜΑΤΩΝ ΑΝΑΖΗΤΗΣΗΣ

Τέλος, η τρίτη προσέγγιση είναι αυτή κατά την οποία ένα site συγκεντρώνει μια σειρά από μηχανές και εργαλεία αναζήτησης στα οποία ο χρήστης μπορεί να εισάγει το ερώτημα του, έχοντας βέβαια υπόψη ότι όσες από τις διαθέσιμες μηχανές θέλει να χρησιμοποιήσει τόσες φορές θα πρέπει να εισάγει το ερώτημα του σε καθεμία από αυτές. Τα πλεονεκτήματα της προσέγγισης αυτής είναι ότι ουσιαστικά ο χρήστης απαλλάσσεται από την ανάγκη να μεταβαίνει από το ένα *site* στο άλλο προκειμένου να χρησιμοποιήσει όλες αυτές τις μηχανές, και επιπλέον υπάρχει περίπτωση να ανακαλύψει μια μηχανή που δε θα είχε σκεφτεί να χρησιμοποιήσει σε άλλη περίπτωση. Βασικό μειονέκτημα αποτελεί το ότι στην ουσία τέτοια *sites* δεν συνιστούν meta-search engine αλλά μια απλή συγκέντρωση συστημάτων αναζήτησης, προκειμένου ο χρήστης να τα βρίσκει όλα συγκεντρωμένα.

Η παρουσίαση των αποτελεσμάτων της αναζήτησης ποικίλει από μηχανή σε μηχανή αναζήτησης. Υπάρχουν τρόποι παρουσίασης που θα αναφερθούν αναλυτικά στο Κεφάλαιο 5 της παρούσας εργασίας για κάθε μηχανή πολλαπλής αναζήτησης ξεχωριστά.

Προφανώς η χρήση των μηχανών meta-search γίνεται ολοένα πιο επιτακτική λόγω της ιλιγγιώδους αύξησης των περιεχομένων του *World Wide Web* και της αδυναμίας μιας και μοναδικής μηχανής αναζήτησης προκειμένου να καλύψει το εύρος

αυτό. Είναι βέβαιο ωστόσο, ότι υπάρχουν σημαντικά περιθώρια βελτίωσης τόσο στο όσο και στα ιδιαίτερα χαρακτηριστικά που προαναφέρθηκαν.

3.4.2. Πλεονεκτήματα και Μειονεκτήματα των Meta-Search Engines

Υπάρχουν πολλές διαθέσιμες meta-search engines και κάθε μία λειτουργεί λίγο διαφορετικά, με αποτέλεσμα να έχει διαφορετικά πλεονεκτήματα και μειονεκτήματα. Το BigHub.com επιτρέπει τη διεξαγωγή έρευνας σε πολλαπλές βάσεις δεδομένων. Επομένως, αυτή η προσέγγιση μπορεί να είναι πολύτιμη εάν κάποιος επιθυμεί εκτεταμένη κάλυψη του Δικτύου. Άλλα εργαλεία συνδέουν το χρήστη αυτόματα με ένα εύρος διαφορετικών εργαλείων αναζήτησης. Το κυριότερο πλεονέκτημά τους είναι ότι προτρέπουν το χρήστη να χρησιμοποιήσει εργαλεία, που σε άλλη περίπτωση δε θα τα λάμβανε καν υπόψη του.

Ωστόσο, υπάρχουν πολλά προβλήματα όσον αφορά τη χρήση μέσων πολλαπλής αναζήτησης. Συγκεκριμένα, αυτό που ο χρήστης χρησιμοποιεί είναι μία συλλογή μηχανών αναζήτησης. Συνεπώς, τα μειονεκτήματα που περιγράφηκαν όσον αφορά τη χρήση των μηχανών αναζήτησης για τον εντοπισμό ποιοτικής πληροφορίας ισχύουν και εδώ. Δηλαδή:

- ✓ *Ο μεγάλος αριθμός των αποτελεσμάτων για κάθε αναζήτηση*
- ✓ *Η έλλειψη επεξηγηματικής πληροφορίας σχετικά με το ανακτώμενο υλικό*
- ✓ *Ο αρκετός απαιτούμενος χρόνος για την εξέταση των αποτελεσμάτων*
- ✓ *Ο περιορισμός στο Web-based υλικό και στο «ορατό Δίκτυο»*
- ✓ *Η έλλειψη διάκρισης του υλικού όσον αφορά την ποιότητα του.*

Εξάλλου, οι meta-search engines δεν επιτρέπουν στο χρήστη την πρόσβαση στις επιλογές αναζήτησης (boolean κ.λ.π.) της κάθε μηχανής ξεχωριστά.

3.5. Θεματικοί Κατάλογοι και Θεματικά Ευρετήρια (Subject Catalogues- Subject Directories)

3.5.1. Ορισμός και Παραλλαγές

Τα *Θεματικά Ευρετήρια* ***(Subject Catalogues- Subject Directories)*** είναι τα εργαλεία εκείνα που παρέχουν μια ιεραρχικά οργανωμένη ταξινόμηση της ανθρώπινης γνώσης σε κατηγορίες, προκειμένου η αναζήτηση στα περιεχόμενα τους να γίνεται με βάση το θέμα. Συχνά ονομάζονται και *δέντρα* ***(trees)*** λόγω της δομής τους, η οποία ξεκινά από ένα βασικό κορμό-κατηγορία και σιγά-σιγά διακλαδώνεται σε υποκατηγορίες. Στα θεματικά ευρετήρια, η ιεραρχική δομή επιτρέπει την αναζήτηση από το γενικότερο όρο-κατηγορία του θέματος προς τον ειδικότερο όρο-κατηγορία. Έτσι ο χρήστης προχωρώντας σταδιακά στις ειδικότερες υποκατηγορίες, περιορίζει και εξειδικεύει την αναζήτηση και κατά συνέπεια την πληροφορία που τελικά επιθυμεί να ανακτήσει. Σε καθεμία από αυτές τις κατηγορίες υπάρχουν *συνδεόμενες ιστοσελίδες* ***(linked pages)***, οι οποίες με τη σειρά τους παραπέμπουν σε άλλες ιστοσελίδες ή έγγραφα. Σε πολλές περιπτώσεις οι συνδεόμενες σελίδες διαθέτουν το δικό τους μηχανισμό αναζήτησης, αφού και οι ίδιες μπορεί να συγκεντρώνουν σημαντικό αριθμό εγγράφων στα περιεχόμενα τους.

Όπως οι «Πύλες» ***(Gateways)*** και οι «Εικονικές» Βιβλιοθήκες ***(Virtual Libraries)*** που θα αναφερθούν στις επόμενες παραγράφους του κεφαλαίου, οι θεματικοί κατάλογοι δημιουργούνται χειρωνακτικά. Τα *sites* υποβάλλονται από τους δημιουργούς τους ή προσδιορίζονται από τους ***site developers*** και μετά εκχωρούνται σε μία κατάλληλη θεματική κατηγορία ή κατηγορίες από τους *υποστηρικτές των καταλόγων* ***(catalogue maintainers)***. Ωστόσο, αντίθετα με τις *Gateways* και τις *Virtual Libraries*, δεν υπάρχει επιλογή και αξιολόγηση της πληροφορίας προκειμένου να εκτιμηθεί η ποιότητα της πριν καταχωρηθεί στον κατάλογο. Η κατάσταση γίνεται ακόμα πιο πολύπλοκη επειδή οι μηχανές αναζήτησης περιλαμβάνουν και ένα στοιχείο θεματικού καταλόγου και πολλοί κατάλογοι συνδέουν τους χρήστες αυτόματα με μία μηχανή αναζήτησης, ώστε να επεκταθούν τα αποτελέσματα της αναζήτησης. Επιπρόσθετα, μερικοί κατάλογοι αξιώνουν την αξιολόγηση ή και την κατάταξη της «ποιότητας» των *sites* που περιλαμβάνονται.

Γνωστά θεματικά ευρετήρια αποτελούν τα: Yahoo!, One-Global, NetSearch, Virtual Yellow Pages, LookSmart, Galaxy, NBCi.com κ.α. Στον παρακάτω πίνακα 3.1 φαίνονται τα πιο δημοφιλή θεματικά ευρετήρια.

Όνομα	Librarians' Index	Infomine	Britannica Web's Best	Yahoo!	Galaxy
Μέγεθος, τύπος	Περίπου 5.000. Καταρτισμένο από τις δημόσιες βιβλιοθήκες περιλαμβάνοντας πληροφορίες για τις επιχειρήσεις. Πάρα πολύ ποιοτικές ιστοσελίδες. Περιέχει σημαντικές σημειώσεις.	Περίπου 16.000. Καταρτισμένο από τις ακαδημαϊκές βιβλιοθήκες.	Περίπου 150.000. Χειρωνακτική δημιουργία του ευρετηρίου, περιέχει σημειώσεις και είναι ταξινομημένο από τους εκδότες της Britannica.	Περίπου 1 εκατομμύριο. Σπάνιες περιγραφές και σχόλια. Το μεγαλύτερο και το πιο γνωστό θεματικό ευρετήριο.	Περίπου 300.000. Γενικά, καλά σχόλια.
Αναζήτηση φράσεων	Όχι	Ναι. Με χρήση των " ".	Ναι. Περισσότερες από μια λέξεις αναζητούνται ως φράση.	Ναι. Με χρήση των " ".	Όχι
Τελεστές Boolean	AND μεταξύ των λέξεων. Επίσης δέχεται τους OR και NOT.	AND μεταξύ των λέξεων. Επίσης δέχεται τον τελεστή OR.	Δέχεται τους τελεστές AND, OR, NOT.	Όχι	OR μεταξύ των λέξεων. Επίσης δέχεται τους AND, OR, NOT
Sub-Searching	Όχι	Όχι	Στα αποτελέσματα.	Ναι. Στα αποτελέσματα, επιλέγει την αναζήτηση με βάση τις κατηγορίες ή το σύνολο του Yahoo.	Όχι

Πίνακας 3.1: Τα πιο δημοφιλή Θεματικά Ευρετήρια (Subject Directories)

3.5.2. Βασικά Χαρακτηριστικά

Η ταξινόμηση της ανθρώπινης γνώσης σε ευρείες κατηγορίες επιτρέπει την περιήγηση ***(browsing)*** των θεματικών αυτών κατηγοριών από το χρήστη, δίνοντας του έτσι την ευκαιρία να εντοπίσει ευκολότερα το ζητούμενο θέμα ή άλλα συναφή θέματα που δε θα ανακάλυπτε εάν χρησιμοποιούσε διαφορετικό τρόπο προσέγγισης. Η σημαντική διαφοροποίηση των θεματικών καταλόγων σε σχέση με τις μηχανές αναζήτησης, είναι ότι διαθέτουν ένα ευρετήριο, μία κατάταξη της ανθρώπινης γνώσης, που διαμορφώθηκε από ειδικό επιστήμονα και όχι από κάποιο *robot*, που χρησιμοποιούν οι μηχανές αναζήτησης. Η διαφορά αυτή είναι ιδιαίτερα σημαντική για το χρήστη που έχει μια γενική ιδέα για το θέμα για το οποίο αναζητά πληροφορίες και βασίζεται στη λογική του προκειμένου να τις εντοπίσει.

Σε ένα θεματικό κατάλογο, όπου η γνώση είναι ιεραρχικά ταξινομημένη, ξεκινώντας από την ευρύτερη κατηγορία και προχωρώντας στις ειδικότερες υποκατηγορίες, ο χρήστης θα μπορέσει να προσεγγίσει με μεγαλύτερη πιθανότητα επιτυχίας το θέμα από ότι αν 0χρησιμοποιούσε μια μηχανή αναζήτησης. Από την άλλη όμως, μια και τα θεματικά ευρετήρια βασίζονται στον ανθρώπινο παράγοντα για την ανανέωση των περιεχομένων τους, είναι φυσικό να μην είναι ενημερωμένα στο βαθμό που είναι οι μηχανές και τα άλλα εργαλεία αναζήτησης που βασίζονται σε λογισμικό,

το οποίο αναλαμβάνει να ανανεώσει τα περιεχόμενα των ευρετηρίων τους. Έτσι, στη σύγκριση μεταξύ μηχανών αναζήτησης και θεματικών κατάλογων, ο χρήστης θα πρέπει να αξιολογήσει από τη μία την πληρότητα και την ενήμερη πληροφορία των μηχανών αναζήτησης, και από την άλλη τη σχετικότητα την οποία προσφέρουν οι θεματικοί κατάλογοι σε σχέση πάντα με τους αναζητήσιμους όρους. Ο συνδυασμός θεματικών καταλόγων και μηχανών αναζήτησης δείχνει βέβαια να αποτελεί τη χρυσή τομή, και για αυτό έχουν γίνει ορισμένες προσπάθειες προς την κατεύθυνση αυτή.

Σε γενικές γραμμές στα θεματικά ευρετήρια διακρίνουμε τα παρακάτω βασικά χαρακτηριστικά:

- ***Δυνατότητα αναζήτησης με βάση τα συμφραζόμενα (context-based searching):*** Όπως προηγούμενα αναφέρθηκε, η δυνατότητα του χρήστη να περιηγείται από την ευρύτερη προς την ειδικότερη θεματική κατηγορία ενδέχεται να οδηγήσει στην ανακάλυψη συναφών θεμάτων και όρων που σε διαφορετική περίπτωση δε θα ανακάλυπτε.
- ***Αναζήτηση σε επιλεγμένες πληροφοριακές πηγές:*** Επειδή η επιλογή των εν λόγω θεμάτων-κατηγοριών γίνεται από ανθρώπους, στις περισσότερες περιπτώσεις τα περιεχόμενα εκτιμώνται ως αξιόπιστες και αξιόλογες πληροφοριακές πηγές. Θα πρέπει ωστόσο να συμπληρωθεί ότι, ακριβώς επειδή η δημιουργία και η σχεδίαση αυτών των ευρετηρίων γίνεται από ανθρώπους, τα αποτελέσματα μιας τέτοιας αναζήτησης ενδέχεται να είναι σημαντικά λιγότερα από εκείνα που θα προκύψουν από τη χρήση μιας μηχανής αναζήτησης.
- ***Αποφυγή διπλών εγγραφών:*** Όπως και στην προηγούμενη περίπτωση επειδή η εισαγωγή και η οργάνωση των θεμάτων και των κατηγοριών γίνονται από ανθρώπους, είναι εξαιρετικά σπάνιο σε μια αναζήτηση να εμφανιστούν διπλές καταχωρήσεις για την ίδια θεματική ενότητα.

Ως μειονέκτημα, εκτός του περιορισμένου αριθμού των αποτελεσμάτων που ενδέχεται να ανακτηθούν, μπορεί να αναφερθεί το γεγονός της έλλειψης ενός *ελεγχόμενου λεξιλογίου* ***(controlled vocabulary)***, η οποία ίσως προκαλέσει σύγχυση στο λιγότερο εξοικειωμένο χρήστη και δυσκολέψει σημαντικά την περιήγηση του στον κατάλογο, όταν διαπιστώσει ότι η ίδια έννοια εμφανίζεται με πολλούς διαφορετικούς όρους.

3.5.3. Εξειδικευμένα Θεματικά Ευρετήρια (Specialized Subject Directories)

Εξαιτίας του τεράστιου μεγέθους του διαδικτύου και του συνεχούς μετασχηματισμού του, η ενημέρωση όλων των θεματικών τομέων είναι ανθρωπίνως αδύνατη. Συνεπώς, ένας οδηγός που δημιουργείται από ένα ειδικό για σημαντικούς πόρους του πεδίου ειδίκευσης του είναι πιο πιθανό να δημιουργήσει σχετική πληροφορία και συνήθως είναι πιο εκτεταμένη από την πληροφορία ενός γενικού οδηγού. Τέτοιοι οδηγοί υπάρχουν ουσιαστικά για κάθε θέμα. Για παράδειγμα, το *Voice of the Shuttle (http://vos.ucsb.edu)* παρέχει ένα άριστο σημείο εκκίνησης για έρευνα κλασσικών μελετών. Οι οπαδοί των ταινιών επίσης, θα πρέπει να λάβουν υπόψη τους ως εκκίνηση της αναζήτησης τους το Internet Movie Database (http://us.imdb.com).

Μερικά *Web sites* λειτουργούν ως συλλογές ή ***clearinghouses*** εξειδικευμένων θεματικών ευρετηρίων. Πολλά από αυτά τα *sites* προσφέρουν ανασκοπήσεις και σχολιασμούς των θεματικών ευρετηρίων που περιλαμβάνουν και τα περισσότερα λειτουργούν βασισμένα στην αρχή ότι οι ειδικοί των γνωστικών τομέων διατηρούν το κάθε ένα θεματικό ευρετήριο. Μερικά *clearinghouses* διατηρούν τους εξειδικευμένους οδηγούς στο δικό τους Web site, ενώ άλλα συνδέονται με οδηγούς που βρίσκονται σε διάφορα απομακρυσμένα sites. Παραδείγματα *clearinghouses* είναι: Argus Clearinghouse (http://www.clearinghouse.net), About.com (http://about.com), WWW Virtual Library (http://www.vlib.org).

3.5.4. Πλεονεκτήματα και Μειονεκτήματα των Subject Catalogues- Subject Directories

Τα θεματικά ευρετήρια και οι κατάλογοι μπορούν και παρέχουν πιο ουσιαστικά αποτελέσματα αναζήτησης από τις μηχανές αναζήτησης. Οι δημιουργοί των *sites* είναι γενικά υπεύθυνοι για την περιγραφή του υλικού τους και ως εκ τούτου οι περιγραφές τους είναι συχνά πιο βοηθητικές από εκείνες που δημιουργούνται αυτόματα από τις μηχανές αναζήτησης. Ωστόσο, υπάρχουν μειονεκτήματα στην ανθρώπινη υποστήριξη. Συγκεκριμένα, οι θεματικοί κατάλογοι και τα ευρετήρια δεν είναι τόσο εκτεταμένα όσον αφορά την κάλυψη τους όσο οι μηχανές αναζήτησης - το Yahoo! κάλυπτε μόνο 1,5 εκατομμύρια ιστοσελίδες (Cooke, 1999) σε σχέση με το AltaVista, το οποίο αξιώνει την καταχώρηση σε ευρετήριο 250 εκατομμυρίων σελίδων. Εντούτοις, είναι σημαντικά

μεγαλύτερα από τις gateways και τις virtual libraries, όπου απαιτείται εκτεταμένη ανθρώπινη προσπάθεια για την επιλογή, αξιολόγηση, περιγραφή και καταχώρηση των πόρων σε καταλόγους.

Ένα άλλο μειονέκτημα είναι ότι τα ευρετήρια δεν ενημερώνονται αυτόματα, και επομένως όταν τα *sites* ή οι ιστοσελίδες μεταβάλλονται, το ευρετήριο δεν ενημερώνεται απαραίτητα τόσο τακτικά όσο μία μηχανή αναζήτησης. Αυτό είναι ένα πρόβλημα το οποίο ισχύει κατά τον ίδιο τρόπο και στις gateways και τις virtual libraries.

❖ ***Έλλειψη ενδείξεων ποιότητας***

Ένα μειονέκτημα ξεχωριστού ενδιαφέροντος είναι ότι οι κατάλογοι και τα ευρετήρια δε διαχωρίζουν απαραίτητα τα sites όσον αφορά την ποιότητά τους. Αυτοί που εμπλέκονται στην ανάπτυξη και υποστήριξη των βάσεων δεδομένων ενδιαφέρονται για την θεματική σχετικότητα του υλικού. Το γεγονός αυτό έρχεται σε αντίθεση με τις gateways και τις virtual libraries, όπου δε δίνεται έμφαση μόνο στην επιλογή των πιθανολογούμενων σχετικών πόρων, αλλά και στην επιλογή υψηλής ποιότητας υλικού για ένα συγκεκριμένο κοινό. Μερικές υπηρεσίες καταλόγων ή ευρετηρίων αξιώνουν την αξιολόγηση των sites, αν και η χρησιμότητα των εκτιμήσεων ποιότητας είναι περιορισμένη.

3.6. Υβριδικές Μηχανές Αναζήτησης (Hybrid Search Engines)

Μερικές μηχανές αναζήτησης διαθέτουν και ένα συνεργαζόμενο ευρετήριο ***(Hybrid Search Engines)*** (Montebello and Ciappara, 2000). Αυτές είναι *sites* που έχουν αναθεωρηθεί ή βαθμολογηθεί. Ως επί το πλείστον, αυτά τα αναθεωρημένα *sites* δεν εμφανίζονται ως «προεπιλεγέντα» όταν ένα ερώτημα υποβάλλεται σε μία υβριδική μηχανή αναζήτησης. Αντίθετα, ο χρήστης πρέπει συνειδητά να επιλέξει να δει τις αναθεωρήσεις. Αυτή είναι μία περίπτωση συγχώνευσης των υπηρεσιών που παρέχονται από τις μηχανές αναζήτησης και τα ευρετήρια σε μία υπηρεσία, αλλά το πρόβλημα είναι ότι τα *sites* πρέπει να υποβληθούν για αναθεώρηση από τα άτομα που διατηρούν το ευρετήριο.

Οι αναθεωρητές συχνά ελέγχουν τα υποβαλλόμενα *sites* και εν συνεχεία επιλέγουν να προσθέσουν εκείνες τις ιστοσελίδες που δείχνουν «ελκυστικές», σύμφωνα με τη γνώμη τους. Αυτού του τύπου οι μηχανές αναζήτησης εξακολουθούν βέβαια να υφίστανται τα προβλήματα των ευρετηρίων ενώ έχουν και το πρόσθετο μειονέκτημα ότι δεν υπάρχει εγγύηση ότι τα υποβαλλόμενα για αναθεώρηση *sites*, που προορίζονται να συμπεριληφθούν στο ευρετήριο, τελικά θα συμπεριληφθούν. Είναι θέμα τύχης και ποιότητας, ενώ δε δίνεται στους χρήστες η επιλογή να αποφασίσουν ελεύθερα σχετικά με το τι είναι σχετικό και τι όχι.

3.7. «Πύλες» και «Εικονικές» Βιβλιοθήκες (Gateways and Virtual Libraries)

Η διαθεσιμότητα μεγάλων ποσοτήτων πληροφορίας ποικίλης ποιότητας μέσω του διαδικτύου έχει αποτελέσει το έναυσμα για κάποιες πρωτοβουλίες που αφορούν την παροχή μίας πιο αποδοτικής πρόσβασης στην πληροφορία. Καθώς το *internet* αναπτύσσεται το ίδιο συμβαίνει και με την πληροφορία που πρέπει να βρεθεί. Ωστόσο, αυτό οδηγεί σε δύο κύρια προβλήματα: πώς να βρεθεί η πληροφορία και πως θα αξιολογηθεί αυτή η πληροφορία όταν εντοπιστεί. Οι *«Πύλες»* και οι *«Εικονικές Βιβλιοθήκες»* είναι η απάντηση στις παραπάνω δύο ερωτήσεις καθώς έχουν σχεδιαστεί να προσφέρουν γρήγορους και εύκολους τρόπους αναζήτησης ποιοτικής πληροφορίας, η οποία βοηθά τους ερευνητές στη δουλειά τους.

Συγκεκριμένα, οι *«Πύλες»* ***Gateways*** και οι *«Εικονικές Βιβλιοθήκες»,* ***Virtual Libraries*** είναι συλλογές πηγών πληροφορίας υψηλής ποιότητας ενός συγκεκριμένου θεματικού τομέα, που αφορούν ένα καθορισμένο κοινό (Cooke, 1999). Είναι δύσκολο να διαχωριστούν περιληπτικά τα εργαλεία αναζήτησης του διαδικτύου διότι το κάθε ένα λειτουργεί με ελάχιστα διαφορετικό τρόπο. Ωστόσο, τα συγκεκριμένα εργαλεία είναι οδηγοί ποιοτικού υλικού. Το καθοριστικό χαρακτηριστικό είναι ότι υπάρχουν προκαθορισμένα κριτήρια για την επιλογή και την αξιολόγηση του υλικού. Εξάλλου, αυτοί οι οδηγοί έχουν αναπτυχθεί από ειδικούς διαφόρων τομέων, συνήθως από άτομα με επαγγέλματα που σχετίζονται με βιβλιοθήκες,

Το κυριότερο πλεονέκτημα της χρήσης μίας *gateway* ή μίας *virtual library* είναι ότι κάποιος με γνώση του χώρου έχει ήδη ερευνήσει το διαδίκτυο και έχει προσπαθήσει να διαχωρίσει το χρήσιμο υλικό από το λιγότερο χρήσιμο. Για αυτό το λόγο, η χρήση μίας *gateway* ή μίας *virtual library* σε αρχικό στάδιο, μπορεί να εξοικονομήσει πολύ χρόνο και προσπάθεια, εφόσον δε χρειάζεται αναζήτηση μέσα σε χιλιάδες ανεπίκαιρους ή άχρηστους δικτυακούς τόπους.

Αξίζει να αναφερθεί σε αυτό το σημείο ότι μια *virtual library* μπορεί να χρησιμοποιηθεί σε διάφορες περιπτώσεις. Εκτός από την περίπτωση που κάποιος γνωρίζει ακριβώς το τι ψάχνει, μπορεί να χρησιμοποιήσει την κατάλληλη *virtual library* βρίσκοντας τις πληροφορίες εύκολα και γρήγορα, αλλά και στην περίπτωση που κάποιος δεν είναι σίγουρος για το τι ψάχνει, γνωρίζοντας όμως ότι ανήκει σε ένα συγκεκριμένο τομέα, αξίζει να ψάξει στην κατάλληλη *virtual library* αποσαφηνίζοντάς το. Καθώς στις *virtual libraries* το ευρετήριο δεν δημιουργείται αυτόματα, όπως στις

περισσότερες μηχανές αναζήτησης, είναι βέβαιο ότι όποιος τις χρησιμοποιεί θα απευθύνεται σε πηγές ποιοτικής πληροφορίας και είναι προτιμότερο από την άσκοπη περιήγηση στο *Web*.

Ωστόσο, εάν κάποιος θέλει να έχει μια κατανοητή άποψη για έναν τομέα και προσπαθεί να ανακτήσει πληροφορίες για τον συγκεκριμένο τομέα, η *virtual library* αποτελεί ένα χρήσιμο σημείο για την εκκίνηση της περιήγησης, χωρίς βέβαια να βρεθούν όλες οι πληροφορίες. Ίσως να χρειαστεί να γίνει έρευνα με τη χρήση μιας μηχανής αναζήτησης για την απόκτηση της γενικής εικόνας.

3.7.1. Βασικά Χαρακτηριστικά

Το 1993, έγινε μία έρευνα στο Ηνωμένο Βασίλειο που αφορούσε τρόπους αντιμετώπισης της συμπίεσης των πόρων της βιβλιοθήκης, η οποία είχε προκληθεί από τη ραγδαία αύξηση του αριθμού των μαθητών και από την παγκόσμια έκρηξη της ακαδημαϊκής γνώσης και πληροφορίας. Το γεγονός αυτό οδήγησε στο *Electronic Libraries Program, eLib* (http://ukoln.ac.uk/services/elib), ένα κεντρικά χρηματοδοτούμενο εθνικό πρόγραμμα έρευνας σχετικά με το ρόλο και την ανάπτυξη της *«ηλεκτρονικής βιβλιοθήκης»*. Υπήρχαν διάφορα παρακλάδια της *eLib*, ένα από τα οποία ήταν η «πρόσβαση σε δικτυακούς πόρους». Ως μέρος αυτής του χώρου, αναπτύχθηκαν μερικές gateways πληροφορίας *βασισμένες σε θέματα* ***(subject-based)***, όπως η ***SOSIG (Social Science Information Gateway)***.

Κάθε μία από τις *gateways* ήταν λίγο διαφορετική, αλλά γενικά όλες μοιράζονταν τα ίδια χαρακτηριστικά:

- Μια ερευνήσιμη και διαχειρίσιμη βάση δεδομένων που περιλάμβανε περιγραφές διαδικτυακών πόρων ενός συγκεκριμένου θεματικού χώρου
- Σαφώς ορισμένα κριτήρια για την αξιολόγηση της ποιότητας του υλικού πριν από την καταχώρηση του στην *gateway*
- Την ανάμειξη επαγγελματιών βιβλιοθήκης και άλλων ειδικών στην ανάπτυξη της υπηρεσίας
- Χειρωνακτική δημιουργία εγγραφών για την εξασφάλιση επεξηγηματικών και ενημερωτικών περιγραφών των πόρων
- Καταχώρηση σε καταλόγους και ταξινόμηση του υλικού χρησιμοποιώντας παραδοσιακές μεθόδους βιβλιοθήκης προκειμένου να υπάρχει δυνατότητα αποτελεσματικής ανάκτησης.

Αυτές οι υπηρεσίες προορίζονταν γενικά για την κοινότητα ανώτερης εκπαίδευσης του Ηνωμένου Βασιλείου και αναπτύχθηκαν από μία διεθνή εταιρική συνεργασία, κυρίως από τον ακαδημαϊκό χώρο.

Στη βιβλιογραφία υπάρχουν εκατοντάδες *πύλες ή εικονικές βιβλιοθήκες* οι οποίες καλύπτουν γενικές ή ειδικές θεματικές ενότητες, πολύ αναλυτικά ή περιληπτικά αναπτυγμένες. Παρακάτω αναφέρονται μερικές διευθύνσεις, από όπου θα μπορούσε κανείς να ξεκινήσει:

- *Μια μεγάλη λίστα από πύλες ή εικονικές βιβλιοθήκες μπορεί να βρεθεί στη* http://www.vlib.org/
- ***The Virtual Reference Collection*** http://libraries.mit.edu/research/virtualref.html
- *Για τις κοινωνικές επιστήμες υπάρχει η* ***The Social Science Information Gateway (SOSIG)*** http://sosig.ac.uk
- *Για θέματα που αφορούν μηχανικούς υπάρχει η* ***Edinburgh Engineering Virtual Library*** http://eevl.ac.uk
- *Για πληροφορίες σε θέματα ιατρικής υπάρχει η* ***Organizing Medical Networked Information (OMNI)*** http://omni.ac.uk
- *Μια άλλη χρήσιμη πύλη για περισσότερες πληροφορίες μπορεί να βρεθεί στην ιστοσελίδα του Πανεπιστημίου του Sussex, που ονομάζεται* ***PIER*** *στη* http://www.susx.ac.uk/library/pier
- *Για τις μαθητικές, διδακτικές και ερευνητικές κοινότητες υπάρχει η* ***The Resource Discovery Network (RDN)*** *(*http://www.rdn.ac.uk/*).*

Θα γίνει μια πιο λεπτομερή αναφορά σε δύο από τις παραπάνω πύλες ή εικονικές βιβλιοθήκες:

- ***The Resource Discovery Network (RDN)*** *(*http://www.rdn.ac.uk/*)*
- ***The Social Science Information Gateway (SOSIG)*** http://sosig.ac.uk

3.7.2. The Resource Discovery Network (RDN)

Η αναγνώριση της σημαντικής συνεισφοράς των *eLib gateways* στην παροχή αποτελεσματικής πρόσβασης σε διαδικτυακό υλικό υψηλής ποιότητας οδήγησε σε πρόσθετη χρηματοδότηση από την κυβέρνηση του Ηνωμένου Βασιλείου για το ***RDN***, ***the Resource Discovery Network*** (http://www.rdn.ac.uk/). Το *RDN* ιδρύθηκε τον Ιανουάριο του 1999 ως μία ελεύθερη υπηρεσία του διαδικτύου, «αφιερωμένη στην

παροχή πρόσβασης σε διαδικτυακούς πόρους υψηλής ποιότητας για τις μαθητικές, διδακτικές και ερευνητικές κοινότητες». Το *RDN* συνέχισε πάνω στη βάση που είχε θέσει η *eLib* - κάθε *eLib gateway* υπέβαλλε μία πρόταση για την ανάπτυξη της υπηρεσίας της σε ένα θεματικό ***'hub'*** *('κομβικό σημείο'}*. Με αυτό τον τρόπο, κάθε *hub* θα είχε ένα πιο ευρύ πεδίο από τις πρωτότυπες *gateways* και όλοι οι σχετικοί με την κοινότητα ανώτερης εκπαίδευσης του Ηνωμένου Βασιλείου τομείς θα καλύπτονταν από τη συλλογή των *hubs*.

Τα παρακάτω *hubs* είχαν αναπτυχθεί ή ήταν υπό ανάπτυξη το 1999 (Cooke, 1999):

- ✓ *Computing* (υπό ανάπτυξη)
- ✓ *Engineering* (http://www.eevl.ac.uk/)
- ✓ *Health and medicine* (http://omni.ac.uk/)
- ✓ *Humanities* (http://www.humbul.ac.uk/)
- ✓ *Mathematics* (υπό ανάπτυξη)
- ✓ *Physical sciences* (http://www.psigate.ac.uk/)
- ✓ *Reference resources* (http://www.rdn.ac.uk/findit/)
- ✓ *Social sciences, including business and law* (http://www.sosig.ac.uk/)

Ενώ κάθε *hub* έχει μία μοναδική «ταυτότητα», είναι δυνατή η αναζήτηση μέσω αρκετών *hubs* ταυτόχρονα και υπάρχει πλέον μία πολιτική ανάπτυξης μίας γενικής συλλογής. Ο ρόλος των πρωτότυπων gateways έχει παραμείνει ο ίδιος- ένας κατάλογος περιγραφών επιλεγμένου και αξιολογημένου διαδικτυακού υλικού ενός καθορισμένου τομέα.

Οι *RDN gateways* καλύπτουν ένα μεγάλο εύρος θεματικών τομέων και η δυνατότητα αναζήτησης μέσα σ' αυτά είναι ένα χρήσιμο χαρακτηριστικό. Ωστόσο, οι *gateways* παραμένουν επικεντρωμένες σε πόρους που ενδιαφέρουν τους χρήστες του διαδικτύου του τομέα ανώτερης εκπαίδευση του Ηνωμένου Βασιλείου, αν και πολύ πιθανόν να προκαλούν το ενδιαφέρον και σε πολύ πιο ευρύ κοινό. Υπάρχουν πολλές gateways και η κάθε μία καλύπτει μία διαφορετική «υποομάδα» του διαδικτύου. Μερικές από αυτές είναι: *BUBL Link, Infomine, The Internet Public Library, Librarians' Index to the Internet, The SCOUT Report Signpost* κ.λ.π.

3.7.3. The Social Science Information Gateway (SOSIG)

Η ***SOSIG*** (http://www.sosig.ac.uk/) ήταν μία από τις υπηρεσίες *gateway* που αναπτύχθηκε αρχικά από την *eLib*, η οποία αργότερα εξασφάλισε χρηματοδότηση από την *RDN* προκειμένου να επεκτείνει την κάλυψη της και να γίνει το *hub* για τις κοινωνικές επιστήμες. Η υπηρεσία απευθύνεται σε κοινωνικούς επιστήμονες της ανώτερης εκπαίδευσης και της έρευνας και καλύπτει ένα μεγάλο εύρος θεμάτων, όπως: επιχειρήσεις, εκπαίδευση, γεωγραφία, νόμοι, φιλοσοφία, πολιτική, ψυχολογία και κοινωνιολογία.

Η ***SOSIG*** προσφέρει δύο κύριες επιλογές πρόσβασης της πληροφορίας – *περιήγηση* ***(browsing)*** τις επικεφαλίδες των θεμάτων ή εισαγωγή λέξεων-κλειδιών για αναζήτηση στη βάση δεδομένων.

Η επιλογή 'Economics', για παράδειγμα, η οποία εμφανίζεται στην αρχική οθόνη οδηγεί σε περαιτέρω σχετικές επικεφαλίδες, όπως 'Economics Systems and Theories' και 'Insurance'. Η επιλογή του δεύτερου οδηγεί σε μία λίστα τίτλων των πόρων. Οι πόροι κατηγοριοποιούνται βάσει είδους, π.χ. 'Data', 'Government Publications' και 'Organisations/Societies'. Κλικάροντας στον τίτλο, ο χρήστης οδηγείται σε μία περιγραφή του συγκεκριμένου πόρου. Διαθέσιμη είναι επίσης και μία 'Advanced Search' επιλογή στην αρχική οθόνη, που δίνει τη δυνατότητα στο χρήστη να καθορίσει το σημείο που επιθυμεί να εμφανίζονται οι όροι της αναζήτησης (π.χ. στον τίτλο του πόρου, στην περιγραφή ή τις λέξεις-κλειδιά) καθώς και το είδος του πόρου που επιθυμεί να ανακτήσει (π.χ. data, government publications και organizations/societies). Ένα πρόσθετο χαρακτηριστικό επιτρέπει με την εισαγωγή της *ρίζας μίας λέξης* ***(word stem)*** να ανακτηθεί κάθε πιθανή παραλλαγή της ρίζας- π.χ. η εισαγωγή της λέξης 'econ' θα ανακτήσει sites που αναφέρουν τις λέξεις 'economy', 'economics', κ.λ.π.

- ***Επιλογή πόρων, αξιολόγηση και περιγραφή***

Ένα σημαντικό χαρακτηριστικό της ***SOSIG***, όπως και άλλων *gateways* και *virtual libraries*, είναι ότι παρέχει πρόσβαση μόνο σε περιγραφές υψηλής ποιότητας πόρων. Πριν καταχωρηθεί στη βάση δεδομένων, κάθε πόρος έχει αξιολογηθεί βάσει σαφών κριτηρίων (http://sosig.ac.uk/desire/ecrit.html). Τα κριτήρια αυτά δεν σχετίζονται μόνο με την παρουσίαση της πληροφορίας, αλλά και με την θεματική *κάλυψη* ***(coverage)***, *ενημερότητα* ***(currency)*** και *ακρίβεια* ***(precision)***. Η διαθεσιμότητα των κριτηρίων online σημαίνει ότι οι χρήστες μπορούν να αξιολογήσουν αν τα κριτήρια που

χρησιμοποιούνται από τη ***SOSIG*** ισοδυναμούν με τις δικές τους απόψεις σχετικά με τις απαιτήσεις ενός ποιοτικού πόρου.

Οι υπεύθυνοι βιβλιοθήκης της ***SOSIG*** συμπληρώνουν μία φόρμα για κάθε πόρο που έχει επιλεγεί για καταχώρηση στη βάση δεδομένων. Η φόρμα είναι παρόμοια με μία εγγραφή σε ένα συνηθισμένο σύστημα βιβλιοθήκης και περιέχει πληροφορία όπως, τίτλο, περιγραφή, λέξεις-κλειδιά και διευθύνσεις διαδικτύου. Οι εγγραφές προστίθενται στη βάση δεδομένων και όταν διενεργείται μία αναζήτηση, ο χρήστης αναζητά μία πληροφορία που βρίσκεται στη βάση δεδομένων σχετικά με κάποιο πόρο και όχι κάθε λέξη του πόρου. Τα σχετικά μέρη αυτής της εγγραφής εμφανίζονται στη συνέχεια στα αποτελέσματα της αναζήτησης. Οι περιγραφές έχουν σκοπό να δώσουν τη δυνατότητα στο χρήστη να αξιολογήσει την αξία και τη χρησιμότητα του κάθε πόρου πριν συνδεθεί με το ίδιο το site. Εκτός από την παροχή μίας λεπτομερής περιγραφής των θεματικών πεδίων και του καλυπτόμενου υλικού, θίγονται και θέματα πρόσβασης. Διάφοροι ειδικοί είναι υπεύθυνοι για τη διαχείριση της συλλογής συγκεκριμένων πεδίων της υπηρεσίας, για την υποστήριξη των τμημάτων καθώς και για την εξέταση σχετικά με τη διαθεσιμότητα και την ενημερότητα των *sites*.

- ***Καταχώρηση σε κατάλογο***

Εκτός από τις περιγραφές για κάθε πόρο, πολλές *gateways* και *virtual libraries* χρησιμοποιούν παραδοσιακές μεθόδους καταχώρησης σε καταλόγους και ταξινόμησης ώστε να κατηγοριοποιηθούν οι πόροι και να εξασφαλιστεί η αποτελεσματική τους ανάκτηση. Οι λέξεις- κλειδιά στη ***SOSIG*** είναι ελεγχόμενοι όροι, που χρησιμοποιούνται για την περιγραφή των θεματικών πεδίων που καλύπτονται από ένα πόρο. Ο λόγος της χρήσης ελεγχόμενων λέξεων-κλειδιών είναι οι διαφορετικές λέξεις που χρησιμοποιούν διαφορετικοί δημιουργοί για την περιγραφή του ίδιου θέματος. Αν υιοθετηθούν προκαθορισμένες λέξεις- κλειδιά, μπορεί ο χρήστης να βρει ποια λέξη- κλειδί χρησιμοποιείται για την περιγραφή του θέματος του και επομένως να ανακτήσει όλες τις σχετικές εγγραφές. Χρησιμοποιείται ένα παραδοσιακό σχήμα από λέξεις- κλειδιά - αυτό είναι ένα συνηθισμένο σχήμα που χρησιμοποιείται στις βιβλιοθήκες σ' όλο τον κόσμο για την περιγραφή του υλικού.

Όταν ο χρήστης κάνει browsing στις επικεφαλίδες της ***SOSIG***, στην πραγματικότητα κάνει *browsing* των ελεγχόμενων λέξεων- κλειδιών. Οι λέξεις- κλειδιά μπορούν να ομαδοποιηθούν βάσει πεδίου και να καταταγούν σε μία ιεραρχία σύμφωνα με το αν ένας όρος είναι περισσότερο ή λιγότερο συγκεκριμένος από έναν άλλο, καθώς

και σύμφωνα με το αν ένας όρος είναι τόσο συγκεκριμένος όσο ένας άλλος αλλά έχει μία σχετική ερμηνεία (όχι ακριβώς την ίδια). Για παράδειγμα, εάν κάποιος επιλέξει 'accountancy' από την αρχική οθόνη, εμφανίζεται η ευρύτερη επικεφαλίδα 'business', ο στενότερος όρος 'auditing', καθώς και οι σχετικοί όροι 'economics' και 'management'.

Εκτός από τις θεματικές επικεφαλίδες, οι πόροι στη ***SOSIG*** κατηγοριοποιούνται σύμφωνα και με το είδος του εγγράφου. Αυτό επομένως σημαίνει ότι όταν ο κατάλογος γίνεται browsed, οι πόροι κατηγοριοποιούνται βάσει του είδους τους. Κατά τον ίδιο τρόπο, είναι δυνατό στην οθόνη της προχωρημένης αναζήτησης η αναζήτηση να περιοριστεί σύμφωνα με ένα συγκεκριμένο είδος πληροφορίας.

- ***Επέκταση της κάλυψης: Μηχανή αναζήτησης κοινωνικών επιστητών***

Εξαιτίας της υψηλού επιπέδου χειρωνακτικής εισόδου που αφορά την επιλογή, την αξιολόγηση και την περιγραφή των πόρων, οι gateways και οι virtual libraries περιορίζονται όσον αφορά τον όγκο του υλικού που μπορούν να καλύψουν, ιδιαίτερα σε σύγκριση με τις μηχανές αναζήτησης. Για αυτό το λόγο, πολλές gateways και virtual libraries πλέον συνδέουν τους χρήστες με μία μηχανή αναζήτησης ώστε να επεκταθούν οι αναζητήσεις τους αυτόματα. Η ***SOSIG*** συνδέεται με μία βάση δεδομένων από 50.000 ιστοσελίδες κοινωνικών επιστημών που έχουν προσδιοριστεί με τη χρήση ενός *λογισμικού* ***(software)***, το 'harvester' (http://sosig.ac.uk./harvester.html). Ωστόσο, δεν υπάρχει φιλτράρισμα ποιότητας του υλικού που περιέχεται σ' αυτό το τμήμα της ***SOSIG*** και οι περιγραφές των πόρων δημιουργούνται αυτόματα.

3.7.4. Πλεονεκτήματα και Μειονεκτήματα των «Πυλών» και των «Εικονικών» Βιβλιοθηκών

- ***Ποιότητα πληροφορίας και όχι ποσότητα***

Το κύριο πλεονέκτημα των *gateways* και των *virtual libraries* είναι ότι παρέχουν πρόσβαση σε περιγραφές μόνο υψηλής ποιότητας πόρων. Έχουν αναπτυχθεί και υποστηρίζονται από επαγγελματίες της πληροφορίας και από ειδικούς και συνεπώς εγγυώνται ότι κάποιος γνώστης του χώρου έχει ήδη προσδιορίσει και αξιολογήσει πόρους υψηλής ποιότητας. Αυτό εξοικονομεί πολύ προσπάθεια από πλευράς του χρήστη όσον αφορά το φιλτράρισμα των πιθανών χρήσιμων πηγών από τις τεράστιες ποσότητες απορριμμάτων που είναι διαθέσιμα στο διαδίκτυο. Επιπρόσθετα, οι

περιγραφές σκοπεύουν στην παροχή ακριβούς, μεστής και σημαντικής ένδειξης της αξίας και της χρησιμότητας κάθε πόρου, αποτρέποντας το χρήστη από άσχετα, ανεπίκαιρα και ανακριβή *sites*.

✓ ***Εστίαση σε περιορισμένο κοινό***

Ένα από τα προβλήματα των γενικών εργαλείων αναζήτησης είναι το μη καθορισμένο κοινό τους- τέτοιες υπηρεσίες στοχεύουν σε κάθε χρήστη του διαδικτύου και συνεπώς θα ήταν δύσκολο, όλοι οι πόροι που περιλαμβάνονται να είναι σχετικοί για τον κάθε έναν. Οι υπηρεσίες των *gateways* και των *virtual libraries* έχουν καθορισμένο κοινό. Για παράδειγμα, η ***SOSIG*** στοχεύει πρωτίστως σε χρήστες των κοινωνικών επιστημών, βασισμένη στην ανώτερη εκπαίδευση ή την έρευνα. Η ποιότητα είναι υποκειμενικό θέμα, δηλαδή αυτό που για κάποιον είναι πηγή ποιοτικής πληροφορίας είναι πιθανόν για κάποιον άλλο να είναι απόρριμμα. Συνεπώς, όσο πιο σαφώς και περιορισμένα καθορισμένη είναι η εστίαση του κοινού μίας υπηρεσίας τόσο ευκολότερο είναι να αποφασίσει κανείς εάν το υλικό που καλύπτεται είναι πιθανόν να τον ενδιαφέρει.

✓ ***Σαφή κριτήρια αξιολόγησης***

Ένα άλλο πλεονέκτημα είναι ότι τα κριτήρια που χρησιμοποιούνται για αξιολόγηση από αυτές τις υπηρεσίες είναι σαφή: υπάρχει συνήθως ένα έγγραφο που περιγράφει τα κριτήρια ή μία πολιτική ανάπτυξης της συλλογής. Για παράδειγμα, τα ακόλουθα κριτήρια μπορούν να αξιολογηθούν online:

- *Evaluating Internet Resources for SOSIG (http://sosig.uk./desire/ecrit.html)*
- *Selection Criteria for the Librarians' Index to the Internet (http://lii.org/search/file/subcriteria)*
- *SCOUT Report Selection Criteria (http://scout.cs.wisc.edu./report/sr/criteria.html)*
- *The Argus Clearinghouse Ratings System (http://www.clearinghouse.net/ratings.html)*

ΚΕΦΑΛΑΙΟ 4
ΟΙ ΚΥΡΙΕΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Με τη γρήγορη ανάπτυξη του διαδικτύου, όλο και περισσότερες πληροφορίες είναι διαθέσιμες στον ιστό και κατά συνέπεια η ανάκτηση πληροφοριών έχει γίνει γεγονός για τους περισσότερους χρήστες διαδικτύου. Εντούτοις, έναντι της κλασικής ανάκτησης πληροφοριών, τα συστήματα ανάκτησης πληροφοριών του ιστού στο σύνολό τους αντιμετωπίζονται ως διαφορετικά σύνολα δεδομένων.

Η μοναδικότητα της ανάκτησης πληροφοριών του ιστού οφείλεται κυρίως στους παρακάτω λόγους (Huang):

- **Όγκος του ιστού (Bulk):** Ο όγκος του διαδικτύου ήταν 350 εκατομμύρια έγγραφα (Ιούλιος, 1998), τα οποία αυξάνονται με ταχύτητα 20 εκατομμύρια το μήνα. Σήμερα εικάζεται ότι στο διαδίκτυο βρίσκονται συνδεδεμένοι περισσότεροι από 50 εκατομμύρια υπολογιστές.
- **Ο Δυναμικός χαρακτήρας του διαδικτύου (Dynamic Internet):** Το διαδίκτυο αλλάζει καθημερινά ενώ τα περισσότερα κλασικά συστήματα ανάκτησης πληροφοριών σχεδιάζονται συνήθως για τις *στατικές βάσεις δεδομένων* ***(static text databases)***.
- **Ετερογένεια του διαδικτύου (Heterogeneity):** Το διαδίκτυο περιέχει μια μεγάλη ποικιλία από τύπους εγγράφων: εικόνες, μουσικά αρχεία, κείμενα, χειρόγραφα κ.λ.π.

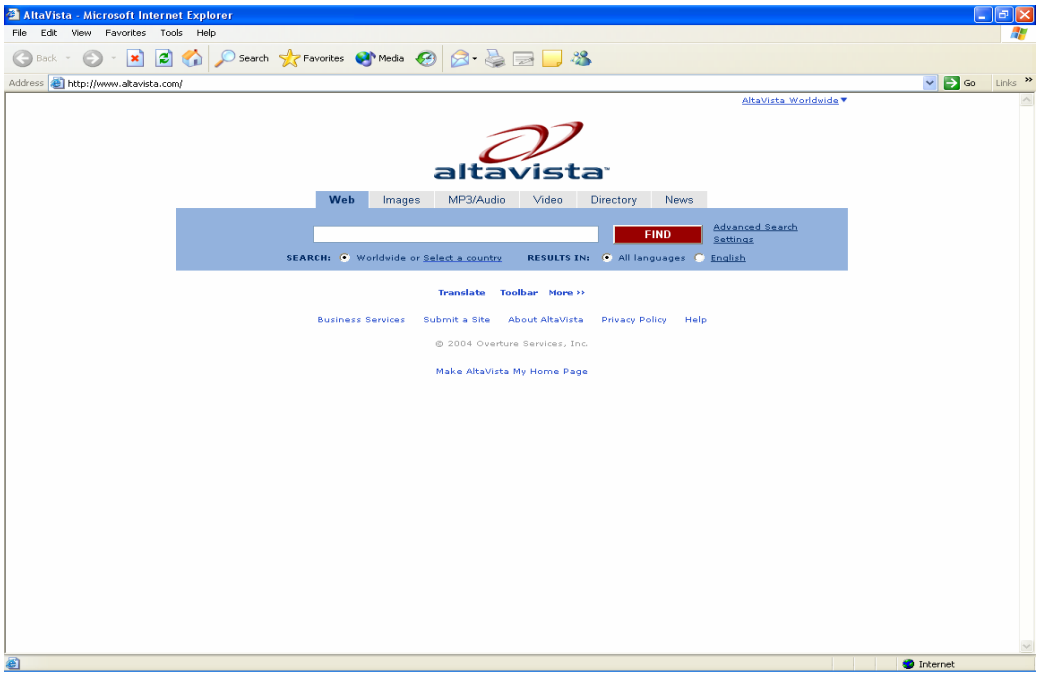
- **Ποικιλία γλωσσών (Variety of Languages):** Οι διάφορες γλώσσες που χρησιμοποιούνται στο διαδίκτυο είναι περισσότερες από 100.
- **Οι διπλές εγγραφές (Duplication):** Η αντιγραφή είναι ένα άλλο σημαντικό χαρακτηριστικό του ιστού. Ισχύει ότι σχεδόν το 30% των ιστοσελίδων είναι όμοιες.
- **Υψηλή συνδεσιμότητα (High Linkage):** Κάθε έγγραφο έχει κατά μέσο όρο περισσότερες από 8 συνδέσεις με άλλες σελίδες.
- **Τα λάθος διατυπωμένα ερωτήματα (Ill-formed queries):** Τα συστήματα ανάκτησης πληροφοριών απαιτούνται να απαντήσουν μικρά και όχι ιδιαίτερα καλά διατυπωμένα ερωτήματα από τους χρήστες του διαδικτύου.
- **Η μεγάλη ποικιλία των χρηστών (Wide Variance in Users):** Κάθε χρήστης ιστού διαφέρει ευρέως στις ανάγκες, στις προσδοκίες και στη γνώση του.
- **Συγκεκριμένη συμπεριφορά των χρηστών (Specific Behavior):** Υπολογίζεται ότι σχεδόν το 85% των χρηστών εξετάζουν μόνο την πρώτη οθόνη των αποτελεσμάτων που προκύπτουν από τις μηχανές αναζήτησης. Το 78% των χρηστών δεν τροποποιούν ποτέ το πρώτο ερώτημά τους.

Έτσι η μεγάλη πρόκληση στην ανάκτηση πληροφοριών του ιστού είναι να ικανοποιηθούν οι ανάγκες των χρηστών σε πληροφορίες δεδομένου της ετερογένειας του ιστού και των λάθος διατυπωμένων ερωτήσεων.

4.1. AltaVista (www.altavista.com)

4.1.1. Εισαγωγή

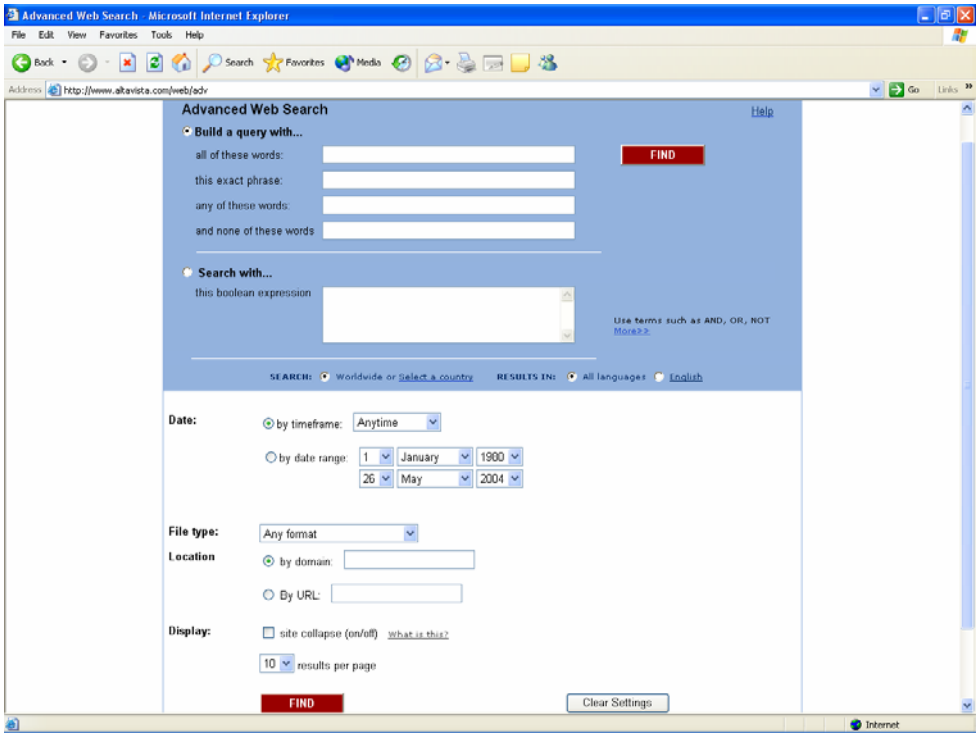
Η AltaVista θεωρείται η πλέον διαδεδομένη μηχανή γενικής αναζήτησης, αφού κατορθώνει με τον τεράστιο όγκο του ευρετηρίου που διαθέτει, το φιλικό ***interface*** και την πολύ καλή φήμη της να παραμένει ανάμεσα στις πρώτες μηχανές αναζήτησης στις προτιμήσεις των χρηστών. Όταν εμφανίστηκε τον Δεκέμβριο του 1995, η AltaVista ισχυριζόταν ότι είχε τη μεγαλύτερη βάση δεδομένων και την κατοχή του μεγαλύτερου ευρετηρίου στο διαδίκτυο. Σύντομα, με την εμφάνιση νέων ανταγωνιστών και της δραστικής αλλαγής του μεγέθους του web, η AltaVista έχασε την αίγλη του πρώτου όμως όχι και τη θέση της ανάμεσα στις καλύτερες μηχανές γενικής αναζήτησης. Πρόσφατα η AltaVista προχώρησε σε ριζικές αλλαγές σε ό,τι αφορά τόσο τον τομέα της αισθητικής όσο και τα περιεχόμενα των βάσεων δεδομένων που ευρετηριάζει. Έτσι, στα τέλη Οκτώβρη του 2000 η μηχανή εμφανίστηκε ριζικά ανανεωμένη, με πλήρως ανασχεδιασμένη κεντρική σελίδα (Σχήμα 4.1.) και πολλές νέες υπηρεσίες και προϊόντα.



Σχήμα 4.1.: Η αρχική σελίδα (Homepage) της AltaVista (www.altavista.com)

Στη νέα αυτή παρουσία της η AltaVista έχει ανασχεδιάσει ή προσθέσει λειτουργίες και επιλογές που αφορούν την παρουσίαση και την ταξινόμηση των αποτελεσμάτων, το μέγεθος των ευρετηριαζόμενων πηγών καθώς και τις δυνατότητες

που προσφέρονται μέσω των εξειδικευμένων αναζητήσεων (Σχήμα 4.2.).



Σχήμα 4.2.: Η σελίδα της προχωρημένης αναζήτησης (Advanced Search) της AltaVista (www.altavista.com)

Ειδικότερα όσον αφορά την παρουσίαση των αποτελεσμάτων, η AltaVista επιστρέφει μόνο μία σελίδα αποτελέσματος από κάθε σχετιζόμενο ***site***, προσφέροντας έτσι στο χρήστη τη δυνατότητα πρόσβασης σε ένα ευρύτερο σύνολο αποτελεσμάτων. Επίσης, κατά την *ταξινόμηση* ***(ranking)*** των αποτελεσμάτων η AltaVista, ακολουθώντας τις σύγχρονες τάσεις και πρακτικές, τείνει να προβάλλει τα αποτελέσματα που θεωρούνται πιο δημοφιλή ως απαντήσεις στην εκάστοτε ερώτηση που έχει τεθεί και άρα ως τα πιο σχετιζόμενα με το θέμα της αναζήτησης. Άμεσα σχετιζόμενη με τη λειτουργία αυτή είναι και η αύξηση του όγκου των πληροφοριακών πηγών που ευρετηριάζονται από την AltaVista, οι οποίες, σύμφωνα με τα δελτία τύπου της εταιρείας, έφτασε τις 250.000.000 εγγραφές από τις 160.000.000 που συγκεντρώνονταν στην προηγούμενη έκδοση. Εξαιτίας του ότι η εκτίμηση αυτή προέρχεται μόνο από την ίδια την εταιρεία-ιδιοκτήτη της μηχανής και επειδή ακριβώς δεν μπορεί να επιβεβαιωθεί πριν το πέρας συγκεκριμένου χρονικού διαστήματος, οπότε και θα έχουν εκτελεστεί συγκεκριμένες έρευνες, η πληροφορία αυτή διατυπώνεται με κάθε επιφύλαξη. Τέλος, σε ό,τι αφορά τις δυνατότητες αναζήτησης, η AltaVista έχει

ανασχεδιάσει την αναζήτηση στο ***Usenet***, μετονομάζοντας τη σε ***Discussion Search***, και έχει δημιουργήσει αυτόνομο ***link*** που αφορά την αναζήτηση αρχείων γραφικών, κινούμενης εικόνας και ήχου. Εντελώς νέα είναι και η δυνατότητα ***αναζήτησης ειδήσεων (News search)*** σε χρονικό διάστημα που εκτείνεται από 6 ώρες έως και 14 ημέρες. Συνολικά, ο πρόσφατος ανασχεδιασμός της AltaVista πρόσθεσε περισσότερο στην αισθητική εικόνα της μηχανής και λιγότερο σης πραγματικές δυνατότητες αναζήτησης.

ΒΑΣΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Όπως και με τις άλλες μηχανές αναζήτησης έτσι και με την AltaVista μπορούμε να διεξάγουμε είτε απλή είτε προχωρημένη αναζήτηση. Στην απλή αναζήτηση η AltaVista δεν υποστηρίζει τη χρήση των τελεστών Boolean, όμως χρησιμοποιεί έναν συνδυασμό διάφορων μορφών σύνταξης, οι οποίες υποδηλώνουν την *εγγύτητα* ***(proximity)*** ανάμεσα στους αναζητήσιμους όρους, τις αναζητήσιμες φράσεις και λέξεις κλειδιά ***(keywords)*** ή ακόμη τις λέξεις που πρέπει να αναζητηθούν και να ανακτηθούν κατά το ερώτημα, όπως και τις λέξεις που πρέπει οπωσδήποτε να παραλειφθούν. Η AltaVista καλύπτει το web ευρετηριάζοντας το *σύνολο των λέξεων* ***(full text indexing)*** που επισκέπτονται τα προγράμματα σάρωσης και καταγραφής (crawlers, robots, spiders) σε 160-250 εκατομμύρια web pages. Καλύπτει επίσης τα ***newsgroups***, όπου επιπλέον γίνεται ευρετηρίαση του συνόλου των λέξεων που βρίσκονται στα μηνύματα και τις ανακοινώσεις. Τέλος, η AltaVista καλύπτει τις ***αυτόματες λίστες (listservs)***, οι οποίες ασχολούνται με διάφορα θέματα και οι χρήστες μπορούν να δηλώσουν τη συμμετοχή τους, λαμβάνοντας έτσι τα μηνύματα που ανταλλάσσονται μεταξύ των υπόλοιπων μελών της συγκεκριμένης λίστας.

Και στις δύο μορφές αναζητήσεων που διαθέτει υποστηρίζεται η αναζήτηση με τη χρήση τελεστών Boolean, αν και η σύνταξη που χρησιμοποιεί είναι διαφορετική σε κάθε περίπτωση. Η απλή αναζήτηση περιλαμβάνει ένα μενού ***pull down*** και ένα απλό ***πεδίο αναζήτησης (search box)***, στο οποίο ο χρήστης μπορεί να εισαγάγει τους αναζητήσιμους όρους σε 25 διαφορετικές γλώσσες, περιορίζοντας έτσι το πεδίο έρευνας. Αντίστοιχα, στη σελίδα της προχωρημένης αναζήτησης υπάρχουν επίσης το μενού pull down, το πεδίο αναζήτησης υπό τον περιορισμό της γλώσσας καθώς επίσης η δυνατότητα να περιοριστεί η αναζήτηση με βάση την ημερομηνία που ο χρήστης μπορεί να εισαγάγει στο κατάλληλο πεδίο, όπως και η ταξινόμηση των αποτελεσμάτων με βάση το θεματικό *όρο κλειδί* ***(keyword)*** που επιθυμεί ο χρήστης.

Τέλος, η AltaVista επιτρέπει στο χρήστη να διαμορφώσει το ***interface*** της αρεσκείας του και στη συνέχεια να το σώσει έτσι ώστε να είναι στη διάθεση του κάθε φορά που χρησιμοποιεί τη συγκεκριμένη μηχανή αναζήτησης.

Σε ό,τι αφορά τη ***βοήθεια (help files)*** που παρέχεται στο χρήστη, η AltaVista διαθέτει ιδιαίτερα αναλυτικές οδηγίες χρήσης τόσο για την απλή όσο και για την προχωρημένη αναζήτηση, όπως επίσης για τους τρόπους με τους οποίους ο χρήστης μπορεί να ***βελτιώσει (refine)*** τα αποτελέσματα της αναζήτησής του. Επίσης παρέχεται μια σελίδα με τις πιο ***συνηθισμένες ερωτήσεις (frequently asked questions, FAQ)*** που μπορεί κάποιος να αναρωτηθεί κατά τη χρήση της συγκεκριμένης μηχανής.

4.1.2. Αρχιτεκτονική της AltaVista

Η αρχιτεκτονική της AltaVista δίνει τη δυνατότητα προσαρμογής των προτιμήσεων του χρήστη σε αρκετά σημεία. Στην αρχιτεκτονική της AltaVista έχουν ενσωματωθεί τα ***AltaVista Search SDK, AltaVista Index Server*** και ***AltaVista Front End Toolkit (AltaVista Enterprise Search 2.1 Technical Overview, May 2002)***, τα οποία περιγράφονται αμέσως παρακάτω και στη συνέχεια γίνεται η ανάπτυξη της αρχιτεκτονικής της AltaVista.

- **AltaVista Search Software Developer's Kit**

Για οργανισμούς που επιδιώκουν μια ιδιαίτερα προσαρμοσμένη λύση στην πρόσβαση πληροφοριών, διατίθεται από την AltaVista το ***λογισμικό αναζήτησης Software Developer's Kit (SDK)***. Το ***SDK*** επιτρέπει τη δημιουργία εφαρμογών αναζήτησης για τις συγκεκριμένες ανάγκες μιας επιχείρησης. Επιτρέπει επίσης την αποθήκευση σημαντικών και δυναμικών χαρακτηριστικών, τομέων και συνηθισμένων παραγόντων ταξινόμησης εγγράφων σε ένα ευρετήριο – επιτρέποντας αναζητήσεις με συγκεκριμένους τρόπους με τους οποίους οι χρήστες έχουν πρόσβαση στις πληροφορίες. Επιπλέον, η χρήση του SDK μπορεί:

- ✓ **Να δημιουργεί αυτόνομες εφαρμογές αναζήτησης**, παραδείγματος χάριν, να μπορούν οι αγοραστές βιβλίων να ψάχνουν με βάση τον τίτλο, τον συγγραφέα, ή τον εκδότη, και τα αποτελέσματα της αναζήτησης να ταξινομούνται με γνώμονα την τιμή του κάθε βιβλίου.
- ✓ **Ενσωμάτωση δυνατοτήτων αναζήτησης μέσα σε άλλες εφαρμογές**, όπως σε ένα φόρουμ συζήτησης ή στα περιεχόμενα ενός συστήματος διαχείρισης.

- ✓ **Επέκταση της λειτουργίας** της αναζήτησης της AltaVista μέσω των κοινών συλλεκτών στοιχείων ή/και φίλτρων.

Η AltaVista με τη χρήση του SDK αποτελείται από ένα σύνολο βιβλιοθηκών και ***interfaces (Application Programming Interfaces (APIs))*** που βοηθούν στη δημιουργία των εφαρμογών αναζήτησης που λειτουργούν με οποιοδήποτε τύπο πληροφοριών. Το SDK παρέχει:

- ✓ **Το Ευρετήριο API** - δημιουργεί το ευρετήριο και προσθέτει τα έγγραφα σε αυτό το ευρετήριο
- ✓ **Το Ερώτημα API** - εκτελεί τα ερωτήματα
- ✓ **Τη Γλωσσολογία API** - επιτυγχάνει την ανίχνευση φράσεων, τη διόρθωση των συλλαβών, τον τρόπο γραφής και την εύρεση συνωνύμων σε μια εφαρμογή αναζήτησης. Η γλωσσολογία API περιέχει, προεπιλεγμένα λεξικά σε 20 γλώσσες.
- ✓ **Την Ομάδα Εγγράφων (Docbundle) API** - προσθέτει έγγραφα σε ένα ευρετήριο που δημιουργήθηκε και ρυθμίστηκε από την AltaVista. Χρησιμοποιώντας αυτά τα APIs δημιουργούνται ολοκληρωμένες πηγές στοιχείων. Το Docbundle API επιτρέπει την υποκλοπή των εγγράφων και την τροποποίησή τους πριν τη σύνταξη ευρετηρίου.
- ✓ **Μετατροπή εγγράφων API** – μετατροπή των πληροφοριών σε περισσότερα από 225 είδη αρχείων σε κείμενο ή HTML για αναζήτηση.
- ✓ **C, Java, COM**, **Perl** και **Tcl** API interfaces.
- ✓ **Παραδείγματα Προγραμμάτων**, συμπεριλαμβανομένης της τεκμηρίωσής τους και ολόκληρου του πηγαίου κώδικα.

- ➢ **AltaVista Index Server**

Το τμήμα του ***Index Server*** παρέχεται ως τμήμα της εγκατάστασης του SDK και επιτρέπει την εύκολη και ολοκληρωμένη αναζήτηση καθώς και τη λειτουργία της ευρετηρίασης. Ο Index Server έχει υλοποιηθεί σε java και υποστηρίζει και την ευρετηρίαση και την αναζήτηση. Είναι μια μηχανή ευρετηρίασης και ερωτήσεων που έχει σχεδιαστεί και έχει βελτιστοποιηθεί για τη γρήγορη και αξιόπιστη επέκταση της τεχνολογίας αναζήτησης της AltaVista. Ο Index Server έχει τα ακόλουθα χαρακτηριστικά:

- ✓ Αναπτύχθηκε ως στοιχείο της ***Java Bean*** ή ως ανεξάρτητη υπηρεσία
- ✓ Η πρόσβασή του γίνεται τοπικά μέσω του ***XML*** και για απομακρυσμένα σημεία

μέσω ***SOAP***

- ✓ Διαμόρφωση των υπηρεσιών του δικτύου για τη λήψη των ***SOAP*** αιτημάτων από απομακρυσμένες πηγές εισόδου
- ✓ Πλήρη υποστήριξη όλων των SDK χαρακτηριστικών ευρετηρίασης
- ✓ Τα αποτελέσματα επιστρέφονται ως ***XML*** για πιο εύκολη πρόσβαση και εμφανίζονται ως ***XSLT***
- ✓ Πλοήγηση για την ενημέρωση και τη χρήση του ***XML*** είναι διαθέσιμα
- ✓ Παραδείγματα εφαρμογών χρήσης του Index Server ως ενσωματωμένο συστατικό της Java Bean.

Με την ενσωμάτωση του ευρετηρίου και της αναζήτησης σε ένα ενιαίο σύστημα και ως μια ενιαία λειτουργία, ο Index Server μειώνει δραματικά το χρόνο και την απαιτούμενη προσπάθεια για την ανάπτυξη μιας συνηθισμένης εφαρμογής αναζήτησης.

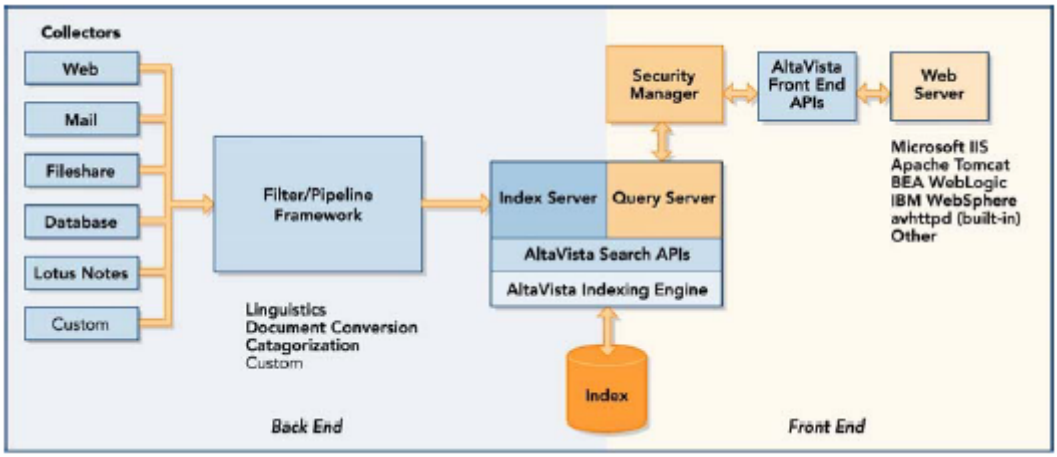
- ➢ **AltaVista Front End Toolkit**

Το ***AltaVista Front End Toolkit (AVFE)*** είναι ένα σύστημα, βασισμένο σε συστατικά μέρη, που επιτρέπει εύκολα την ενσωμάτωση της εφαρμογής της αναζήτησης με τη λειτουργία του server στο περιβάλλον εργασίας του καθενός. Οι βιβλιοθήκες (Java Servlet, JSP, C και COM interfaces) μπορεί να εγκατασταθούν είτε με το ***pre-packaged Alta Vista Search Application*** είτε με τα στοιχεία του Index Server. Το AltaVista Front End Toolkit έχει τις ακόλουθες δυνατότητες:

- ✓ Βιβλιοθήκες που αναπτύσσουν παραμετροποιημένα interface αναζήτησης (C, Java, JSP και COM interfaces)
- ✓ Γλωσσολογικές δυνατότητες (συλλαβισμός, συνώνυμα, παρενθέσεις και αναγνώριση εκφράσεων) που επεκτείνουν τις ερωτήσεις του χρήστη.
- ✓ Μια τοπική σελίδα αναζήτησης για χρήση από έναν διεθνή χρήστη της βάσης
- ✓ Υποστηρίζει την κατηγοριοποίηση και τα χαρακτηριστικά ασφαλείας που παρέχονται από την AltaVista Search Application.
- ✓ Τονίζει και κάνει πιο ευκρινείς τους όρους της αναζήτησης κατά την εμφάνιση των αποτελεσμάτων.
- ✓ Συμβατό και με την AltaVista Search Application και με τα στοιχεία του Index Server.
- ✓ Περιέχει παραδείγματα εφαρμογών τα οποία αποδεικνύουν τις πρόσθετες κατηγοριοποιήσεις.

Ενσωματώνοντας στην AltaVista τα ***AltaVista Search SDK, AltaVista Index Server*** και ***AltaVista Front End Toolkit*** δίνεται η δυνατότητα:

- ✓ Δημιουργίας και αναβάθμισης των συνηθισμένων προγραμμάτων συλλογής δεδομένων που συγκεντρώνουν πληροφορίες και τις αποθηκεύουν σε αποθήκες που δεν υποστηρίζονται απευθείας από την εφαρμογή αναζήτησης της AltaVista.
- ✓ Πρόσθετων προγραμμάτων συλλογής δεδομένων που βοηθούν στον εντοπισμό αντιπροσωπευτικότερων πληροφοριών.
- ✓ Δημιουργίας και ολοκλήρωσης των συνηθισμένων μετατροπέων αρχείων διαχειρίζοντας πληροφορίες ειδικών τύπων που δεν περικλείονται στους 225 τύπους αρχείων που υποστηρίζονται ως τώρα.
- ✓ Ανάπτυξης φίλτρων για την πληροφορία που υπάρχει στο ευρετήριο.
- ✓ Δημιουργίας ενός προσαρμοζόμενου interface ερωτήματος και τροποποίηση των αποτελεσμάτων του αρχικού interface αναζήτησης.



Σχήμα 4.3.: Η αρχιτεκτονική της AltaVista (AltaVista Enterprise Search 2.1 Technical Overview, May 2002)

- ✓ Ολοκλήρωση του ***AltaVista Highlighter*** παρέχοντας καλύτερα αποτελέσματα αναζήτησης και μετατροπή των αρχείων σε HTML.
- ✓ Χρήση της αυτόματης κατηγοριοποίησης με σκοπό την αυτόματη ταξινόμηση των εγγράφων σε κατηγορίες. Οι ταξινομήσεις καθορίζονται μέσω ενός interface διαχείρισης και το σύνολο των κανόνων μπορεί να δημιουργηθεί χειρωνακτικά ή μέσω της στατιστικής ανάλυσης των υπό δοκιμή συνόλων.

Ένα άλλο επίσης πολύ σημαντικό ζήτημα που η AltaVista έχει επιλύσει είναι η

Ασφαλής Πρόσβαση στα Αποτελέσματα Αναζήτησης ***(Securing Access to Search Results)***. Η AltaVista παρέχει δυνατότητες ασφάλειας των αποτελεσμάτων μιας αναζήτησης. Αρκετά συχνά μερικές πληροφορίες δεν πρέπει να είναι προσιτές σε όλους τους χρήστες επομένως, δεν πρέπει να επιστρέφονται σε όλους τους χρήστες. Υπάρχουν διάφορες προγραμματιστικές μέθοδοι φιλτραρίσματος των αποτελεσμάτων αναζήτησης που βασίζονται στην τροποποίηση της ερώτησης ή/και φιλτράροντας τα αποτελέσματα κατά την εμφάνισή τους. Όταν αλλάζουν τα προνόμια ενός χρήστη ή σύνθετες πολιτικές πρέπει να εφαρμοστούν, η απλή τροποποίηση της ερώτησης δεν αρκεί. Για να υποστηρίξει αυτούς τους σύνθετους ελέγχους πρόσβασης, η AltaVista παρέχει τις ακόλουθες μεθόδους για την ασφαλή πρόσβαση στα αποτελέσματα της αναζήτησης:

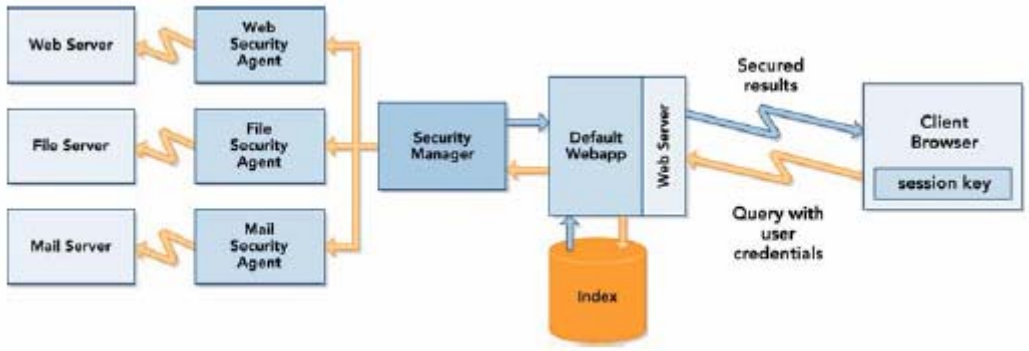
- **Ασφάλεια σε Επίπεδο Ευρετηρίου** ***(Index-level Security):*** Περιορίζει το ευρετήριο στο οποίο ένας συγκεκριμένος χρήστης μπορεί να ψάξει.
- **Σε Πραγματικό Χρόνο Ασφάλεια σε Επίπεδο Εγγράφου** ***(Real-time document-level security):*** Τα όρια αναζήτησης επεκτείνονται μόνο στα έγγραφα που ο χρήστης έχει το δικαίωμα να έχει πρόσβαση. Αυτός είναι ένας εξαιρετικά εύκολος και ισχυρός τρόπος να υπάρξει ασφάλεια σε περιβάλλοντα όπου οι άδειες πρόσβασης είναι πολύ μπερδεμένες (έγγραφο σε επίπεδο χρηστών) και μπορεί να αλλάζουν συχνά.

Η **ασφάλεια σε επίπεδο ευρετηρίου** ***(Index-level Security)*** επιτρέπει την πρόσβαση στους μεμονωμένους χρήστες σε ευρετήρια που τους επιτρέπονται. Απαιτεί τον προσδιορισμό των χρηστών (και των χαρακτηριστικών τους) που επικυρώνεται από τον ***web server***, και από τον κατάλογο των εξουσιοδοτημένων ευρετηρίων για κάθε χρήστη. Όταν το ***Index-level Security*** ενεργοποιείται το ***interface*** της αναζήτησης απαριθμεί μόνο εκείνα τα ευρετήρια που ο χρήστης έχει πρόσβαση.

Η **σε πραγματικό χρόνο ασφάλεια σε επίπεδο εγγράφου** ***(Real-time document-level security)*** ουσιαστικά ελέγχει εάν ένας χρήστης έχει ή όχι πρόσβαση σε ένα έγγραφο τη στιγμή της ερώτησης. Η AltaVista προσπαθεί μέσω κατάλληλων διεργασιών να επικυρώσει και να εγκρίνει εάν ένας χρήστης έχει την άδεια να έχει πρόσβαση σε κάθε έγγραφο που περιλαμβάνεται στα αποτελέσματα που ταιριάζουν με την ερώτηση του χρήστη. *Ο διαχειριστής ασφαλείας* ***(Security Manager)*** αλληλεπιδρά με το ***interface*** της αναζήτησης (AVFE) και τους *πράκτορες ασφάλειας της αναζήτησης*

(Search security agents) της AltaVista για να εξασφαλίσει την πρόσβαση στα προστατευμένα έγγραφα. Οι αναφορές σε οποιαδήποτε έγγραφα που ο χρήστης δεν είναι εξουσιοδοτημένος να έχει πρόσβαση δεν εμφανίζονται στα αποτελέσματα αναζήτησης.

Όταν ένας χρήστης θέτει μια ερώτηση, τα αποτελέσματα οργανώνονται μέσω ενός φίλτρου ασφάλειας. Αρχικά, η πιστοποίηση του χρήστη λαμβάνεται από τον διαχειριστή ασφαλείας. Κατόπιν, για κάθε έγγραφο, ο κατάλληλος πράκτορας ασφάλειας επιλέγεται βασισμένος στα έγγραφα ***URI***, ο οποίος ελέγχει την απομακρυσμένη πηγή πληροφοριών για να ελέγξει εάν ο χρήστης είναι εξουσιοδοτημένος ή όχι για να δει το έγγραφο. Εάν το έγγραφο είναι προσβάσιμο (ή δεν χαρακτηρίζεται ως ασφαλισμένο) προστίθεται στη λίστα των αποτελεσμάτων που θα επιστραφεί στο χρήστη.



Σχήμα 4.4.: Ο διαχειριστής ασφαλείας αλληλεπιδρά με τους πράκτορες ασφαλείας στη διαδικασία της αναζήτησης με σκοπό την ελεγχόμενη πρόσβαση (AltaVista Enterprise Search 2.1 Technical Overview, May 2002)

Η διαδικασία της έγκρισης της αυθεντικότητας για το σύνολο των αποτελεσμάτων έχει επιπτώσεις στην απόδοση της αναζήτησης, και περιορίζεται από την απόδοση του υπάρχοντος συστήματος ασφάλειας της αποθήκευσης των εγγράφων. Η AltaVista περιορίζει τον αντίκτυπο με τη διατήρηση μιας διαμορφώσιμης λίστας εγγράφων και των αποτελεσμάτων πιστοποίησης των προηγούμενων ερωτήσεων του χρήστη.

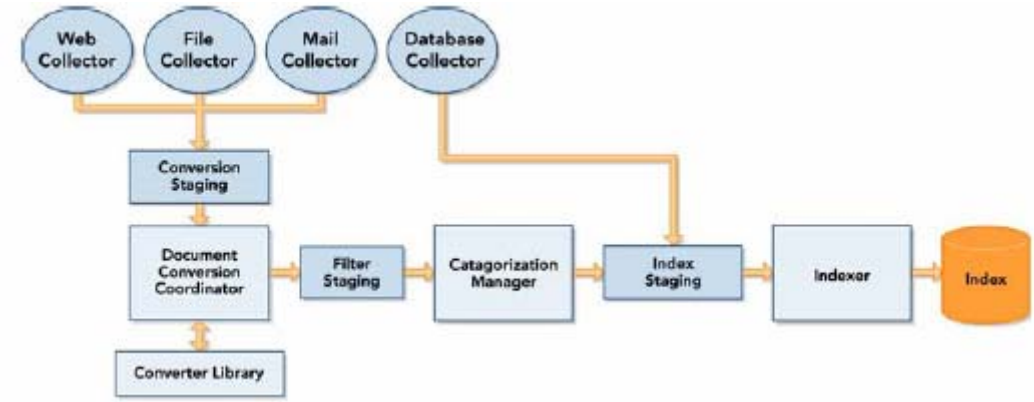
4.1.3. Crawling

- Ο Webmaster μπορεί να υποβάλει ιστοσελίδες στην AltaVista για ευρετηρίαση. Έναν περίπου μήνα μετά την υποβολή της σελίδας, η AltaVista θα επισκεφτεί την ιστοσελίδα και θα ψάξει και για τυχόν άλλες σελίδες για να προστεθούν. Κατόπιν επιστρέφει κάθε τέσσερις έως έξι εβδομάδες για να ελέγξει για τυχόν αλλαγές και για νέες σελίδες.
- Η AltaVista συλλέγει πολλές από τις σελίδες του ***web***. Δεν έχει κανένα ιδιαίτερο όριο, και πολλές ιστοσελίδες έχουν εκατοντάδες ή χιλιάδες σελίδες απαριθμημένες.

Εάν οι σελίδες κλειδιά ***(key pages)*** λείπουν για κάποιους λόγους, μπορεί να καταχωρηθούν χειρωνακτικά (Bikkannavar, 1999).

4.1.4. Ευρετήριο (Indexing)

Η διαδικασία της σύνταξης του ευρετηρίου της AltaVista είναι απλή και κατανοητή.



Σχήμα 4.5.: Συλλογή, μετατροπή και σύνταξη ευρετηρίου (AltaVista Enterprise Search 2.1 Technical Overview, May 2002)

1. Αρχικά τα έγγραφα μαζεύονται από την απομακρυσμένη πηγή στοιχείων με χρήση των καθορισμένων ***συλλεκτών (ειδικό λογισμικό- collectors)*** για κάθε ευρετήριο. Αυτά τα έγγραφα συλλέγονται σε ξεχωριστές ομάδες ***(docbundle)*** (παραδείγματος χάριν, τα διάφορα αρχεία HTML θα συλλεχθούν σε ένα ενιαίο docbundle) για γρηγορότερη επεξεργασία κατά ομάδες και στη συνέχεια για μετατροπή.

2. Ο *υπεύθυνος της μετατροπής* ***(Conversion Manager)*** ελέγχει έναν κατάλογο, και

επεξεργάζεται τα docbundles καθώς τοποθετούνται σε αυτόν τον κατάλογο. Αμέσως μόλις το περιεχόμενο των σύνθετων εγγράφων (MS-WORD, Excel, κ.τ.λ.) έχει μετατραπεί σε κείμενο, οδηγούνται στο *διαχειριστή κατηγοριοποίησης* ***(Categorization Manager)***.

3. Εάν είτε η μονάδα RegExCategorizer είτε η μονάδα AltaVista AutoCategorizer ενεργοποιηθεί, το περιεχόμενο κάθε εγγράφου εξετάζεται και ταξινομείται σε μια ή περισσότερες κατηγορίες. Η ταξινόμηση βασίζεται σε ένα σύνολο κανόνων στη μονάδα κατηγοριοποίησης ***(Categorization Module)***.
4. Τέλος, το docbundle οδηγείται στον ***Indexer***, το οποίο προσθέτει το έγγραφο στο ευρετήριο.

Σε οποιοδήποτε σημείο αυτής της διαδικασίας, τα docbundles μπορεί να παρεμποδιστούν και να υποβληθούν σε επεξεργασία χρησιμοποιώντας το Docbundle API ή με άμεση τροποποίηση των κειμένων των αρχείων. Αυτό επιτρέπει στους υπεύθυνους για την ανάπτυξη αυτής της διαδικασίας την πρόσθεση φίλτρων ή συλλεκτών σε οποιοδήποτε στάδιο της σύνταξης ευρετηρίου.

Οι γενικοί κανόνες που εφαρμόζονται από την AltaVista για τη σύνταξη του ευρετηρίου είναι οι παρακάτω:

1. Η AltaVista δεν συντάσσει ευρετήριο λαμβάνοντας υπ' όψιν της τα σημεία στίξης, έτσι η φράση "webmaster's guide" γίνεται "web master s guide".
2. Ο τονισμός επηρεάζει μόνο εάν ο χρήστης τον χρησιμοποιήσει.
3. Τα σχόλια δεν ευρετηριάζονται.
4. Μόνο τα πρώτα 100Kb του κειμένου σε μια σελίδα ευρετηριάζεται. Μετά από αυτό μόνο οι συνδέσεις προς άλλες σελίδες συντάσσονται, μέχρι ένα μέγιστο, τα 4MB. Δεδομένου ότι οι περισσότερες ιστοσελίδας καταλαμβάνουν κάτω από 100Kb, αυτοί οι περιορισμοί δεν είναι πρόβλημα για τους περισσότερους webmasters.
5. Οι σελίδες με γράμματα κειμένου πολύ μικρά σε μέγεθος μπορεί να μη ληφθούν υπ' όψιν.

Εκτός όμως από τη διαδικασία που ακολουθεί και τους κανόνες ευρετηρίασης που εφαρμόζει η AltaVista, θα αναφερθούμε και στους *συλλέκτες, το ειδικό λογισμικό των πληροφοριών* ***(collectors)***. Ένας *συλλέκτης* ***(collector)*** είναι ένα μέρος της εφαρμογής αναζήτησης AltaVista που μπορεί να συνδεθεί με μια πηγή στοιχείων (όπως μια βάση δεδομένων, έναν mail server, ή μια ιστοσελίδα) και να εξάγει πληροφορίες

ευρετηρίασης. Η AltaVista έχει τους συλλέκτες στοιχείων που παρατίθενται στον πίνακα 4.1.

Οι συλλέκτες προστίθενται σε ένα ευρετήριο χρησιμοποιώντας το ***Management User Interface***. Χρησιμοποιώντας ο ***administrator*** το ***wizard*** του κάθε συλλέκτη διαμορφώνει τις παραμέτρους του και επιτρέπει το συντονισμό της διαδικασίας συλλογής στοιχείων. Σε άλλες περιπτώσεις, μπορεί να είναι επιθυμητή η μεγιστοποίηση της ταχύτητας συλλογής προκειμένου να ολοκληρωθεί η συλλογή μέσα από ένα συγκεκριμένο παράθυρο.

Ως τμήμα καθορισμού της διαδικασίας συλλογής, υπάρχει η δυνατότητα καθορισμού ενός προγράμματος για την έναρξη της διαδικασίας συλλογής. Αυτό το πρόγραμμα μπορεί να καθοριστεί να τρέχει κάθε ημέρα μια συγκεκριμένη στιγμή, ή να επαναλαμβάνεται κάποια ημέρα ή ημερομηνία. Μπορεί επίσης να σταματήσει ο συλλέκτης σε έναν ορισμένο χρόνο ή μετά την ολοκλήρωση της διαδικασίας.

Ειδικό Λογισμικό	**Τύπος Δεδομένων**	**Πλατφόρμα Συστήματος**
Συλλέκτης Ιστού (web collector)	HTML και άλλα έγγραφα που βρίσκονται στους web servers, περιέχοντας spreadsheets, επεξεργασία κειμένου, PDF αρχεία, κ.τ.λ. Ο Web Collector της AltaVista είναι ο γρηγορότερος, ο πιο ικανός, ο πιο ευφυής και δυναμικός Web crawler που υπάρχει στην αγορά σήμερα	Υποστηρίζει τα Windows και UNIX
Συλλέκτης Αρχείων (File Collector)	Αρχεία που βρίσκονται στο δίκτυο για κοινή χρήση	Υποστηρίζει τα Windows μόνο
Συλλέκτης Βάσεων Δεδομένων (Database Collector)	Περιεχόμενα σχεσιακών βάσεων, περιεχόμενα ευρετηρίου που τα διαχειρίζονται οι RDBMS, όπως οι Oracle, SQLServer, Informix, DB2, κ.τ.λ.	Υποστηρίζει τα Windows και UNIX
Συλλέκτης των Mail (Mail Collector)	Συλλέγει και ευρετηριάζει πληροφορίες από τους IMAP mail servers. Ευρετηριάζει και τα κοινά και τα ιδιωτικά αρχεία που ευρετηριάζονται από τους mail servers όπως το Microsoft Exchange	Υποστηρίζει τα Windows μόνο
Άλλοι Συλλέκτες	Συλλέκτες για διαφορετικούς αποθηκευτικούς χώρους, όπως τους Documentum και Lotus Notes, είναι διαθέσιμα από το λογισμικό της AltaVista	Δίαφορα

Πίνακας 4.1.: Οι συλλογείς στοιχείων της AltaVista (AltaVista Enterprise Search 2.1 Technical Overview, May 2002)

- ✓ Ο **Συλλέκτης Ιστού (Web Collector)** είναι ένας πολύπλοκος και υψηλής απόδοσης ***web crawler*** που υποστηρίζει πλήρως το ***HTTP*** και τα ***Πρωτόκολλα HTTPS***. Η περιήγησή του μπορεί να αρχίσει με την άμεση πρόσβαση σε απομακρυσμένες ιστοσελίδες και τοποθεσίες ή μέσω ενός ***proxy server***. Είναι δυνατή η παραμετροποίηση του *συλλέκτη* ***(collector)*** ώστε να ευρετηριάζει συγκεκριμένα ***META tags*** ως τομείς μέσα στο ευρετήριο της AltaVista. Για παράδειγμα, εάν οι σελίδες περιλαμβάνουν ένα meta tag ***"author"***, θα μπορούσε να διατυπωθεί μια ερώτηση για τον εντοπισμό όλων των εγγράφων του ***"author: Shakespeare"***. Κατά τη διάρκεια της επανευρετηρίασης, μόνο εκείνα τα έγγραφα που έχουν τροποποιηθεί θα ενημερωθούν στο ευρετήριο.
- ✓ Ο **Συλλέκτης Αρχείων (File Collector)** μπορεί να χρησιμοποιηθεί για το πέρασμα και τον ευρετηριασμό των αρχείων από την κοινή χρήση των αρχείων ***(fileshares)*** στην πλατφόρμα των ***windows***. Το *wizard του συλλέκτη αρχείων* ***(File Collector wizard)*** παρέχει ένα εργαλείο περιήγησης του δικτύου για την επιλογή του περιεχομένου και των τύπων των αρχείων που περιλαμβάνονται στο ευρετήριο.
- ✓ Ο **Συλλέκτης των Mail (Mail Collector)** επιτρέπει την πρόσβαση και τη σύνταξη ευρετηρίου των μηνυμάτων του ηλεκτρονικού ταχυδρομείου (και των *επισυναπτόμενων αρχείων* ***(attachments)*** τους) που αποθηκεύονται στους ***mail servers***. Ο συλλέκτης των ***Mail*** χρησιμοποιεί το ***πρωτόκολλο IMAP*** για πρόσβαση στα περιεχόμενα του ***server***. Αυτό επιτρέπει τη σύνταξη ευρετηρίου mail από το ***Microsoft Exchange*** και το ***Lotus Notes mail servers***. Ο συλλέκτης ευρετηριάζει το κύριο μέρος των μηνυμάτων καθώς επίσης και τις κεφαλίδες των mail (από ποιο οργανισμό στάλθηκε και το θέμα του) και τις ημερομηνίες για περισσότερο εύστοχα ερωτήματα (τα *επισυναπτόμενα αρχεία* κληρονομούν τις meta πληροφορίες από το μήνυμα που εσωκλείεται).
- ✓ Χρησιμοποιώντας το **Συλλέκτη Βάσεων Δεδομένων (Database Collector)**, μπορεί να ευρετηριαστεί το περιεχόμενο ενός πίνακα ή να πάρει πληροφορίες από ένα παρόμοιο σύστημα βάσεων δεδομένων (RDBMS). Το περιεχόμενο κάθε επιλεγμένης στήλης του πίνακα ευρετηριάζεται σε χωριστούς τομείς (βασισμένοι στο όνομα των στηλών). Υπάρχουν μερικοί τρόποι που ευρετηριάζονται τα έγγραφα ξανά: συλλέγουν όλες τις σειρές και ενημερώνουν μόνο εκείνες τις σειρές που έχουν αλλάξει (βασισμένοι σε έναν κατάλογο ή μια timestamp στήλη). Παρέχει επίσης έναν τρόπο απομακρύνσεων των παλαιών αρχείων του ευρετηρίου.

4.1.5. Ταξινόμηση (Ranking)

Η AltaVIsta χρησιμοποιεί τους ακόλουθους κανόνες για την ταξινόμηση των σελίδων (Bikkannavar, 1999):

1. Πόσες φορές οι λέξεις του ερωτήματος του χρήστη εμφανίζονται
2. Εάν οι λέξεις του ερωτήματος του χρήστη βρίσκονται στους τίτλους ή στα ***meta-tags***
3. Η εγγύτητα των λέξεων του ερωτήματος του χρήστη στο έγγραφο

Κάποιοι άλλοι κανόνες είναι:

1. Εάν όλοι οι όροι που αναζητούνται εμφανίζονται, ή μόνο μερικοί από αυτούς.
2. Εάν οι όροι της αναζήτησης εμφανίζονται στον τίτλο, στα metatags ή στις πρώτες γραμμές της σελίδας
3. Εάν οι όροι της αναζήτησης εμφανίζονται συχνά στο έγγραφο.
4. Η AltaVista δίνει μεγαλύτερη έμφαση στις διαδρομές των αρχικών σελίδων, εν' αντιθέσει των τεχνικών μιας μεμονωμένης λέξης και των δημοφιλών αναζητήσεων.
5. Μικροί, προσαρμοζόμενοι τίτλοι, χρήση των φράσεων στις λέξεις κλειδιά των metatags, και ελάχιστος αριθμός επαναλήψεων των λέξεων στα metatags.
6. Ταξινομεί υψηλότερα τις πιο παλιές σελίδες, όταν υπάρχουν διάφορες σελίδες με τον ίδιο βαθμό σχετικότητας. Αυτό δε γίνεται σε κάθε περίπτωση, αλλά συμβαίνει συχνά.
7. Η AltaVista χρησιμοποιεί επίσης την *αυτόματη έρευνα φράσης* ***(automatic phrase searching)***. Η υπηρεσία έχει ένα λεξικό με αρκετά εκατομμύρια φράσεις, δημιουργώντας τη χρήση μιας αυτοματοποιημένης διαδικασίας που χρησιμοποιεί γλωσσικά στοιχεία για τον καθορισμό της φράσης. Κατά συνέπεια, οι σελίδες όπου οι όροι της αναζήτησης εμφανίζονται στη σελίδα με τη σειρά που υπάρχουν στο ερώτημα είναι πιθανότερο να εμφανιστούν από εκείνες που απλά τους περιέχουν, αλλά όχι με τη συγκεκριμένη σειρά.

Υπάρχουν βέβαια και εξαιρέσεις για όλους αυτούς τους παράγοντες.

4.1.6. Spamming

Τα παρακάτω η AltaVista τα θεωρεί spamming:

- Υποβάλλοντας την URL επανειλημμένα οποιαδήποτε ημέρα.

- Υποβολή ενός μεγάλου αριθμού URLs από την ίδια περιοχή οποιαδήποτε ημέρα.
- Υποβολή των ίδιων σελίδων από την ίδια περιοχή.
- Επαναλαμβάνοντας τις λέξεις κλειδιά επανειλημμένως, χωρίς λόγο.
- Γεμίζοντας τις σελίδες με λέξεις-κλειδιά ανεξάρτητες όμως από το πραγματικό περιεχόμενο της σελίδας.
- Χρησιμοποίηση αόρατου ή πάρα πολύ μικρού κειμένου για να το διαβάσει.

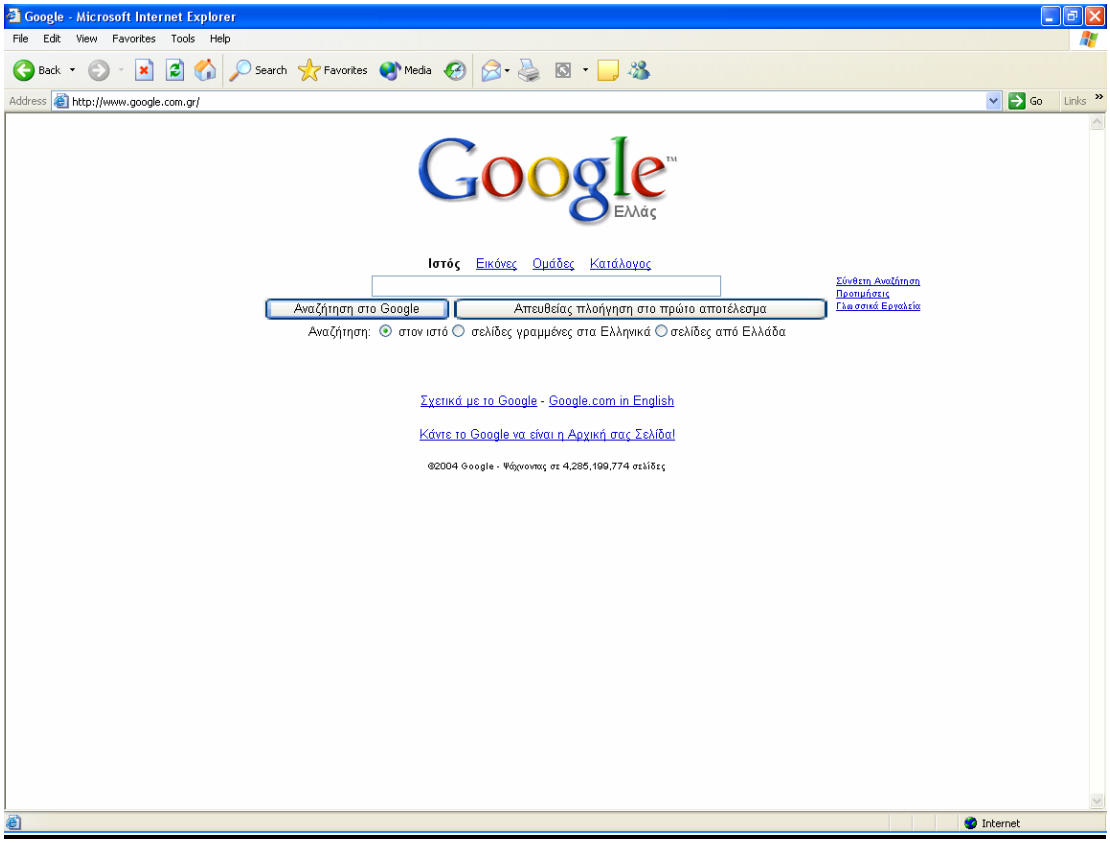
4.1.7. Γλωσσική Ανίχνευση

Η AltaVista κατηγοριοποιεί αυτόματα τις ιστοσελίδας με βάση τη γλώσσα. Το spider αυτής προσπαθεί να καθορίσει τη γλώσσα της ιστοσελίδας την ίδια στιγμή που την ανιχνεύει. Η τεχνολογία είναι *βασισμένη σε λεξικό* ***(dictionary-based)***. Η AltaVista εξετάζει μια σελίδα για να δει εάν ο όγκος των λέξεων ταιριάζει με εκείνους μιας συγκεκριμένης γλώσσας.

4.2. Google (www.google.com)

4.2.1. Εισαγωγή

Η Google είναι μια μεγάλης κλίμακα μηχανή αναζήτησης, η οποία υπερνικά πολλά από τα προβλήματα των υπαρχόντων συστημάτων. Κάνει μεγάλη χρήση της πρόσθετης δομής που υπάρχει στο υπερκείμενο για την παροχή ποιοτικότερων αποτελεσμάτων. Ο στόχος του συστήματος είναι να εξεταστούν πολλά από τα προβλήματα, και στην ποιότητα και στην εξέλιξη, που εμφανίζονται με την κλιμακωτή εφαρμογή της τεχνολογίας της μηχανής αναζήτησης σε τέτοιους μεγάλους αριθμούς. Η Google θεωρείται ότι έχει συντάξει ευρετήριο με περίπου 85 εκατομμύρια ιστοσελίδες, το οποίο είναι περίπου 10,6% του συνολικού ιστού. Η αρχική σελίδα του Google φαίνεται στο σχήμα 4.6.



Σχήμα 4.6.: Η αρχική σελίδα (Homepage) της Google (www.google.com)

Και ο ρυθμός ανάπτυξης του ιστού και οι τεχνολογικές αλλαγές εξετάζονται στο σχεδιασμό της Google. Η Google σχεδιάστηκε να ανταποκριθεί σε εξαιρετικά μεγάλους όγκους δεδομένων. Κάνει αποδοτική χρήση του χώρου αποθήκευσης για την

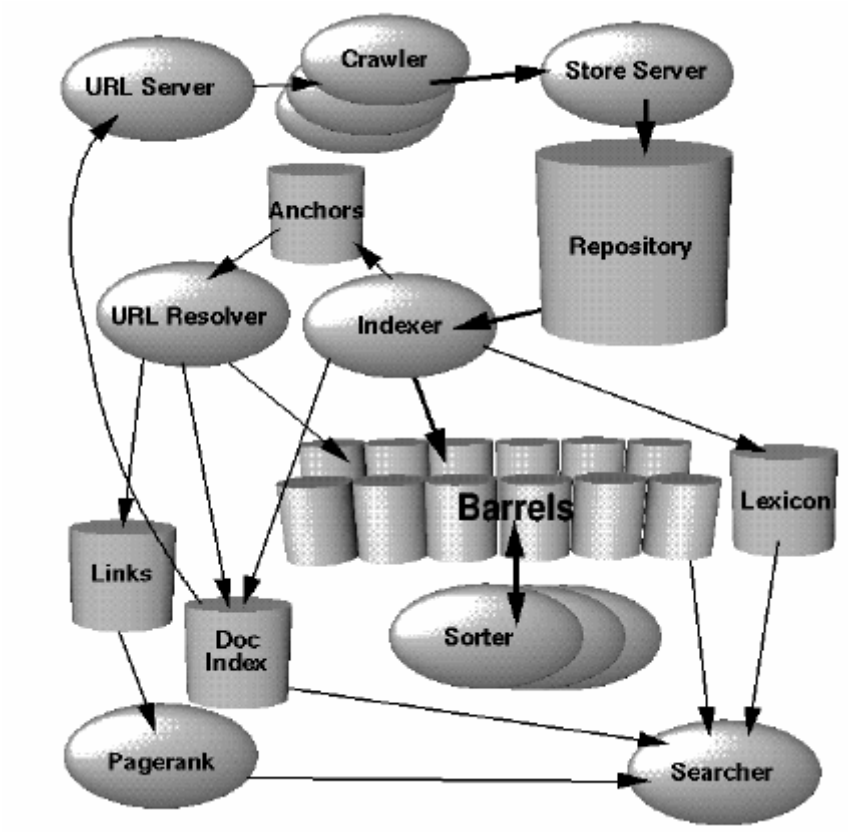
αποθήκευση του ευρετηρίου. Οι δομές δεδομένων βελτιστοποιούνται για γρήγορη και αποδοτική πρόσβαση. Ο κύριος στόχος της μηχανής αναζήτησης Google είναι να βελτιωθεί η ποιότητα της αναζήτησης του ιστού.

Η μηχανή αναζήτησης Google έχει δύο σημαντικά χαρακτηριστικά που την βοηθούν να παραγάγει υψηλή ακρίβεια στα αποτελέσματα της αναζήτησης:

- Χρησιμοποιεί τη *δομή συνδέσεων* ***(link structure)*** του ιστού για να υπολογίσει την ποιότητα και έτσι να ταξινομήσει κάθε ιστοσελίδα. Αυτή η ταξινόμηση καλείται *ταξινόμηση σελίδας* ***(PageRank)***.
- Χρησιμοποιεί τις *αναφορές σε σελίδες* ***(links)*** για την βελτίωση των αποτελεσμάτων της αναζήτησης.

4.2.2. Αρχιτεκτονική του Google

Το παρακάτω σχήμα 4.7. (Huang) δείχνει τη δομή του συστήματος της μηχανής αναζήτησης Google. Το μεγαλύτερο μέρος της Google έχει προγραμματισθεί σε C ή C++ για μεγαλύτερη αποδοτικότητα και μπορεί να τρέξει σε *Solaris* ή *Linux*. Στη Google, διάφορα κατανεμημένα προγράμματα *crawlers* εκτελούν την ανίχνευση του ιστού. Υπάρχει ένας ***URL server*** ο οποίος στέλνει λίστες από *URLs* που θα πρέπει να ανιχνευθούν από τους *crawlers*. Οι ιστοσελίδες αυτές που προσκομίζονται στέλνονται έπειτα στον ***storeserver***. Ο *storeserver* συμπιέζει έπειτα και αποθηκεύει τις ιστοσελίδες σε μια αποθήκη. Κάθε ιστοσελίδα έχει έναν σχετικό κωδικό αριθμό αποκαλούμενο ***docID***, ο οποίος ορίζεται όποτε ένα νέο *URL* λαμβάνεται από μια ιστοσελίδα. Ο ***indexer*** και ο ***sorter*** εκτελούν την ευρετηρίαση. Ο ***indexer*** εκτελεί διάφορες λειτουργίες. Διαβάζει την αποθήκη, αποσυμπιέζει τα έγγραφα, και τα αναλύει. Κάθε έγγραφο μετατρέπεται σε ένα σύνολο από εμφανίσεις λέξεων που ονομάζονται ***hits***. Τα *hits* καταγράφουν τη λέξη, τη θέση της στο έγγραφο, μια προσέγγιση του μεγέθους των χαρακτήρων της, και αν είναι γραμμένη με κεφαλαία γράμματα. Ο ***indexer*** διανέμει αυτά τα *hits* σε ένα *σύνολο βαρελιών* ***(set of barrels)***, που δημιουργεί έναν μερικώς *ταξινομημένο προς τα μπρος ευρετήριο* ***(sorted forward index)***. Ο ***indexer*** εκτελεί και μια άλλη σημαντική λειτουργία. Αναλύει όλες τις συνδέσεις κάθε ιστοσελίδας και αποθηκεύει σημαντικές πληροφορίες για αυτές σε ένα αρχείο ***anchors***. Αυτό το αρχείο περιέχει αρκετές πληροφορίες για τον καθορισμό της κάθε σύνδεσης, από πού ξεκινά και που καταλήγει, και για το κείμενο της σύνδεσης.



Σχήμα 4.7.: Αρχιτεκτονική Google (Huang)

Ο ***URL resolver*** διαβάζει το αρχείο των ***anchors*** και μετατρέπει τις σχετικές *URLs* σε πραγματικές *URLs* και στη συνέχεια *σε* ***docIDs***. Βάζει το ***anchor*** κείμενο στον μπροστινό δείκτη, που συνδέεται με το *docID* στο οποίο δείχνει το ***anchor***. Δημιουργεί επίσης μια βάση δεδομένων των συνδέσεων, οι οποίες είναι ζευγάρια *docIDs*. Η βάση αυτή των συνδέσεων χρησιμοποιείται για τον υπολογισμό των *PageRanks* όλων των εγγράφων.

Ο ***sorter*** παίρνει τα βαρέλια, που ταξινομούνται με βάση το *docID*, και δημιουργεί από το ***wordID*** τον *αντίστροφο δείκτη* ***(inverted index)***. Αυτό γίνεται γιατί απαιτείται προσωρινή μικρή δέσμευση χώρου. Ο ***sorter*** παράγει επίσης έναν κατάλογο από *wordIDs* και τον συμψηφίζει με τον *αντίστροφο δείκτη*. Ένα πρόγραμμα που ονομάζεται ***DumpLexicon*** παίρνει αυτόν τον κατάλογο μαζί με το λεξικό που παράγει ο ***indexer*** και παράγει ένα νέο λεξικό που θα χρησιμοποιηθεί από τον *ερευνητή* ***(Searcher)***. Ο ερευνητής οργανώνεται από έναν ***web server*** και χρησιμοποιεί το λεξικό που δημιουργήθηκε από το *DumpLexicon* μαζί με τον *αντίστροφο δείκτη* και το ***PageRanks*** για να απαντήσει στις ερωτήσεις.

4.2.3. Crawling

Προκειμένου να αντιμετωπίσει τις εκατοντάδες των εκατομμυρίων των ιστοσελίδων, η Google έχει ένα γρήγορο κατανεμημένο ***σύστημα crawling***. Ένας ***URLserver*** παρέχει τους καταλόγους των *URLs* σε διάφορους *crawlers*. Κάθε *crawler* χειρίζεται ταυτοχρόνως περίπου 300 συνδέσεις. Αυτό είναι απαραίτητο για την ανάκτηση των ιστοσελίδων με έναν αρκετά γρήγορο ρυθμό. Χρησιμοποιώντας τη μέγιστη ταχύτητα, το σύστημα μπορεί να ανιχνεύσει πάνω από 100 ιστοσελίδες ανά δευτερόλεπτο χρησιμοποιώντας τέσσερις *crawlers*. Αυτό σημαίνει κατά προσέγγιση 600Kb ανά δευτερόλεπτο δεδομένα. Μια σημαντική παράμετρος της απόδοσης είναι το ***DNS lookup***. Κάθε *crawler* διατηρεί το δικό του ***DNS cache*** έτσι δεν χρειάζεται να κάνει *DNS lookup* πριν από την ανίχνευση κάθε εγγράφου. Κάθε μια από τις εκατοντάδες των συνδέσεων μπορεί να είναι σε διαφορετικά κράτη: κοιτάζοντας το *dns*, σύνδεση με τον *οικοδεσπότη* ***(host)***, αποστολή του αιτήματος, και λήψη της απάντησης. Αυτοί οι παράγοντες κάνουν τον *crawler* ένα σύνθετο συστατικό του συστήματος. Χρησιμοποιεί ασύγχρονες IO για να διαχειριστεί τα γεγονότα, και διάφορες σειρές αναμονής για τη μετακίνηση της σελίδας που προσκομίζεται από κράτος σε κράτος.

4.2.4. Ευρετήριο (Indexing)

Οποιοσδήποτε ***parser*** που σχεδιάζεται για να τρέξει σε ολόκληρο τον ιστό πρέπει να μπορεί να χειριστεί έναν τεράστιο αριθμό πιθανών λαθών. Αυτά τα λάθη κυμαίνονται από τύπους στα ***HTML tags*** ως τα ***kilobytes*** των μηδενικών στη μέση ενός tag, ***μη-ASCII*** χαρακτήρες, ***HTML tags*** τοποθετημένα και καλά κρυμμένα, καθώς και μια μεγάλη ποικιλία από άλλα λάθη. Για τη μέγιστη ταχύτητα, αντί της χρήσης του ***YACC*** για τη δημιουργία ***CFG parser***, χρησιμοποιείται ο ***Flex*** για να δημιουργηθεί ένας λεκτικός αναλυτής.

Αφού αναλυθεί κάθε έγγραφο, κωδικοποιείται σε διάφορα βαρέλια. Κάθε λέξη μετατρέπεται σε ***wordID*** με τη χρήση ενός ***hash-table*** - το *λεξικό* ***(lexicon)***. Νέες προσθήκες στο λεξικό του *hash-table* καταγράφονται σε ένα αρχείο. Μόλις μετατραπούν οι λέξεις σε *wordID*, οι εμφανίσεις τους στο τρέχον έγγραφο μεταφράζονται σε ***hit lists*** *(λίστες χτυπημάτων)* και γράφονται στα μπροστινά βαρέλια. Η κύρια δυσκολία με τον παραλληλισμό της φάσης της ευρετηρίασης είναι το γεγονός ότι το λεξικό πρέπει να μοιραστεί. Αντί της διανομής του λεξικού, η Google υιοθέτησε

τη μέθοδο της καταχώρησης σε έναν τομέα όλων των πρόσθετων λέξεων που δε βρίσκονται σε ένα λεξικό βάσης, το οποίο καθορίζει 14 εκατομμύρια λέξεις. Με αυτό τον τρόπο πολλοί ***indexers*** μπορούν να τρέχουν παράλληλα, και έπειτα ένας τελικός *indexer* μπορεί να επεξεργαστεί το μικρό αρχείο των πρόσθετων λέξεων.

Προκειμένου να παραχθεί ο *inverted index*, ο *sorter* παίρνει κάθε ένα από τα μπροστινά *βαρέλια* και τα ταξινομεί με βάση τα *wordID* για να δημιουργηθεί ένα *inverted barrel* για τον τίτλο και τα ***anchor hits*** και ενός πλήρους κειμένου *inverted barrel*. Αυτή η διαδικασία πραγματοποιείται με ένα βαρέλι τη φορά, απαιτώντας κατά συνέπεια προσωρινά λίγο αποθηκευτικό χώρο. Επίσης, η φάση της ταξινόμησης πραγματοποιείται με τη χρήση πολλών μηχανών ταξινόμησης με το τρέξιμο πολλών *sorters*, οι οποίοι μπορούν να επεξεργαστούν τους διαφορετικούς κάδους συγχρόνως. Από αυτή τη στιγμή τα βαρέλια δεν χωρούν στην κύρια μνήμη, ο *sorter* τα υποδιαιρεί περαιτέρω σε καλάθια που χωρούν στη μνήμη βασιζόμενα στα *wordID* και *docID*. Έπειτα ο *sorter* φορτώνει κάθε καλάθι στη μνήμη, ταξινομεί και γράφει το περιεχόμενό τους στο μικρό *inverted barrel* και στο *full inverted barrel*.

4.2.5. Ταξινόμηση (Ranking)

Η Google διατηρεί πολλές πληροφορίες για τα έγγραφα ιστού. Κάθε ***hit list*** περιλαμβάνει τη θέση, το μέγεθος των χαρακτήρων και πληροφορίες εάν είναι κεφαλαία ή όχι. Επιπλέον, τα *hits* στοιχειοθετούνται από το κείμενο ***anchor*** και το ***PageRank*** του εγγράφου. Ο συνδυασμός όλων αυτών των πληροφοριών ως ένα βαθμό είναι δύσκολη. Οι λειτουργίες της ταξινόμησης σχεδιάζονται έτσι ώστε κανένας συγκεκριμένος παράγοντας να μην τις επηρεάζει σημαντικά.

Ερώτημα μιας λέξης:

Προκειμένου να ταξινομηθεί ένα έγγραφο με μια ερώτηση αποτελούμενη από μια λέξη, η Google εξετάζει τη *hit list* του εγγράφου για αυτή τη λέξη. Η Google θεωρεί ότι κάθε ένα hit μπορεί να είναι ένας από τους διάφορους διαφορετικούς τύπους (τίτλος, *anchor, URL, ...*), κάθε ένας από τους οποίους έχει το δικό του βάρος. Τα βάρη των τύπων δημιουργούν ένα διάνυσμα. Η Google μετρά τον αριθμό των χτυπημάτων κάθε τύπου στη *hit list*. Κατόπιν κάθε αρίθμηση μετατρέπεται σε μια αρίθμηση βαρών. Τα αριθμημένα βάρη αυξάνουν γραμμικά στην αρχή και μετά μικραίνουν, έτσι μόνο μια μέτρηση δεν αρκεί. Ο υπολογισμός των βαρών με τον πίνακα των βαρών των τύπων

υπολογίζουν τον τελικό βαθμό της ανάκτησης του εγγράφου. Τελικά αυτός ο βαθμός μαζί με το *PageRank* θα δώσει την τελική κατάταξη του εγγράφου.

Ερώτημα πολλών λέξεων:

Στη συγκεκριμένη περίπτωση οι πολλαπλές *hit lists* πρέπει να ανιχνευθούν αμέσως έτσι ώστε, τα *hits* που εμφανίζονται κοντά το ένα με το άλλο σε ένα έγγραφο τους αποδίδεται μεγαλύτερο βάρος από τα *hits* που εμφανίζονται μακριά το ένα από το άλλο. Τα *hits* από τις πολλαπλές *hit lists* αντιστοιχίζονται έτσι ώστε τα κοντινά *hits* να αντιστοιχίζονται μαζί. Για το κάθε αντιστοιχισμένο σύνολο *hits*, υπολογίζεται η *εγγύτητά του* ***(proximity)***. Η εγγύτητα βασίζεται στο πόσο μακριά εμφανίζονται τα *hits* στο έγγραφο (ή anchor) αλλά ταξινομείται με 10 διαφορετικές τιμές "bins" κυμαινόμενη από την αντιστοιχία φράσης έως "όχι στενή αντιστοιχία". Οι μετρήσεις υπολογίζονται όχι μόνο για τον κάθε *hit* τύπο, αλλά για κάθε τύπο και εγγύτητα. Κάθε ζευγάρι τύπων και εγγύτητας έχει ένα type prox - βάρος. Οι μετρήσεις μετατρέπονται σε count-weights και το συνολικό αποτέλεσμα των count-weights και των type-prox-weight λαμβάνονται για τον υπολογισμό ενός αποτελέσματος IR.

4.2.6. Γενικά Χαρακτηριστικά του Google

Εκτός από την ποιότητα της αναζήτησης, η Google σχεδιάστηκε ώστε να συμπεριφέρεται ικανοποιητικά καθώς ο ιστός μεγαλώνει. Μια πτυχή αυτού είναι να χρησιμοποιηθεί ο αποθηκευτικός χώρος αποτελεσματικά. Λόγω της συμπίεσης, το συνολικό μέγεθος της *αποθήκης* ***(repository)*** είναι περίπου 53 GB, ακριβώς μεγαλύτερος κατά ένα τρίτο των συνολικών δεδομένων που αποθηκεύει. Οι τρέχουσες τιμές των δίσκων κάνουν την αποθήκευση μια σχετικά φτηνή πηγή των χρήσιμων στοιχείων. Το πιο σημαντικό, το σύνολο όλων των στοιχείων που χρησιμοποιούνται από τη μηχανή αναζήτησης απαιτεί ένα συγκρίσιμο όγκο αποθήκευσης, περίπου 55 GB. Επιπλέον, οι περισσότερες ερωτήσεις μπορούν να απαντηθούν με τη χρήση του ***short inverted index***. Με καλύτερη κωδικοποίηση και συμπίεση του ***Document Index***, μια υψηλής ποιότητας μηχανή αναζήτησης ιστού μπορεί να εγκατασταθεί επάνω σε έναν 7GB drive ενός καινούριου PC.

Ο ***Google Indexer*** βλέπει κατά προσέγγιση 54 σελίδες ανά δευτερόλεπτο. Οι ***sorters*** μπορούν να τρέχουν παράλληλα χρησιμοποιώντας τέσσερις μηχανές, ολόκληρη η διαδικασία του ***sorting*** διαρκεί περίπου 24 ώρες. Η τωρινή έκδοση του Google

απαντά στις περισσότερες ερωτήσεις μεταξύ του ενός και των δέκα δευτερολέπτων. Αυτός ο χρόνος εξαρτάται συνήθως από τα IO των δίσκων άνω των NFS (δεδομένου ότι οι δίσκοι βρίσκονται σε διάφορες μηχανές). Επιπλέον, το Google δεν έχει κάποιες βελτιστοποιήσεις όπως εναποθήκευση της ερώτησης, υπο-ευρετήρια με κοινούς όρους, και άλλες κοινές βελτιστοποιήσεις. Οι σχεδιαστές του Google σκοπεύουν να το επιταχύνουν αρκετά μέσω της διανομής και του υλικού, του λογισμικού, και των αλγοριθμικών βελτιώσεων. Στοχεύουν στο να είναι να είσαι σε θέση να χειριστεί αρκετές εκατοντάδες ερωτήσεις ανά δευτερόλεπτο.

Το Google σχεδιάστηκε με σκοπό να ανταποκριθεί και να φτάσει τις 100 εκατομμύρια ιστοσελίδες. Όλα τα μέρη του συστήματος είναι δουλεύουν παράλληλα και κατά προσέγγιση σε γραμμικό χρόνο. Αυτά τα μέρη περιλαμβάνουν τα τους crawlers, τους indexers, και τους sorters. Εντούτοις, στις 100 εκατομμύρια ιστοσελίδες το Google δεν θα μπορεί να τρέχει εύκολα σε όλους τους κοινούς τύπους λειτουργικών συστημάτων (στη σημερινή του μορφή τρέχει σε Solaris και σε Linux). Αυτό περιλαμβάνει θέματα όπως προσπελάσιμη μνήμη, αριθμός από τους περιγραφείς αρχείων, υποδοχών δικτύων και εύρους ζώνης αυτών, και πολλά άλλά. Η επέκτασή του σε πολύ περισσότερες από 100 εκατομμύρια σελίδες θα αύξανε πολύ την πολυπλοκότητα του συστήματος.

4.3. HotBot (www.hotbot.com)

4.3.1. Εισαγωγή

Ένας από τους νεότερους «παίκτες» στην αγορά των μηχανών αναζήτησης η HotBot, έκανε την εμφάνιση της στις αρχές του 1996 ως συστατικό του site του περιοδικού ***«wired»***. Με τον καιρό το *site* διευρύνθηκε και αντίστοιχα η μηχανή αναζήτησης βελτιώθηκε και εμπλουτίστηκε με νέες προσθήκες και διευκολύνσεις, ενώ σιγά- σιγά εξελίχθηκε σε αυτόνομο *site*. Πρόσφατες έρευνες αποδεικνύουν πως όχι μόνο διατήρησε την αυτονομία της, αλλά ταυτόχρονα πέτυχε να καθιερωθεί στη δύσκολη αγορά των εργαλείων αναζήτησης ως μία από τις πιο μοντέρνες και εύχρηστες μηχανές γενικής αναζήτησης.

Συγκεντρώνοντας περισσότερες από 100 εκατομμύρια σελίδες, οι οποίες ανανεώνονται περίπου κάθε δύο εβδομάδες, η HotBot καταφέρνει να συντηρεί μία από τις πλέον ενημερωμένες βάσεις δεδομένων - χαρακτηριστικό που αποκτά ιδιαίτερη σημασία, εάν αναλογιστούμε τους ρυθμούς ανανέωσης των περιεχομένων του *world wide web*. Η HotBot καλύπτει το ***web***, το ***Usenet*** και επίσης τις ***listservs***. Με τα σημερινά όμως δεδομένα αυτή η μεγάλη βάση δεδομένων, τώρα είναι σημαντικά μικρότερη από αρκετές άλλες μηχανές αναζήτησης – και αυτή είναι η πιο σημαντική αδυναμία της.

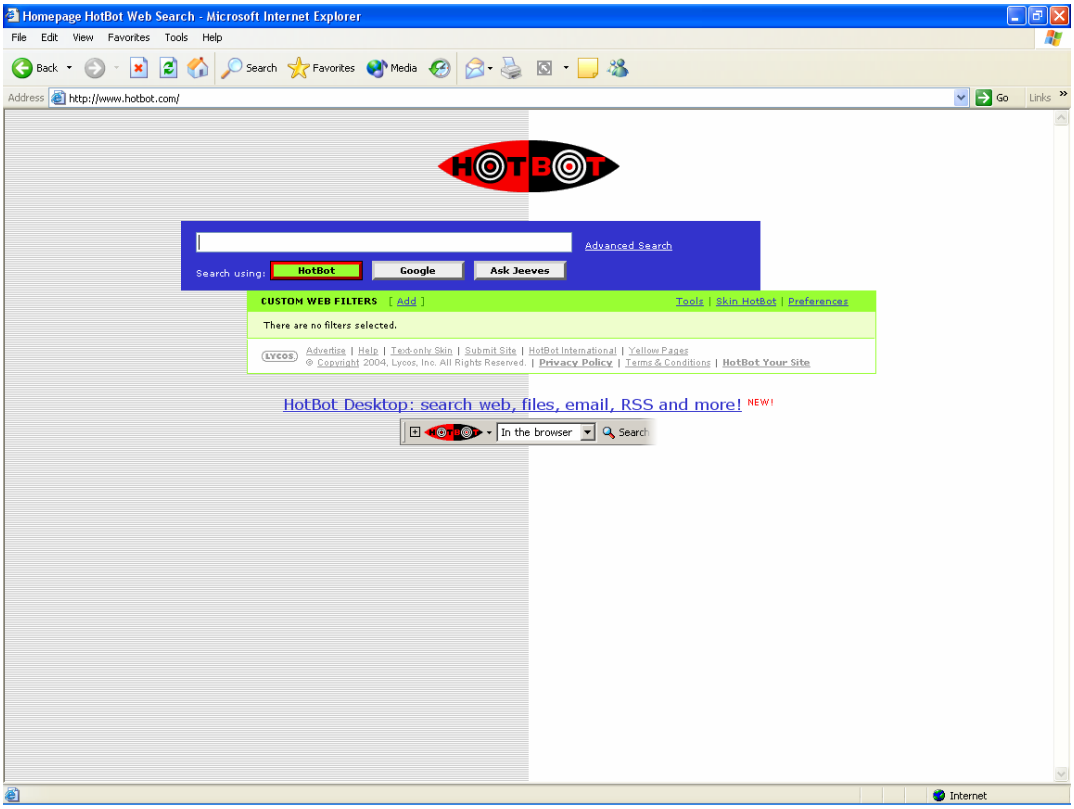
Η HotBot ανήκει στη Lycos, και μπορεί να είναι συχνή, και πιθανώς να αυξηθεί στο μέλλον, η επικάλυψη και οι ομοιότητες μεταξύ των χαρακτηριστικών γνωρισμάτων και των υπηρεσιών αυτών των μηχανών αναζήτησης.

4.3.2. Βασικά Χαρακτηριστικά

Προχωρώντας στην παρουσίαση των γενικών χαρακτηριστικών της μηχανής, παρατηρούμε ότι, όπως και στις περισσότερες μηχανές αναζήτησης, στη HotBot υπάρχουν δύο διαφορετικά ***interfaces*** αναζήτησης, ένα για την απλή και ένα για την προχωρημένη αναζήτηση. Σε ό,τι αφορά το *interface* της απλής αναζήτησης, η HotBot παραδίδει μαθήματα ευχρηστίας και πληρότητας, αφού λίγες είναι οι μηχανές που ακόμη και στην οθόνη της προχωρημένης αναζήτησης διαθέτουν όλες τις ευκολίες και τα ειδικά πεδία που προσφέρει η HotBot στην οθόνη της απλής αναζήτησης. Εκτός του *βασικού πεδίου αναζήτησης* ***(query box)*** στην πρώτη σελίδα, ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει έναν συνδυασμό από ***pull down menus*** και ***check boxes*** προκειμένου να περιορίσει όσο το δυνατό περισσότερο το εύρος της αναζήτησής του

(Σχήμα 4.8.).

Όπως στη σελίδα της απλής έτσι και σε εκείνη της προχωρημένης αναζήτησης

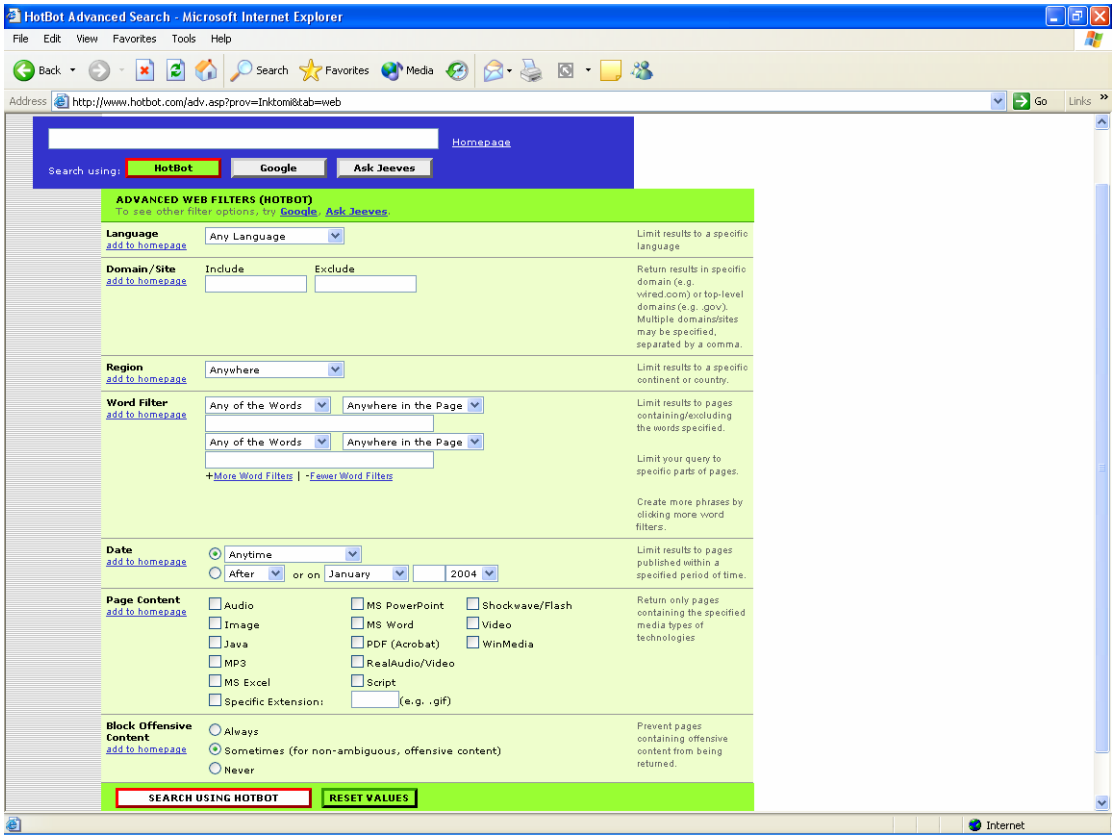


Σχήμα 4.8.: Η αρχική σελίδα (Homepage) της HotBot (www.hotbot.com)

(Σχήμα 4.9.) κυριαρχούν τα *pull down menus*. Αυτά, εκτός του ότι είναι πολύ περισσότερα, μπορούν να συνδυαστούν με τα *check boxes* και τους λογικούς τελεστές ***Boolean***. Η HotBot, όπως και άλλες μηχανές αναζήτησης, προκειμένου να διευκολύνει τους λιγότερο εξοικειωμένους χρήστες, δε χρησιμοποιεί μόνο την καθιερωμένη ορολογία των τελεστών αλλά και μια σειρά από λέξεις και φράσεις που υποκαθιστούν τους τελεστές όπως αναφέρεται στην παράγραφο των χαρακτηριστικών αναζήτησης.

Σε ό,τι αφορά τους περιορισμούς που έχει στη διάθεση του ο χρήστης προκειμένου να μειώσει το εύρος της αναζήτησης, η μηχανή προσφέρει μια πλειάδα επιλογών και ρυθμίσεων εύχρηστων ακόμη και από τον αρχάριο χρήστη της μηχανής. Η HotBot μαζί με την AltaVista είναι από τις μηχανές που διαθέτουν τους περισσότερους *τελεστές περιορισμού* ***(meta-tags ή limit operators)***. Ειδικότερα, η HotBot προσφέρει τη δυνατότητα περιορισμού της αναζήτησης ανάλογα με την ημερομηνία, τη γλώσσα του εγγράφου καθώς και το μέσο, δηλαδή το εάν πρόκειται για αρχείο εικόνας, ήχου, ακόμη και ***javascript***. Στην οθόνη της προχωρημένης αναζήτησης η HotBot προσφέρει μια διευρυμένη εκδοχή των διευκολύνσεων που συναντήσαμε

στην απλή αναζήτηση, με την προσθήκη του περιορισμού ανάλογα με το ***domain name***, τη γεωγραφική περιοχή, όπως επίσης το *«βάθος σελίδας»* ***(page depth)***, δηλαδή το επίπεδο στο οποίο θα κληθεί να φτάσει η μηχανή προκειμένου να ικανοποιήσει το ερώτημα του χρήστη. Τέλος, η επιλογή ***Word Stemming*** αναλαμβάνει να εντοπίσει παραλλαγές και παράγωγα των όρων που είχε εισαγάγει αρχικά ο χρήστης στο πεδίο αναζήτησης.



Σχήμα 4.9.: Η σελίδα προχωρημένης αναζήτησης (Advanced Search) της HotBot

4.3.3. Ευρετήριο (Indexing)

Στη βάση δεδομένων που διαθέτει, τα *crawlers* της HotBot αναλαμβάνουν να ευρετηριάσουν σχεδόν το σύνολο των λέξεων που εμπεριέχονται σε κάθε *web page* που συναντούν, με μοναδική εξαίρεση τις λέξεις που περιέχονται σε αρχεία εικόνων.

4.3.4. Ταξινόμηση (Ranking)

Η HotBot ταξινομεί τα ανακτημένα έγγραφα βασισμένη στους ακόλουθους παράγοντες (Hock, 2001):

- Συχνότητα των όρων του ερωτήματος στο έγγραφο
- Εμφάνιση των όρων αναζήτησης στον τίτλο του εγγράφου
- Εμφάνιση των όρων αναζήτησης ως λέξεις κλειδιά - οι όροι ως λέξεις κλειδιά στα metatag συμβάλλουν περισσότερο στο αποτέλεσμα της ταξινόμησης από ότι οι λέξεις κειμένου, αλλά λιγότερο από τις λέξεις που βρίσκονται στον τίτλο
- Το μέγεθος του εγγράφου- ένα μικρό έγγραφο με Ν εμφανίσεις της λέξης ταξινομείται υψηλότερα από ένα άλλο έγγραφο με τον ίδιο αριθμό εμφάνισης των όρων
- Το anti-spamming -διάφορα τεχνάσματα χρησιμοποιούνται από τους δημιουργούς μιας ιστοσελίδας για αυξήσουν σε ένα έγγραφο τη σχετικότητά του και να λάβει τεχνητά καλύτερη θέση ταξινόμησης όταν εξετάζεται από τις μηχανές αναζήτησης. Ένα παράδειγμα είναι η χρήση μιας λέξης εκατοντάδες φορές σε ένα έγγραφο. Εάν η HotBot εντοπίσει το *spam*, μικραίνει τη θέση ταξινόμησής του.
- Η HotBot έχει έναν κατάλογο μερικών κοινών λέξεων που δεν ευρετηριάζονται ***(stop words)***. Αυτός ο κατάλογος έχει γίνει πολύ μικρός κατά τη διάρκεια του χρόνου.

Direct Hit – Η HotBot χρησιμοποιεί τη μηχανή ***Direct Hit*** για να προσδιορίσει "τις περισσότερο δημοφιλείς ιστοσελίδες". Για κοινές αναζητήσεις, μέχρι 10 "τις πιο συχνές ιστοσελίδες" θα εμφανιστούν στην κορυφή της λίστας των αποτελεσμάτων. **Αυτό μπορεί να είναι πολύ χρήσιμο, ειδικά για απλές, κοινές αναζητήσεις -** για την εύρεση των αρχικών σελίδων των επιχειρήσεων.

4.3.5. Χαρακτηριστικά Αναζήτησης (Search Features)

Λογική Αναζήτησης και Σύνταξης

Η HotBot υποστηρίζει τη χρήση όλων των τελεστών Boolean AND, NOT και OR, ενώ κάνει χρήση των συμβόλων + και - στην απλή αναζήτηση και ειδική ορολογία στην προχωρημένη αναζήτηση. Έτσι, στη θέση του λογικού τελεστή AND ο χρήστης της HotBot μπορεί να χρησιμοποιήσει τη φράση «All the words», στη θέση του τελεστή OR τη φράση «Any of the words», ενώ προκειμένου να προσδιορίσει την *εγγύτητα* ***(proximity)*** μεταξύ των αναζητήσιμων όρων θα πρέπει να χρησιμοποιήσει τη φράση «The Person». Φυσικά, για τους χρήστες που είναι περισσότερο εξοικειωμένοι με τους λογικούς τελεστές Boolean (AND, OR, NOT, AND NOT) και τη χρήση τους, η HotBot

επιτρέπει την εισαγωγή των τελεστών αυτών στο πεδίο αναζήτησης με την καθιερωμένη μορφή τους.

Αναζήτηση Φράσεων (Phrase Searching)

Η μηχανή υποστηρίζει την αναζήτηση φράσεων. Οι φράσεις μπορούν να αναζητηθούν είτε με την είσοδο της φράσης και επιλέγοντας *"the phrase"* ως τύπο αναζήτησης στο *"Look for"* παράθυρο, είτε με τη χρησιμοποίηση των ανωφερή εισαγωγικών γύρω από τη φράση. Εκτός από τις ακριβείς φράσεις, η αναζήτηση με βάση την εγγύτητα δεν είναι διαθέσιμη.

Παράδειγμα: ""πίνακες κυκλωμάτων"

Αποκοπή (Truncation)

Ένας αστερίσκος (*) χρησιμοποιείται για την αποκοπή των λέξεων προκειμένου να αναζητηθούν τα παράγωγα των λέξεων προς αναζήτηση. Η HotBot παρέχει και την *ελεγχόμενη από τον χρήστη αποκοπή* ***(user- controlled truncation)*** και την αυτόματη ρίζα της λέξης ***(automatic word-stemming)***. Και στις δύο περιπτώσεις η HotBot χρησιμοποιεί τον **αστερίσκο.** Έτσι, χρησιμοποιεί το χαρακτήρα (*) προκειμένου να αναζητήσει απεριόριστο αριθμό χαρακτήρων μετά τη ρίζα της λέξης (π.χ. book*) και το χαρακτήρα (?) προκειμένου να αναζητήσει μόνο έναν χαρακτήρα μετά τη ρίζα της λέξης (π.χ. book?)

Στην προηγμένη έκδοση, υπάρχει η επιλογή της αυτόματης ρίζας της λέξης που επιτυγχάνεται κλικάροντας ***"Enable Word Stemming"***. Έπειτα, η HotBot θα εμφανίσει κάποιους γραμματικούς κανόνες.

Η χρήση της ρίζας και η αποκοπή δεν μπορούν να χρησιμοποιηθούν από κοινού. Στις περισσότερες περιπτώσεις, θα είναι απλούστερο και πιθανώς αποτελεσματικότερο να επιλεχθεί η αποκοπή.

Αναγνώριση Πεζών/Κεφαλαίων

Αν και η HotBot υποστηρίζει ότι είναι δυνατή η διάκριση των κεφαλαίων και των πεζών χαρακτήρων, αλλά η χρήση της αποδεικνύει το αντίθετο, έτσι δεν μπορεί κάποιος να βασιστεί σε αυτή την υπηρεσία. Έτσι, αν εισαγάγουμε τον αναζητήσιμο όρο με πεζούς χαρακτήρες, η μηχανή θα τον ανακτήσει στην περίπτωση που είναι γραμμένος τόσο με πεζούς όσο και με κεφαλαίους χαρακτήρες. Αν εισαγάγουμε τον όρο με κεφαλαίους χαρακτήρες, η μηχανή θα αναζητήσει και θα ανακτήσει τον όρο

μόνο όπου τον συναντήσει με κεφαλαίους χαρακτήρες.

Αναζήτηση με Βάση το Όνομα

Για την αναζήτηση του ονόματος ενός προσώπου επιλέγεται στο ***"Look for"*** η επιλογή ***"the person"***. Όταν διευκρινίζεται "το πρόσωπο", η HotBot ψάχνει αυτόματα για τη δοθείσα μορφή καθώς επίσης και για την κανονική μορφή του ονόματος. *Παράδειγμα:* Το Winston Churchill θα ανακτήσει και τους δύο:

Winston Churchill

Churchill, Winston

4.3.6. Παρουσίαση Αποτελεσμάτων

Η HotBot εμφανίζει μόνο ένα αποτέλεσμα ανά *site*. Για να εμφανιστούν περισσότερες σελίδες από το *site*, πρέπει να γίνει κλικ στο ***"See results from this site only"*** στο τέλος του αρχείου. Η HotBot εμφανίζει μέχρι 1.000 αποτελέσματα. Στη δεύτερη σελίδα των αποτελεσμάτων υπάρχει μια ένδειξη για τα συνολικά ανακτημένα αρχεία.

Όπως αναφέρθηκε νωρίτερα, η HotBot χρησιμοποιεί το ***"Direct Hit"***. Για αναζητήσεις που είναι αρκετά κοινές, στις σελίδες των αποτελεσμάτων θα εμφανιστεί το μήνυμα ***"Top 10 Matches"*** στην κορυφή του καταλόγου. Αυτά τα πρώτα 10 αρχεία προέρχονται από το ***"Direct Hit"***. Εάν το ***"Direct Hit"*** δεν έχει ανακτήσει 10 αποτελέσματα (ή εάν έχει τροποποιηθεί η ερώτηση με κάποιο τρόπο, όπως με την ημερομηνία), θα εμφανιστεί μια σύνδεση ***"Get the Top N sites"***.

Για πολλές αναζητήσεις, οι σελίδες των αποτελεσμάτων της HotBot θα περιλαμβάνει τα ακόλουθα:

- ✓ **Προτεινόμενες Εναλλακτικές Αναζητήσεις (Suggested alternate searches)**.
- ✓ **Σχετικές Κατηγορίες (Related Categories)** - από το ***Open Directory***.
- ✓ **Συνδέσεις με τους Πόρους του Lycos Network (Links to Lycos Network resources)** - για αναζητήσεις που κάνουν οι επιχειρήσεις.
- ✓ **Τα Καλύτερα 10 Αποτελέσματα (Top 10 matches) -** Τα αποτελέσματα του ***Direct Hit***.
- ✓ **Αποτελέσματα από Καταλόγους (Matching Directory Records)** - από το ***Open Directory***.
- ✓ **Η βάση Δεδομένων των Αποτελεσμάτων (Web database results)** - από τη

βάση δεδομένων της Inktomi.

Στις σελίδες αποτελεσμάτων, δίπλα στο ***"query box"***, η HotBot δίνει τη δυνατότητα αναζήτησης μέσα στα αποτελέσματα ***(within these results)***. Και σε μερικές άλλες μηχανές αναζήτησης που παρέχουν μια παρόμοια υπηρεσία, αυτό μπορεί να θεωρηθεί όχι και κάτι το σημαντικό. Σε οποιαδήποτε μηχανή αναζήτησης μπορεί να γίνει αναζήτηση μέσα στα αποτελέσματα με την προσθήκη πρόσθετων όρων στην ερώτηση.

4.3.7. HotBot Directory

Η HotBot κάνει χρήση του ***Open Directory*** για τον κατάλογο του Ιστού της. Όταν επιλεχθεί μια από τις κατηγορίες καταλόγου που βρίσκονται στην αρχική σελίδα της HotBot γίνεται πρόσβαση στο ***Open Directory***. Η εφαρμογή της HotBot είναι αρκετά τυποποιημένη, και σε κάθε επίπεδο μπορεί να ψαχθεί είτε ολόκληρος ο κατάλογος είτε το τρέχον επίπεδο. Η HotBot κάνει αποτελεσματικότερη τη χρήση του ***Open Directory***, όπως αναφέρθηκε παραπάνω, ενσωματώνοντας αυτόματα οποιεσδήποτε σελίδες από τον κατάλογο στα αποτελέσματα αναζήτησης (μετά από τα καλύτερα Direct Hit αποτελέσματα και πριν από τα αποτελέσματα της μεγάλης βάσης δεδομένων της Inktomi). Εκτός από τις περιοχές του καταλόγου που απαριθμούνται μέσα στα αποτελέσματα, οι "σχετικές κατηγορίες" από τον κατάλογο θα εμφανιστούν στην πρώτη σελίδα των αποτελεσμάτων.

4.3.8. Ειδικές Επιλογές/ Χαρακτηριστικά

Η HotBot παρέχει διάφορες χρήσιμες πρόσθετες επιλογές και στην αρχική σελίδα της και στη σελίδα προχωρημένης αναζήτησης. Από την πλευρά των χαρακτηριστικών πύλης ***(portal features)***, η HotBot δεν πρέπει πιθανώς να θεωρηθεί ως μια γενική πύλη δεδομένου ότι δεν παρέχει οποιεσδήποτε επιλογές ή *χαρακτηριστικά εξατομίκευσης* ***(personalization options)***, όπως οι σημαντικότεροι τίτλοι ειδήσεων, ο καιρός, και άλλες υπηρεσίες πυλών που βρίσκονται σε μερικά άλλα *sites*. Εντούτοις, παρέχει ένα πλήθος *χρήσιμων συνδέσεων* ***(useful links)***. Αναφέρουμε μερικές ενδεικτικά: News Headlines, Yellow Pages, White Pages, Email Addresses, Email and Homepages, Calendar, Greeting, Cards, Lycos SHOP, FTP Search, Free Downloads, Road Maps, Books, Hardware, Domain Names, Classifields, Music Search, Jobs & Resumes, Travel, Autos,

κ.τ.λ.

Με τη χρήση της επιλογής του βάθους σελίδων στην προχωρημένη αναζήτηση της HotBot, μπορεί να διευκρινιστεί εάν θα ανακτηθεί μια ***top - level*** σελίδα ενός *site*, ή μια σελίδα κατώτερης ιεραρχίας. Μια εφαρμογή αυτού είναι η αναζήτηση της αρχικής σελίδας μιας επιχείρησης. Εάν, παραδείγματος χάριν, αναζητείται η επιχείρηση Biogen, τοποθετείται "Biogen" ως λέξη τίτλου και γίνεται κλικ στο *Top Page*. Μπορεί να ανακτήσει άμεσα αυτό που ψάχνουμε χωρίς να πρέπει να εξεταστούν πάνω από 100 άλλες σελίδες. Αυτό, εντούτοις, είναι ένα από τα χαρακτηριστικά που δεν λειτουργεί πάντα.

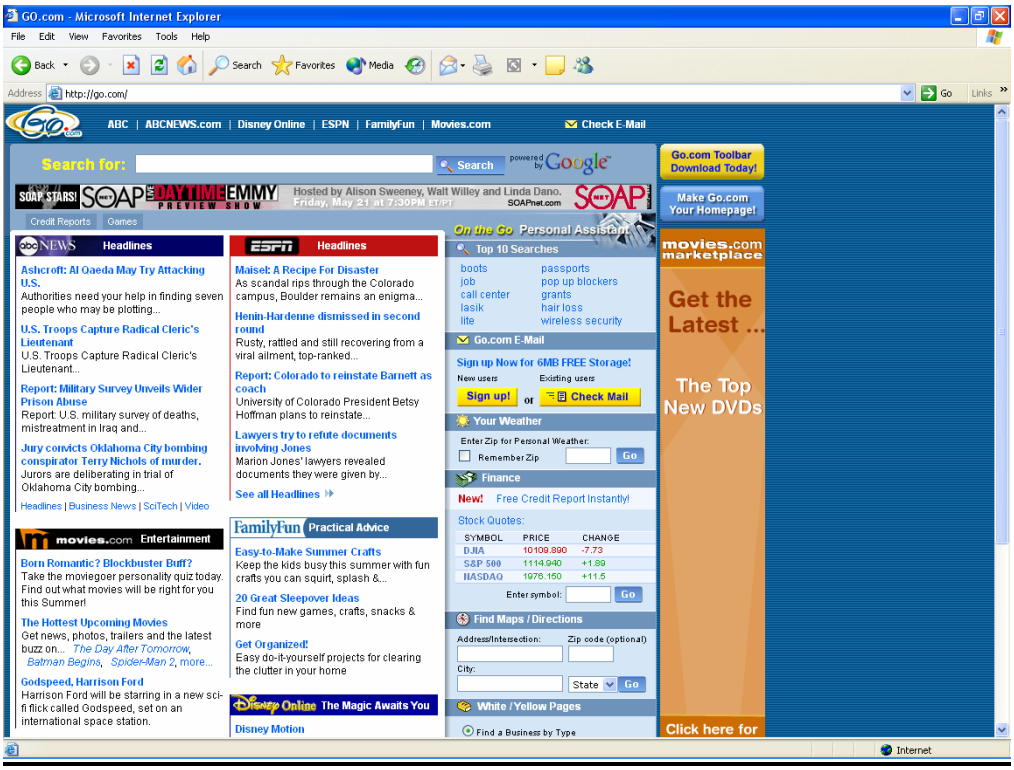
Όσον αφορά τα αρχεία βοήθειας και οδηγιών, η HotBot παρέχει μια σελίδα με *συνηθισμένες ερωτήσεις* ***(frequently asked questions, FAQ)***, η οποία όμως θα μπορούσαμε να αναφέρουμε ότι μόνο οδηγίες προς τους ερευνητές δεν περιέχει, αφού οι περισσότερες από αυτές απευθύνονται σε δημιουργούς και συντηρητές *sites* και όχι στον επίδοξο ερευνητή που θα θελήσει να μάθει τον τρόπο να χρησιμοποιεί καλύτερα τη μηχανή. Στην παραπλανητική επιλογή ***search tips*** ο χρήστης θα βρει δύο βασικές επιλογές, τις ***Getting Started*** και ***Advanced Search Features***. Η τελευταία επιλογή είναι αυτή που ενεργοποιεί το πραγματικό αρχείο βοήθειας για την προχωρημένη αναζήτηση, ενώ περιγράφει με αναλυτικά παραδείγματα και οδηγίες τον τρόπο χρήσης και τα εξελιγμένα χαρακτηριστικά της μηχανής, δίνοντας έτσι στο χρήστη μια πλήρη και κατατοπιστική εικόνα των δυνατοτήτων του συγκεκριμένου εργαλείου.

Για το σοβαρό ερευνητή-αναζητητή, η HotBot είναι εύκολη στη χρήση της, και η ολοκλήρωση των αποτελεσμάτων με τη χρήση του καταλόγου την κάνουν ένα πολύτιμο εργαλείο αναζήτησης. Οι επιλογές της είναι σαφείς και έτσι ο χρήστης ξέρει τα πιθανά αποτελέσματα. Έτσι η HotBot μπορεί να χρησιμοποιηθεί παραγωγικά και από τους συχνούς ερευνητές και από τους περιστασιακούς ερευνητές. Η μεγαλύτερη αδυναμία της είναι το μέγεθος της βάσης δεδομένων της, η οποία μπορεί βελτιωθεί στο μέλλον.

4.4. Infoseek (www.infoseek.com)

4.4.1. Εισαγωγή

Η μηχανή αναζήτησης Infoseek αποτελεί έναν «βετεράνο» ανάμεσα στα εργαλεία αναζήτησης, αφού υφίσταται από τον Ιανουάριο του 1994, εποχή ακόμη πρώιμη για το διαδίκτυο και ιδιαίτερα για τα εργαλεία αναζήτησης τέτοιου τύπου. Η Infoseek καλύπτει περίπου 75.000.000 διευθύνσεις ***(URL)*** (Ιούνιος 1999), το οποίο είναι περίπου το 9,4% του συνόλου του διαδικτύου και, διαθέτοντας ένα ιδιαίτερα απλό και εύχρηστο user interface, καταφέρνει να κερδίσει το χρήστη από την πρώτη σελίδα. Σύμφωνα με εταιρείες έρευνας, η μηχανή αναζήτησης Infoseek συγκαταλέγεται ανάμεσα στις πέντε πιο δημοφιλείς μηχανές αναζήτησης.



Σχήμα 4.10.: Η αρχική σελίδα (Homepage) της Infoseek- Go.com (www.go.com)

Η υπηρεσία αναζήτησης της Infoseek αποτελείται από δύο ενσωματωμένες υπηρεσίες (Bikkannavar, 1999):

1. Την Υπηρεσία Καταλόγου Infoseek (Infoseek directory service): Η Infoseek έχει ταξινομήσει τα εκατομμύρια των σελίδων σε έναν κατάλογο με βάση το θέμα τους. Αυτός ο κατάλογος έχει περίπου 15 κατηγορίες. Αυτό είναι ο κατάλογος των ιστοσελίδων που υποβάλλονται από τους ***webmasters*** και τους χρήστες

2. Την Ultraseek Μηχανή Αναζήτησης και το Ευρετήριο (Ultraseek Search Engine and Index): Η βάση δεδομένων της *Ultraseek* περιέχει πάνω από 50 εκατομμύρια σελίδες.

Το πιο σημαντικό σημείο της Infoseek είναι η συνεργασία μεταξύ του καταλόγου και του *Ultraseek Server*, με σκοπό την επίτευξη μιας ολοκληρωμένης αναζήτησης. Όταν εκτελείται μια τυπική αναζήτηση από την προεπιλεγμένη αρχική σελίδα της Infoseek (Σχήμα 4.10.), η αναζήτηση ενεργοποιεί και τον κατάλογο και την *Ultraseek* μηχανή.

4.4.2. Βασικά Χαρακτηριστικά

Σύμφωνα με στοιχεία της ίδιας της εταιρείας, η μηχανή αναζήτησης ανανεώνει τα περιεχόμενα της κάθε δύο με τρεις εβδομάδες, πρόκειται λοιπόν για μία από τις πλέον σύγχρονες μηχανές αναζήτησης. Σε ό,τι αφορά τα γενικά χαρακτηριστικά της, εύκολα διαπιστώνουμε ότι πρόκειται για ένα από τα πλέον φιλικά και εύχρηστα εργαλεία, αφού το βασικό *interface* είναι ιδιαίτερα απλό, χωρίς ωστόσο να διακρίνουμε κάποια σημαντική έλλειψη σε αυτό. Ο χρήστης έχει τη δυνατότητα να αναζητήσει τις πληροφορίες του με βάση είτε το θέμα είτε κάποια *λέξη-κλειδί* ***(keyword)***.

Η Infoseek προσφέρει δύο *interfaces* αναζήτησης, με το πρώτο να χρησιμοποιείται για την απλή και το δεύτερο για την προχωρημένη αναζήτηση. Στην απλή αναζήτηση ο χρήστης έχει στη διάθεση του το *πεδίο αναζήτησης* ***(query box)*** καθώς και ένα μενού επιλογών ***pull down***, από το οποίο μπορεί να επιλέξει τη θεματική κατηγορία μέσα από την οποία θέλει να αντλήσει τις πληροφορίες που θα εισαγάγει στο πεδίο αναζήτησης. Οι κατηγορίες αυτές μπορεί να είναι *web sites*, *newsgroups* καθώς και ονόματα εταιρειών. Στη σελίδα της προχωρημένης αναζήτησης, ο χρήστης επίσης έχει στη διάθεση του ένα *πεδίο αναζήτησης* καθώς και ένα μενού επιλογών ***pull down***, από όπου μπορεί να επιλέξει την επιλογή πεδίων όπως τον *τίτλο της σελίδας* ***(web page title)***, *τη διεύθυνση* ***(URL)***, τους όρους που θα πρέπει ή δε θα πρέπει να συμπεριλαμβάνονται στις σελίδες των αποτελεσμάτων, κύρια ονόματα ή ακόμη το όνομα κάποιου domain.

Και στις δύο μορφές αναζητήσεων που διαθέτει η Infoseek υποστηρίζεται η αναζήτηση με τη χρήση τελεστών Boolean, αν και η σύνταξη είναι διαφορετική σε κάθε περίπτωση. Ενώ στην απλή αναζήτηση η σύνταξη για τον τελεστή AND είναι το σύμβολο + και για τον τελεστή OR το σύμβολο -, στην προχωρημένη αναζήτηση χρησιμοποιείται ειδική ορολογία προκειμένου να χρησιμοποιήσουμε τους λογικούς

τελεστές Boolean. Για παράδειγμα, στη θέση του τελεστή AND ο χρήστης θα πρέπει να εισαγάγει τον όρο Must, αντί του τελεστή OR θα πρέπει να εισαγάγει τη λέξη Should και, τέλος, αντί του τελεστή NOT τη φράση Should Not. Γενικά, ως *αυτόματο* ***(default)*** τελεστή η μηχανή αναζήτησης της Infoseek χρησιμοποιεί τον τελεστή OR.

Τα αποτελέσματα της αναζήτησης επιστρέφονται από το λογισμικό της μηχανής, το οποίο χρησιμοποιεί στατιστικές μεθόδους μέτρησης και άλλες τεχνικές προκειμένου να προσδιορίσει τη σχετικότητα και τη σημασία των όρων που ανακτήθηκαν σε σχέση με τους όρους που αρχικά ζητήθηκαν.

Σε ό,τι αφορά τη *βοήθεια* ***(Help Files)*** που παρέχεται στο χρήστη, η Infoseek διαθέτει αναλυτικές και ιδιαίτερα λεπτομερείς οδηγίες χρήσης τόσο για την απλή όσο και για την προχωρημένη αναζήτηση, όπως επίσης για τους τρόπους με τους οποίους ο χρήστης μπορεί να *βελτιώσει* ***(refine)*** τα αποτελέσματα της αναζήτησης του.

4.4.3. Crawling

Αφότου μια ιστοσελίδα υποβάλλεται, *το ειδικό λογισμικό της Infoseek* ***(Infoseek spider)*** συνήθως θα επισκεφθεί την ιστοσελίδα μέσα σε μια εβδομάδα για να ελέγξει τις σελίδες και συχνά να προσθέσει και άλλες από το ίδιο ***site***. Μετά από αυτό, το spider υποτίθεται ότι επισκέπτεται το ***site*** τουλάχιστον κάθε δύο μήνες. Μπορεί να το επισκεφθεί πιο σύντομα εάν οι σελίδες αλλάζουν συχνά.

4.4.4. Ταξινόμηση (Ranking)

Οι παρακάτω είναι οι σημαντικότεροι παράγοντες ταξινόμησης που εφαρμόζει η Infoseek (Bikkannavar, 1999):

- **Λέξεις-κλειδιά στην αρχή του εγγράφου (Early Keywords):** Σελίδα με λέξεις κλειδιά στους τίτλους και στην αρχή της.
- **Metatags:** Οι λέξεις-κλειδιά στα ***metatags*** βοηθούν στον εντοπισμό της σχετικότητας του περιεχομένου της σελίδας.
- **Συχνότητα (Frequency):** Οι σελίδες με υψηλή συχνότητα εμφάνισης των λέξεων-κλειδιών ταξινομούνται πιο υψηλά.
- **Λίστα καταλόγου (Directory Listing):** Ένα σημαντικότερο πλεονέκτημα δίνεται στις σελίδες που είναι συνδεδεμένες στον κατάλογό της.

- **Πλήθος συνδέσεων με άλλες σελίδες (Link Popularity):** Η Infoseek χρησιμοποιεί το ***Link Popularity*** για να ταξινομήσει μερικές ιστοσελίδες υψηλότερα.

4.4.5. Ευρετήριο (Indexing)

Η Infoseek συντάσσει ευρετήριο με βάση τους παρακάτω κανόνες (Bikkannavar, 1999):

- Βρίσκει τις αντιστοιχίες για μερικές λέξεις που συλλαβίζονται είτε ως μια λέξη, είτε ως δύο λέξεις. Παραδείγματος χάριν, μια αναζήτηση των λέξεων "hard disk" επίσης θα βρει και τις λέξεις "harddisk", και αντίστροφα.
- Αυτόματα αναζητά τη *ρίζα των λέξεων* ***(stemming)***, έτσι μια αναζήτηση για "swim" θα εμφανίσει επίσης σελίδες με τις λέξεις "swims" και "swimming".
- Οι περιγραφές διαμορφώνονται από τα metatags ή από τους πρώτους χαρακτήρες μιας σελίδας. Γενικά είναι περίπου 200 χαρακτήρες, αν και το μήκος μπορεί να ποικίλει μεταξύ των 170 και 240 χαρακτήρων.

4.4.6. Spamming

Η Infoseek δεν καταχωρεί τις σελίδες που κάνουν χρήση των διάφορων τεχνικών spamming. Οι σημαντικότεροι παράγοντες που λαμβάνονται ως spamming είναι:

- Κατάχρηση ή επανάληψη των λέξεων
- Η χρήση των meta ενημερώνει γρηγορότερα από ότι το ανθρώπινο μάτι μπορεί να δει.
- Χρήση οποιουδήποτε επαναπροσανατολισμού.
- Χρήση του ίδιου χρώματος κειμένου με το ***background***.
- Χρήση λέξεων-κλειδιών που δεν αφορούν το περιεχόμενο της σελίδας.
- Όμοιες σελίδες με διαφορετικά ***URLs***.
- Χρήση σελίδων που καταλήγουν στο ίδιο ***URL***.
- Απορρίπτει σελίδες που χρησιμοποιούν μέγεθος γραμματοσειράς πάρα πολύ μικρό, όπως μέγεθος -1.
- Οι σελίδες που υποβάλλονται πάνω από μία φορά σε μια περίοδο 24 ωρών μπορούν επίσης να θεωρηθούν ως *spam* και αφαιρούνται από το ευρετήριο.

Ο εντοπισμός του *Spamming* γίνεται και στο επίπεδο των spiders και στο επίπεδο

των συντακτών, οι οποίοι διενεργούν ελέγχους σημείων ***(spot checks)*** για τα πιο δημοφιλή θέματα.

4.4.7. Άλλα χαρακτηριστικά της Infoseek

Φιλικότητα προς το χρήστη

Η Infoseek έχει κάνει το καλύτερο δυνατό να καταστήσει όλες τις πτυχές των προϊόντων της φιλικές προς το χρήστη. Από όλες τις σημαντικότερες μηχανές γενικής αναζήτησης, η Infoseek έχει την καλύτερη σχεδίαση interface. Η φιλική προς το χρήστη ατμόσφαιρα της Infoseek επεκτείνεται και στο λογισμικό αναζήτησης του Ultraseek server. Η εγκατάσταση, η συντήρηση και η χρήση είναι όλα ευκολότερα και περισσότερο διαισθητικά απ' οποιαδήποτε από τα άλλα εμπορικά πακέτα των μηχανών αναζήτησης.

Ευελιξία και πλήθος χαρακτηριστικών

Η Infoseek έχει ισορροπήσει την ευελιξία και το πλήθος των χαρακτηριστικών με την εύκολη παραμετροποίηση, με συνέπεια να υπάρξει ένα προϊόν που χρειάζεται ελάχιστο χρονικό διάστημα διαμόρφωσης σύμφωνα με τις προτιμήσεις του χρήστη.

Τα πλεονεκτήματά της

Η μηχανή αναζήτησης Infoseek υπερέχει σε κάποια σημεία έναντι των άλλων μηχανών αναζήτησης αφού:

1. Μπορεί να καθοριστεί ο αριθμός των spiders ***(threads)*** που συντάσσουν ευρετήριο ή αναζητούν τον ιστό οποιαδήποτε στιγμή. Μπορούν να αρχίσουν και να σταματήσουν τα ***threads*** μεμονωμένα.

2. Η Ultraseek επιτρέπει επίσης την αναζήτηση με βάση τον τίτλο, με βάση τη *URL*, με βάση το όνομα των ***hosts***, με βάση την εικόνα, και με βάση τα ***metatags***.

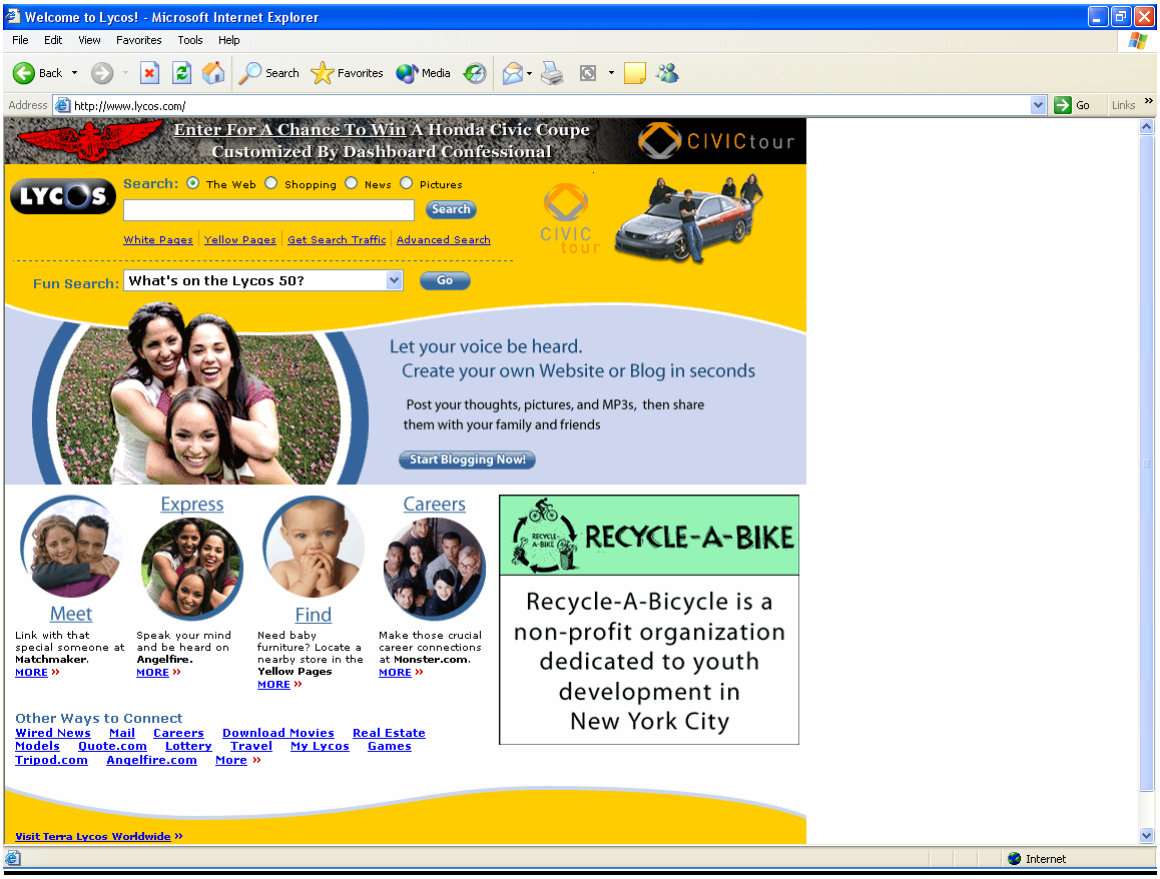
3. Η Ulrtaseek παρέχει υποστήριξη για πολλαπλά ευρετήρια, αποκαλούμενα *συλλογές* ***(collections) Ultraseek speak***. Κάθε συλλογή μπορεί να συντάξει ευρετήριο από ένα διαφορετικό τμήμα ενός ***intranet***. Οι χρήστες μπορούν να ψάξουν κάθε συλλογή χωριστά ή να επιτύχουν οποιαδήποτε ομαδοποίηση των συλλογών συγχρόνως.

4.5. Lycos (www.lycos.com)

4.5.1. Εισαγωγή

Ο Dr Michael Mauldin ήταν υπεύθυνος του προγράμματος Lycos για τη *Μηχανή Μετάφρασης* ***(Machine Translation)*** στο πανεπιστήμιο του ***Carnegie Mellon*** ως πείραμα *της πρώτης καλύτερης αναζήτησης* ***(best first search)*** πληροφοριών του ιστού. Βοηθά τους χρήστες να εντοπίσουν έγγραφα που περιέχουν τις συγκεκριμένες λέξεις-κλειδιά που παρέχονται από τον χρήστη. Λόγω της περιεκτικότητας της βάσης δεδομένων της, η Lycos έγινε γρήγορα πολύ δημοφιλής στους χρήστες ιστού που έπρεπε να διεξαγάγουν αναζητήσεις με βάση το περιεχόμενο των εγγράφων στο χώρο που διαμορφώθηκε από τον ιστό. Η Lycos θεωρείται ότι έχει συντάξει ευρετήριο με 50 εκατομμύρια ιστοσελίδας, το οποίο είναι περίπου 6,3% του συνολικού όγκου του ιστού.

Το σχήμα 4.11. παρουσιάζει την αρχική σελίδα της μηχανής αναζήτησης Lycos, όπου εκτός από το συνηθισμένο απλό *πεδίο αναζήτησης* ***(search box)***, υπάρχει και η δυνατότητα επιλογής κάποιων ενεργειών πέρα από την αναζήτηση.



Σχήμα 4.11.: Η αρχική σελίδα (Homepage) της Lycos (www.lycos.com)

Η Lycos απαντά στην ερώτηση του χρήστη ερευνώντας τον τεράστιο *κατάλογό* της από διευθύνσεις του ιστού. Ομάδες λογισμικού αποκαλούμενες spiders ψάχνουν το WWW καθημερινά (συμπεριλαμβανομένων των περιοχών gopher και FTP), και δημιουργούν έναν κατάλογο. Ο κατάλογος είναι κάτι λιγότερο από μια βάση δεδομένων με διευθύνσεις ιστού που περιέχει πληροφορίες για το τι είναι σημαντικό σε κάθε διεύθυνση. Τα προγράμματα των spider διασφαλίζουν ότι τα δημοφιλέστερα προγράμματα ευρετηριάζονται πρώτα. Το ευρετήριο που δημιουργείται από τους spiders ελέγχεται σε σχέση με τον κατάλογο, έτσι ώστε να ελεγχθεί εάν νέες ιστοσελίδες πρέπει να προστεθούν, να διαγραφούν ή να αφαιρεθούν. Η Lycos σήμερα καταχωρεί τρία είδη αρχείων - αρχεία HTTP, αρχεία gopher και αρχεία FTP.

4.5.2. Crawling

Το τμήμα του προγράμματος που είναι υπεύθυνο για την ανίχνευση του ιστού αρχικά προήλθε από ένα πρόγραμμα αποκαλούμενο ***Longlegs***, που γράφτηκε από τους John Leavitt και Eric Nyberg στο πανεπιστήμιο *Carnegie Mellon*. Η Lycos χρησιμοποιεί ένα καινοτόμο, πιθανολογικό σχέδιο για τη μετάβαση από *κεντρικό υπολογιστή* ***(server)*** σε κεντρικό υπολογιστή στον ιστό. Αυτό βοηθά στην αποφυγή υπερφόρτωσης ενός κεντρικού υπολογιστή με μεγάλο αριθμό αιτημάτων, και επιτρέπει επίσης στη Lycos να δίνει προτεραιότητα σε *URLs* που κρίνονται πιο σημαντικές σε πληροφορίες. Τα βασικά βήματα του αλγορίθμου είναι τα εξής (Bikkannavar, 1999):

1. Όταν μια URL προσκομίζεται, η Lycos ανιχνεύει το περιεχόμενό της εάν πρόκειται για νέα URL, την οποία προσθέτει σε μια εσωτερική σειρά αναμονής.

2. Για να επιλέξει την επόμενη URL που θα εξερευνήσει, η Lycos κάνει μια τυχαία επιλογή μεταξύ των HTTP, gopher, και FTP αναφορών της λίστας αναμονής.

Η Lycos προτιμά να αναζητήσει τα δημοφιλή έγγραφα, δηλαδή εκείνα στα οποία καταλήγουν πολλά *links*. Η Lycos επίσης προτιμά τις μικρές URLs, οι οποίες είναι γενικά σημαντικοί κατάλογοι και έγγραφα πιο κοντά στην αρχική *ρίζα* ***(root)*** της ιεραρχίας. Σύμφωνα με τον Mauldin, η φιλοσοφία της Lycos έγκειται στο να κρατήσει ένα πεπερασμένο πρότυπο του ιστού το οποίο επιτρέπει να προχωρεί γρηγορότερα στις επόμενες αναζητήσεις. Η ιδέα είναι να μικρύνει το *δέντρο* ***(tree)*** των εγγράφων και να αντιπροσωπευθούν οι κομμένες άκρες αυτού του δέντρου με μια περίληψη των εγγράφων που βρίσκονται κάτω από τον συγκεκριμένο κόμβο. Οι 100 σημαντικότερες λέξεις από διάφορα έγγραφα μπορούν να συνδυαστούν για να παραγάγουν έναν

κατάλογο των 100 σημαντικότερων λέξεων ενός σύνολο εγγράφων. Η Lycos ακολουθεί τα πρότυπα για τον αποκλεισμό των ρομπότ, και οριοθετεί τη χρήση των HTTP πρακτόρων. Κατ' αυτό τον τρόπο, οι Webmasters είναι σε θέση να γνωρίζουν πότε η Lycos έχει επισκεφθεί τον κεντρικό υπολογιστή τους.

Η Lycos, εντούτοις, δεν ψάχνει και δεν συντάσσει ευρετήριο για εφήμερα *sites*, για *διαφορετικούς χρόνους* ***(time-varying)***, ή για τις άπειρες εικονικές περιοχές. Επομένως, η Lycos αγνοεί τις ακόλουθες περιοχές:

- ✓ Βάσεις δεδομένων WAIS (WAIS Databases)
- ✓ Ειδήσεις USENET (USENET News)
- ✓ Περιοχές Mailto (Mailto Space)
- ✓ Υπηρεσίες Telnet (Telnet Services)
- ✓ Τοπικές περιοχές αρχείων (Local File Space)

Η Lycos αγνοεί επίσης τα αρχεία που αρχίζουν με "/dev/tty" ή καταλήγουν με αυτές τις επεκτάσεις: AU, AVI, BIN, DAT, DVI, EXE, FLI, GIF, GZ, HDF, HQX, JPEG, LHA, MAC, MPEG, PS, TAR, TGA, TIFF, UU, UUE, WAV, Z, ή ZIP.

4.5.3. Ευρετήριο (Indexing)

Για να μειώσει το ποσό πληροφοριών που πρέπει να αποθηκευτεί, η Lycos εξάγει τις ακόλουθες πληροφορίες από κάθε έγγραφο που ανακτά (Bikkannavar, 1999):

- Τίτλους
- Κεφαλίδες και υπότιτλους
- Τις 100 σημαντικότερες λέξεις
- Τις πρώτες 20 γραμμές
- Μέγεθος εγγράφου σε bytes
- Τον αριθμό των λέξεων

Οι 100 σημαντικότερες λέξεις επιλέγονται χρησιμοποιώντας τον ***Tf*IDf αλγόριθμο απόδοσης βάρους σε κάθε λέξη***, που εξετάζει τη θέση και τη συχνότητα εμφάνισης της κάθε λέξης, μεταξύ και άλλων παραγόντων. Οι λέξεις, παραδείγματος χάριν, βαθμολογούνται από το πόσο πιο κάτω από την αρχή του εγγράφου εμφανίζονται. Κατά συνέπεια, οι λέξεις του τίτλου ή της πρώτης παραγράφου αποκτούν μεγαλύτερο βάρος.

Σε μια συλλογή Ν εγγράφων, η *συχνότητα εμφάνισης του όρου* ***(Term Frequency Tf)*** είναι ο αριθμός της εμφάνισης του συγκεκριμένου όρου στη συλλογή, και η *συχνότητα εγγράφων* ***(Document Frequency Df)*** είναι ο αριθμός των εγγράφων στη συλλογή στην οποία ο συγκεκριμένος όρος εμφανίζεται. Η ιδέα μιας *αντίστροφης συχνότητας εγγράφων* ***(Inverse Document Frequency IDf)*** μετρά πως οι συγκεκριμένοι όροι βοηθούν στο διαχωρισμό των εγγράφων, δηλαδή διακρίνουν τα λίγα έγγραφα στα οποία εμφανίζονται από τα περισσότερα στα οποία είναι απόντες. Ένας τυπικός παράγοντας *IDf* δίνεται από τον ***log(N/DF)***.

Στον *Tf*IDf* αλγόριθμο απόδοσης βαρών, η βασική ιδέα είναι ότι οι καλύτεροι όροι ευρετηρίασης είναι εκείνοι που εμφανίζονται συχνά σε ξεχωριστά έγγραφα, αλλά σπάνια στο υπόλοιπο της συλλογής. Η σημασία, ή το βάρος, ενός όρου ορίζεται έτσι ως το αποτέλεσμα του πολλαπλασιασμού της συχνότητας εμφάνισης του όρου (*TF*), με την αντίστροφη συχνότητα των εγγράφων *(IDf)*. Δηλαδή:

$$\textbf{\textit{Βάρος = TF * IDf = TF * log(N/DF)}}$$

Η Lycos δε λαμβάνει καθόλου υπ' όψιν της τα metatags.

4.6. Northernlight (www.northernlight.com)

4.6.1. Εισαγωγή

Η μηχανή NorthernLight αποτελεί σήμερα μία από τις καλύτερες επιλογές στις μηχανές γενικής αναζήτησης, καθώς προσφέρει ευκολίες και χαρακτηριστικά που δε διαθέτουν άλλες μηχανές. Η NorthernLight κατέχει κυρίαρχη θέση στις προτιμήσεις των χρηστών των προερχομένων κυρίως από τον επιχειρηματικό χώρο, αφού περιλαμβάνει όλα εκείνα τα χαρακτηριστικά που θα πρέπει να διαθέτει ένα σύγχρονο εργαλείο αναζήτησης, Η NorthernLight, συγκεντρώνοντας και ευρετηριάζοντας περισσότερες από 160 εκατομμύρια σελίδες web, πετυχαίνει το υψηλότερο ποσοστό κάλυψης του world wide web από οποιαδήποτε άλλη μηχανή ή εργαλείο αναζήτησης αυτή τη στιγμή. Παράλληλα, η NorthernLight διατηρεί μία από τις πιο ενημερωμένες και σύγχρονες βάσεις δεδομένων καθώς ανανεώνει τις εγγραφές της κάθε δύο εβδομάδες περίπου. Το γεγονός αυτό σε συνδυασμό με τα ειδικά χαρακτηριστικά και το φιλικό και εύχρηστο interface που διαθέτει της προσδίδουν ξεχωριστή αξία. Η NorthernLight δημιουργήθηκε το Σεπτέμβρη του 1995 και κατάφερε μέσα σε μία τριετία να καθιερωθεί ως μία από τις καλύτερες μηχανές αναζήτησης, ειδικότερα στον επιχειρηματικό χώρο, αφού οι 5.500 και πλέον πληροφοριακές πηγές που ευρετηριάζει απευθύνονται κυρίως στον επαγγελματία ***information broker***, όπως έχει καθιερωθεί να ονομάζεται ο *σύγχρονος διαχειριστής της πληροφορίας*.

4.6.2. Βασικά Χαρακτηριστικά

Η NorthernLight, μέσα από ένα ιδιαίτερα απλό αλλά ταυτόχρονα φιλικό interface, καταφέρνει να κερδίσει το χρήστη από την πρώτη σελίδα (Σχήμα 4.12.). Πέρα από *το βασικό πεδίο αναζήτησης* ***(query box)*** και τις απαραίτητες επιλογές περιορισμού της αναζήτησης, η NorthernLight στην κεντρική πρώτη σελίδα παρουσιάζει μια σειρά από ιδιαίτερα χαρακτηριστικά, που πραγματικά προσφέρουν υπηρεσίες και μοναδικές διευκολύνσεις. Τα δύο πιο σημαντικά από αυτά είναι η *Ειδική Βιβλιοθήκη των Επιχειρήσεων* ***(Business Research Library)*** που διαθέτει η NorthernLight καθώς και οι *Ειδικές Υπηρεσίες Αναζήτησης* ***(Custom Search Services)***. Η Ειδική Βιβλιοθήκη των Επιχειρήσεων που διαθέτει η NorthernLight θα μπορούσε να είναι μια βιβλιοθήκη online σύμφωνα με τον ορισμό της ίδιας της NorthernLight στην οποία συγκεντρώνονται και αποδελτιώνονται περισσότερες από 5.500 πληροφοριακές πηγές, όπως βιβλία, reports, περιοδικά, και newswires, καλύπτοντας έτσι ένα τεράστιο

θεματικό εύρος πληροφοριών και γνώσεων, τα οποία προσφέρει στο χρήστη με μικρό αντίτιμο. Σε ό,τι αφορά τις Ειδικές Υπηρεσίες Αναζήτησης, δεν είναι τίποτα περισσότερο από ***θεματικά directories***, όπου συγκεντρώνονται και ομαδοποιούνται τα αποτελέσματα της αναζήτησης με τέτοιο τρόπο ώστε να είναι εύκολα προσπελάσιμα και ανακτήσιμα από το χρήστη. Τα χαρακτηριστικά αυτά αποτελούν καινοτομίες στο χώρο των εργαλείων αναζήτησης, καθώς κατορθώνουν να προσφέρουν με φιλικό και εύχρηστο τρόπο επιχειρηματικές, επιστημονικές και άλλες αξιόπιστες πληροφορίες με μηδαμινό κόστος.

Σχήμα 4.11.: Η αρχική σελίδα (Homepage) της NorthernLight (www.northernlight.com)

Σε ό,τι αφορά τα υπόλοιπα χαρακτηριστικά της NorthernLight, ιδιαίτερη εντύπωση προκαλούν τα πολλαπλά interfaces και query boxes που διαθέτει για κάθε κατηγορία αναζήτησης που υποστηρίζει. Έτσι, υπάρχουν ξεχωριστές σελίδες για την αναζήτηση *Επιχειρηματικών Πληροφοριών* (Business Research). Έτσι, μπορεί να γίνει αναζήτηση στην ειδική βάση πληροφοριών Investext, η οποία συγκεντρώνει *εταιρικές εκθέσεις* ***(company reports και company profiles)*** και στοιχεία για χιλιάδες

επιχειρήσεις ανά τον κόσμο.

4.6.3. Ειδικά Χαρακτηριστικά

Λογική αναζήτησης και σύνταξης

Η NorthernLight υποστηρίζει τη χρήση όλων των τελεστών Boolean, AND, NOT και OR, όπως επίσης των συμβόλων "-" και "+". Επίσης αναγνωρίζει τον τελεστή εγγύτητας ***(proximity operator)*** NEAR. Η μηχανή χρησιμοποιεί τον τελεστή AND ως default, δηλαδή αυτόματα θα ελέγξει και τους όρους, εφόσον βέβαια τους βρει κατά την πορεία της αναζήτησης.

Αναζήτηση φράσεων

Η μηχανή υποστηρίζει πλήρως την αναζήτηση φράσεων. Θεωρεί τους όρους που περικλείονται μέσα σε ανωφερή εισαγωγικά "" ως φράση, την οποία και αναζητά αυτούσια.

Τελεστές περιορισμού αναζήτησης (Limit operators)

Η NorthernLight μπορεί να εστιάσει την αναζήτηση σε συγκεκριμένες περιοχές της σελίδας, όπως την ημερομηνία, τη γλώσσα του εγγράφου, ακόμη το domain name και τη γεωγραφική τοποθεσία. Ακόμη, υποστηρίζει πλήρως την αναζήτηση *ειδικών πεδίων* ***(Field Search)***.

Αποκοπή (Truncation)

Η NorthernLight υποστηρίζει πλήρως την αποκοπή των λέξεων προκειμένου να αναζητά τα παράγωγα των λέξεων προς αναζήτηση.

Έτσι, χρησιμοποιεί το χαρακτήρα «*» προκείμενου να αναζητήσει απεριόριστο αριθμό χαρακτήρων μετά τη ρίζα της λέξης, π.χ. book*, ενώ χρησιμοποιείτο χαρακτήρα «%» προκειμένου να αναζητήσει μόνο έναν χαρακτήρα μετά τη ρίζα της λέξης. Και τα δυο σύμβολα μπορούν να χρησιμοποιηθούν είτε στην αρχή είτε στο τέλος της λέξης. Η μηχανή υποστηρίζει επίσης την αναζήτηση με βάση τον αριθμό του όρου, δηλαδή το εάν πρόκειται για όρο στον ενικό ή τον πληθυντικό.

Αναγνώριση πεζών/ κεφαλαίων (Case Sensitivity)

Η μηχανή αναζήτησης υποστηρίζει την αναγνώριση πεζών και κεφαλαίων χαρακτήρων. Έτσι, στην περίπτωση που εισάγουμε τον αναζητήσιμο όρο με πεζούς χαρακτήρες, η μηχανή θα τον ανακτήσει αν είναι γραμμένος με πεζούς αλλά κι αν τον συναντήσει με κεφαλαίους χαρακτήρες. Στην περίπτωση που εισάγουμε τον όρο με κεφαλαίους χαρακτήρες, η μηχανή θα αναζητήσει και θα ανακτήσει τον όρο μόνο όπου τον συναντήσει με κεφαλαίους χαρακτήρες. Επιπλέον, κατά την παρουσίαση των αποτελεσμάτων η μηχανή θα κατατάξει υψηλότερα στην παρουσίαση τα αποτελέσματα εκείνα που εισήγαγε ο χρήστης με κεφαλαίους χαρακτήρες και ανακτήθηκαν.

4.6.4. Παρουσίαση Αποτελεσμάτων

Εμφάνιση και διάταξη των αποτελεσμάτων (Display & Order of results)

Στη NorthernLight τα αποτελέσματα της αναζήτησης παρουσιάζονται με σειρά αντιστρόφως ανάλογη της σχετικότητας, δηλαδή η πιο σχετική web page να εμφανίζεται στην αρχή της σελίδας. Κάθε εγγραφή περιλαμβάνει τον τίτλο του εγγράφου ή της web page που ανακτήθηκε, το είδος του εγγράφου, μια μικρή περίληψη, την ημερομηνία καθώς και τη *διεύθυνση* ***(URL)*** του *site* στο οποίο βρέθηκε. Στην *προ-χωρημένη αναζήτηση* ***(Power Search)*** υπάρχει η δυνατότητα ταξινόμησης των αποτελεσμάτων κατά ημερομηνία.

Ειδικά χαρακτηριστικά

Εκτός της απλής και της προχωρημένης αναζήτησης, η NorthernLight προσφέρει μια σειρά διευκολύνσεων προς το χρήστη, προκειμένου η αναζήτηση του να γίνει πιο εύκολη και αποδοτική. Οι σελίδες ειδικών αναζητήσεων News Search, Stock Search, Investext Search καθώς και η αναζήτηση στην Ειδική Συλλογή πληροφοριακών πηγών που διαθέτει η NorthernLight, σε συνδυασμό με τις επιλογές ως προς το είδος και τη μορφή των web pages που αναζητούνται, αποτελούν σημαντικούς περιοριστικούς παράγοντες του εύρους της αναζήτησης, μειώνοντας έτσι σημαντικά τα ανακτήσιμα αποτελέσματα και το χρόνο αναζήτησης.

4.7. Σύγκριση των Μηχανών Αναζήτησης

Είναι γεγονός ότι πάρα πολλές συγκρίσεις έχουν πραγματοποιηθεί από τους ερευνητές για τις μηχανές αναζήτησης, ώστε να μπορέσουν να συμπεράνουν ποιες από αυτές υπερέχουν και σε ποια σημεία έναντι των άλλων. Στην ενότητα 1.3.4 αναφέρονται αναλυτικά οι διάφορες μελέτες που έχουν γίνει. Γενικά μπορούμε να αναφέρουμε ότι έχουν γίνει συγκρίσεις για τους *χρόνους απόκρισης* ***(response times)*** *της κάθε μηχανής,* αφού εμφανίζονται να είναι στην κορυφή της λίστας των σημαντικότερων θεμάτων για τους χρήστες ιστού, καθώς επίσης και για τον αριθμό των πιο πολύτιμων-σχετικών σελίδων που εμφανίζονται στην πρώτη σελίδα των ανακτημένων αποτελεσμάτων (δηλ., τα κορυφαία 8, 10, ή 12 αποτελέσματα), έτσι ώστε το κουμπί που μας οδηγεί στην *επόμενη σελίδα* ***(next page)*** δεν θα είναι απαραίτητο να χρησιμοποιηθεί για την ανεύρεση των καλύτερων αποτελεσμάτων.

Άλλες συγκρίσεις βασίστηκαν στο βασικό πρότυπο των παραδοσιακών συστημάτων ανάκτησης πληροφοριών που αναγνωρίζει τη σχέση μεταξύ τριών παραγόντων: της *ταχύτητας* της ανάκτησης των πληροφοριών ***(speed)***, της *ακρίβειας* ***(precision)***, και της *ανάκλησης* ***(recall)*** (Gordon and Pathak, 1999).

Επίσης έχουν γίνει συγκρίσεις για το ποσοστό επικάλυψης του ιστού από την κάθε μηχανή ξεχωριστά, από το σύνολο των μηχανών αλλά και το ποσοστό επικάλυψης της κάθε μηχανής με τις άλλες μηχανές αναζήτησης (Lawrence and Giles, 1998).

Η σύγκριση των παραπάνω κύριων μηχανών αναζήτησης γενικού περιεχομένου πραγματοποιείται για κάθε σημαντικό τμήμα αυτών, όπως είναι το crawling, το ευρετήριο, η διάταξη και ταξινόμηση των σελίδων αποτελεσμάτων κ.τ.λ. Εκτός όμως από την παραπάνω σύγκριση θα γίνει και μια σύγκριση με βάση τα χαρακτηριστικά των μηχανών που έχουν σχεδιασθεί για το χρήστη. Αυτή η σύγκριση συνοψίζει τις κυριότερες εντολές της κάθε μηχανής καθώς και τα χαρακτηριστικά που διευκολύνουν τους χρήστες.

❖ Σύγκριση με Βάση το Crawling

Crawling	Ναι	Όχι	Παρατηρήσεις
Εξαντλητικό Crawl (Deep Crawl)	AltaVista, Google	Lycos	
Άμεση Ευρετηρίαση (Instant Indexing)	AltaVista	Lycos, Google	Οι σελίδες εμφανίζονται σε μια ή δυο μέρες μετά την καταχώρησή τους
Υποστήριξη Πλαισίων (Frames Support)	AltaVista, Google	Lycos	Lycos παρέχει την ελάχιστη υποστήριξη
Καταγραφή Εικόνων (Image Maps)	AltaVista	Google, Lycos	
Robots.txt	Όλες	-	
Meta Robots Tag	Όλες	-	
Πλήθος Συνδέσεων που Βοηθά στο Εξαντλητικό Crawl	Lycos	AltaVista	
Συχνότητα Αλλαγών της Σελίδας	AltaVista	Google, Lycos	

Πίνακας 4.2.: Η σύγκριση των μηχανών αναζήτησης με βάση το Crawling (Bikkannavar, 1999)

✓ **Εξαντλητικό Crawl (Deep Crawl)**

Οι μηχανές αναζήτησης που υποστηρίζουν το εξαντλητικό ή πλήρες Crawling θα τοποθετήσουν σε μια λίστα πολλές σελίδες από μια ιστοσελίδα, ακόμα και αν οι σελίδες αυτές δεν έχουν υποβληθεί ρητά σε αυτές. Οι υπόλοιπες μηχανές αναζήτησης θα απαριθμήσουν συνήθως πολύ λιγότερες σελίδες από ένα *site*. Γενικά, όσο μεγαλύτερο είναι το ευρετήριο μιας μηχανής αναζήτησης, το πιθανότερο να έχει πολλές σελίδες ανά *site*.

✓ **Άμεση Ευρετηρίαση (Instant Indexing)**

Σε μια άμεσης ευρετηρίασης μηχανή αναζήτησης, συνήθως οποιαδήποτε σελίδα υποβάλλεται σε αυτή θα εμφανιστεί μέσα σε μια ημέρα ή δύο μετά από την υποβολή.

✓ **Υποστήριξη Πλαισίων (Frames Support)**

Αυτό παρουσιάζει ποιες μηχανές αναζήτησης μπορούν να ακολουθήσουν τις συνδέσεις πλαισίων. Εκείνες που δεν μπορούν πιθανώς να μην απαριθμήσουν ένα μεγάλο μέρος της υποβληθείσας ιστοσελίδας.

✓ **Καταγραφή Εικόνων (Image Maps)**

Δείχνει ποιες μηχανές αναζήτησης μπορούν να ακολουθήσουν από την πλευρά του πελάτη την καταγραφή των εικόνων. Όπως με τα πλαίσια, εκείνες οι μηχανές αναζήτησης που δεν μπορούν να ακολουθήσουν τις εικόνες θα έχουν χάσει πιθανώς ένα μεγάλο μέρος του υποβληθείσας ιστοσελίδας.

✓ **robots.txt**

Το αρχείο robots.txt είναι ένας τρόπος για τους webmasters να κρατούν τις μηχανές αναζήτησης μακριά από τις ιστοσελίδες τους.

✓ **Meta Robots Tag**

Αυτό είναι ένα ειδικό *meta tag* που επιτρέπει στους ιδιοκτήτες ιστοσελίδων να καθορίσουν ότι μια σελίδα δεν πρέπει να ευρετηριασθεί. Είναι ιδανικό για εκείνους που δεν μπορούν να δημιουργήσουν ένα αρχείο *robots.txt*. Για να κρατήσουν τους *spiders* μακριά, απλά προσθέτουν αυτό το κείμενο μεταξύ των κεφαλίδων σε κάθε σελίδα που δεν θέλουν να ευρετηριαστεί:

< META NAME="ROBOTS" CONTENT="NOINDEX" >

Δεν είναι αναγκαίο να χρησιμοποιηθούν παραλλαγές αυτού του *tag* για να μπορεί να ευρετηριαστεί ευκολότερα η σελίδα. Και δεν είναι αναγκαίο να χρησιμοποιηθεί αυτό το *tag* εάν το αρχείο *robots.txt* χρησιμοποιείται ήδη.

✓ **Το Πλήθος των Συνδέσεων Βοηθά στο Εξαντλητικό Crawl (Link Popularity Helps Deep Crawl)**

Όλες οι μηχανές αναζήτησης μπορούν να καθορίσουν τη δημοτικότητα μιας σελίδας με την ανάλυση του αριθμού των *συνδέσεων **(links)*** που καταλήγουν σε αυτή από άλλες σελίδες. Μερικές μηχανές χρησιμοποιούν αυτό ως μέσο να καθορίσουν ποιες σελίδες θα συμπεριληφθούν στο ευρετήριο. Αυτό δεν είναι το ίδιο με την ταξινόμηση

των σελίδων υψηλότερα με βάση το μεγαλύτερο αριθμό συνδέσεων.

✓ **Συχνότητα Αλλαγών της Σελίδας (Learns Frequency)**

Διάφορες μηχανές αναζήτησης μπορούν να μάθουν πόσο συχνά οι σελίδες αλλάζουν (ανανεώνουν) τα περιεχόμενά τους. Σελίδες που αλλάζουν τα περιεχόμενά τους συχνά, επισκέπτονται περισσότερο.

❖ Σύγκριση με Βάση τον Τρόπο Ευρετηρίασης (Indexing)

Indexing	Ναι	Όχι	Παρατηρήσεις
Πλήρες Κείμενο (Full Body Text)	Όλες	-	Μερικές stop words μπορεί να μην ευρετηριάζονται
Stop Words	AltaVista, Lycos, Google	Καμιά	
Meta Description	Όλες,	Google, Lycos	
Meta Keywords	Όλες,	Google, Lycos	
ALT Text	AltaVista, Lycos	Google	
Σχόλια	Καμιά	Άλλες	
Stemming			

Πίνακας 4.3.: Η σύγκριση των μηχανών αναζήτησης με βάση το Indexing (Bikkannavar, 1999)

✓ **Ευρετηρίαση με Βάση το Σύνολο των Λέξεων (Full Body Text)**

Όλες οι σημαντικές μηχανές αναζήτησης λένε ότι συντάσσουν ευρετήριο με βάση το πλήρες κείμενο μιας σελίδας, αν και μερικές δεν ευρετηριάζουν τις ***stop words*** ή αποκλείουν όμοιες σελίδες που κρίνονται ότι ανήκουν στην κατηγορία του *spamming*.

✓ **Stop Words**

Μερικές μηχανές αναζήτησης παραλείπουν λέξεις όταν ευρετηριάζουν μια σελίδα ή μπορεί να μην αναζητήσουν αυτές τις λέξεις σε ένα ερώτημα. Αυτές οι *stop words* παραβλέπονται με σκοπό την εξοικονόμηση αποθηκευτικού χώρου ή για την επιτάχυνση των αποτελεσμάτων αναζήτησης.

✓ **Meta Description & Meta Keywords**

Δείχνει ποιες μηχανές αναζήτησης υποστηρίζουν την περιγραφή των meta και των meta λέξεις-κλειδιά. Δε σημαίνει ότι η χρήση αυτών ταξινομεί τις σελίδες υψηλότερα. Αυτό αναφέρεται παρακάτω.

✓ **Κείμενο ALT/ Σχόλια (ALT Text/Comments)**

Αυτό δείχνει ποιες μηχανές αναζήτησης συντάσσουν ευρετήριο με βάση το κείμενο ALT που συνδέεται με εικόνες ή με κείμενο σχολίων.

✓ **Stemming**

Μερικές μηχανές αναζήτησης ψάχνουν για παραλλαγές μιας λέξης βασισμένες στη ρίζα της. Παραδείγματος χάριν, θέτοντας τη λέξη "κολύμπι" μπορεί να ψάξει και να εμφανίσει τις λέξεις "κολυμπώ" και ίσως "κολυμπώντας", ανάλογα με τη μηχανή αναζήτησης.

❖ Σύγκριση με Βάση τον Τρόπο Ταξινόμησης (Ranking)

Οι περισσότερες μηχανές αναζήτησης κάνουν χρήση της θέσης και της συχνότητας εμφάνισης των λέξεων-κλειδιών σε μια ιστοσελίδα με στόχο την ταξινόμησή της. Ο ακριβής μηχανισμός λειτουργεί λίγο διαφορετικά για κάθε μηχανή αναζήτησης.

Ranking	Ναι	Όχι	Παρατηρήσεις
Meta Tags βοηθούν στην Ταξινόμηση	Καμιά	AltaVista, Lycos, Google	
Reviewed Status βοηθά στην Ταξινόμηση	Καμιά	AltaVista, Lycos, Google	
Το πλήθος συνδέσεων βοηθά στην Ταξινόμηση	AltaVista, Google	Lycos	Πολύ σημαντικό για το Google
Direct Hit βοηθά στην Ταξινόμηση	Καμιά	Όλες	

Πίνακας 4.4.: Σύγκριση των μηχανών με βάση τα χαρακτηριστικά ταξινόμησης (Bikkannavar, 1999)

Σε αντίθεση της θέσης και της συχνότητας, μερικές μηχανές αναζήτησης δίνουν μεγαλύτερη σημασία σε άλλους παράγοντες. Αυτοί οι παράγοντες βοηθούν λίγο και δεν

εγγυώνται την καλύτερη δυνατή ταξινόμηση. Μερικοί σημαντικοί παράγοντες φαίνονται στον πίνακα 4.4.

✓ **Meta Tags Βοηθούν στην Καλύτερη Ταξινόμηση (Meta Tags Boost Ranking)**

Μερικές μηχανές αναζήτησης που υποστηρίζουν την περιγραφή των meta και των λέξεων-κλειδιών θα ταξινομήσουν καλύτερα μια σελίδα εάν οι αναζητήσιμοι όροι εμφανίζονται σε αυτές τις περιοχές. Όλες οι μηχανές αναζήτησης που υποστηρίζουν αυτά τα *tags* δεν ταξινομούν υψηλότερα τις σελίδες.

✓ **Reviewed Status βοηθά στην Ταξινόμηση (Reviewed Status Boosts Rankings)**

Μερικές μηχανές αναζήτησης επισκέπτονται τις ιστοσελίδες ή τις καταχωρούν σε έναν σχετικό κατάλογο. Μπορούν να ταξινομήσουν αυτές τις σελίδες υψηλότερα.

✓ **Το πλήθος Συνδέσεων βοηθά στην Ταξινόμηση (Link Popularity Boosts Rankings)**

Όπως περιγράφεται παραπάνω, όλες οι μηχανές αναζήτησης μπορούν να καθορίσουν τη δημοτικότητα μιας σελίδας αναλύοντας πόσες συνδέσεις (αναφορές) καταλήγουν σε αυτές από άλλες σελίδες. Μερικές μηχανές ταξινομούν υψηλότερα σελίδες με πολλές συνδέσεις, ή συνδέσεις από σημαντικές ιστοσελίδες.

✓ **Direct Hit βοηθά στην Ταξινόμηση (Direct Hit Boosts Rankings)**

Το Direct Hit είναι ένα σύστημα που μετρά ποιοι χρήστες κάνουν *click* στη σελίδα που βρίσκεται στα αποτελέσματα της αναζήτησης προκειμένου να επαναπροσδιοριστεί ο βαθμός σχετικότητας της σελίδας.

❖ Σύγκριση με Βάση το Spamming

Όλες οι σημαντικές μηχανές αναζήτησης απορρίπτουν την ευρετηρίαση σελίδων που κάνουν χρήση τεχνικών ***"spam"*** για τη βελτίωση της θέσης τους. Μια κοινή τεχνική είναι η "συσσώρευση" ή το "γέμισμα" λέξεων σε μια σελίδα. Αυτό γίνεται όταν μια λέξη επαναλαμβάνεται πολλές φορές σε μια σειρά. Εάν μια μηχανή αναζήτησης εντοπίσει μια τεχνική *spamming*, μπορούν να ταξινομήσουν τη σελίδα χαμηλότερα ή να την διαγράψουν από τις λίστες ολοκληρωτικά. Τα παρακάτω στοιχεία θα μπορούσαν να προκαλέσουν να θεωρηθούν τεχνικές *spam*.

Spam	Ναι	Όχι
Meta Refresh	AltaVista, Lycos	Google
Αόρατο Κείμενο (Invisible Text)	Άλλες	Google
Μικροσκοπικό Κείμενο	AltaVista, Lycos	Google

Πίνακας 4.5.: Σύγκριση των μηχανών με βάση τα χαρακτηριστικά spam (Bikkannavar, 1999)

✓ **Meta Refresh**

Μερικοί ιδιοκτήτες ιστοσελίδων δημιουργούν σελίδες-στόχου ***(target pages)*** που οδηγούν αυτόματα τους επισκέπτες σε διαφορετικές σελίδες μέσα σε ένα *web site*. Το ***meta refresh tag*** είναι ένας χαρακτηριστικός τρόπος να επιτευχθεί το παραπάνω. Κάποιες μηχανές αναζήτησης δεν ευρετηριάζουν μια σελίδα με υψηλό δείκτη *meta refresh*. Το Google δεν λαμβάνει υπόψη πολύ το *meta refresh* ή τα παρακάτω στοιχεία επειδή το σύστημα ταξινόμησης με βάση το πλήθος των συνδέσεών εντοπίζει τις τεχνικές του *spam*.

✓ **Αόρατο Κείμενο (Invisible Text)**

Είναι η τεχνική της εισαγωγής κειμένου σε μια σελίδα με χρώμα ίδιο με το ***background*** καθιστώντας το αόρατο στους ανθρώπινους θεατές. Πολλές μηχανές αναζήτησης δεν ευρετηριάζουν αυτό το κείμενο ή δεν ευρετηριάζουν οποιαδήποτε σελίδα που περιέχει αόρατο κείμενο.

✓ **Μικροσκοπικό Κείμενο (Tiny Text)**

Είναι η τεχνική της εισαγωγής κειμένου σε μια σελίδα με πολύ μικρό μέγεθος γραμμάτων. Σελίδες που είναι γεμάτες με τέτοιου τύπου κείμενα μπορεί να απομακρυνθούν ως *spam* ή το μικροσκοπικό κείμενο να μην ευρετηριαστεί. Γενικά δεν ευρετηριάζονται σελίδες με κείμενα όπου το μέγεθος των γραμμάτων τους είναι κυρίως μικρότερο από το κανονικό.

Οι παρακάτω πίνακες συγκρίνουν τα χαρακτηριστικά των μηχανών που έχουν σχεδιασθεί για το χρήστη. Αυτή η σύγκριση συνοψίζει τις κυριότερες εντολές της κάθε μηχανής καθώς και τα χαρακτηριστικά που διευκολύνουν τους χρήστες.

❖ **Εντολές Μαθηματικών των Μηχανών Αναζήτησης**

Εντολές	Με Ποίον Τρόπο	Υποστηρίζονται από
Πρόσθεση Όρου	+	Όλες
Αφαίρεση Όρου	-	Όλες
Φράση	“ ”	Όλες (Παρατήρηση: Ημιαυτόματα στην AltaVista, Google)
Σύγκριση Κάθε Όρου	Αυτόματα	AltaVista, Infoseek
	Από το menu	Lycos
	Αλλιώς	Google
Σύγκριση Όλων των Όρων	Αυτόματα	Google, Lycos
	Αλλιώς	Επιτυγχάνεται σε όλες με χρήση του τελεστή + η με επιλογές από το menu

Πίνακας 4.6.: Σύγκριση των μηχανών με βάση τις εντολές μαθηματικών (Bikkannavar, 1999)

❖ **Εντολές για Καλύτερη Αναζήτηση**

Εντολή	Τρόπος	Υποστηρίζεται από
Αναζήτηση με βάση τον τίτλο	Τίτλος	AltaVista, Infoseek
	Διαφορετικός	Μέσω menus, Lycos
	Κανένας	Google
Αναζήτηση Σελίδας	Πεδίο	Καμιά
	Διαφορετικός	AltaVista, Infoseek, Lycos
	Κανένας	Google
Αναζήτηση URL	URL	AltaVista, Infoseek
	Διαφορετικός	Lycos
	Κανένας	Google
Αναζήτηση Συνδέσεων	Συνδέσεων	AltaVista, Infoseek, Google
	Linkdomain	Καμιά
	Κανένας	Google
Wildcard	*	AltaVista
	Κανένας	Infoseek, Google, Lycos

Πίνακας 4.7.: Σύγκριση των μηχανών με βάση τις εντολές για επίτευξη καλύτερης αναζήτησης (Bikkannavar, 1999)

❖ **Βοηθητικά Χαρακτηριστικά Αναζήτησης**

Χαρακτηριστικά	Ναι	Όχι	Παρατηρήσεις
Σχετικές Αναζητήσεις	AltaVista, Infoseek	Άλλες	
Ομαδοποίηση	Infoseek	Άλλες	Google έχει μερικά χαρακτηριστικά ομαδοποίησης
Βρίσκουν τις όμοιες	Infoseek	Άλλες	
Stemming	Infoseek, Lycos	AltaVista	
Εύρος ημερομηνιών	AltaVista	Άλλες	
Αναζήτηση στα αποτελέσματα	Infoseek, Lycos	Άλλες	
Αναγνώριση Κεφαλαίων	AltaVista, Infoseek	Lycos	
Direct Hit	Lycos	Άλλες	

Πίνακας 4.8.: Σύγκριση των μηχανών με βάση τα βοηθητικά χαρακτηριστικά αναζήτησης (Bikkannavar, 1999)

❖ **Χαρακτηριστικά Εμφάνισης**

Χαρακτηριστικά	Ναι	Όχι	Παρατηρήσεις
Ταξινόμηση με βάση την ημερομηνία	Infoseek	Άλλες	
Εμφάνιση Ημερομηνίας	AltaVista, Infoseek	Άλλες	
Αύξηση αριθμού των αποτελεσμάτων	Infoseek, Lycos,Google	AltaVista	

Πίνακας 4.9.: Σύγκριση των μηχανών με βάση τα χαρακτηριστικά εμφάνισης (Bikkannavar, 1999)

❖ **Boolean εντολές**

Εντολές	Τρόπος	Υποστηρίζεται από:
Or	OR	Όλες εκτός…
	Κανένας	Infoseek, Google
And	AND	Όλες εκτός…
	Κανένας	Infoseek, Google
Not	NOT	Όλες εκτός…
	AND NOT	AltaVista
	Κανένας	Infoseek, Google
Πολλοί όροι	( )	Όλες εκτός…
	Κανένας	Infoseek, Google
Κοντά	NEAR	AltaVista (10 λέξεις), Lycos (25 λέξεις)
	Κανένας	Άλλες

Πίνακας 4.10.: Σύγκριση των μηχανών με βάση τα Boolean χαρακτηριστικά τους (Bikkannavar, 1999)

Συμπερασματικά θα μπορούσαμε να σημειώσουμε ότι με το πέρασμα του χρόνου οι μηχανές αναζήτησης έχουν εξελιχθεί σε πολύ πολύπλοκα εργαλεία αναζήτησης του διαδικτύου. Κατά τη διάρκεια των πρώτων ετών, μια μηχανή αναζήτησης κρινόταν και θεωρούνταν υψηλού επιπέδου από το ποσοστό του ιστού που είχε συντάξει το ευρετήριό της. Τώρα όμως δόθηκε έμφαση στην ύπαρξη του ποιοτικότερου ευρετηρίου, παρά του μεγαλύτερου σε μέγεθος ευρετηρίου. Αυτό σημαίνει συνήθως ότι είναι μεγάλο σε μέγεθος το ευρετήριο, επειδή οι περισσότερες από τις μηχανές αναζήτησης θέλουν να έχουν μια πλήρη εικόνα του τι υπάρχει στον ιστό. Αλλά αυτό σημαίνει επίσης, ότι η πλήρης ενημέρωσή του είναι δύσκολο να επιτευχθεί, όταν εκατομμύρια ιστοσελίδων πρέπει να ελεγχθούν. Καμία από τις μηχανές αναζήτησης δεν είναι σε θέση να ευρετηριάζει τα πάντα που υπάρχουν στον ιστό. Καμία μηχανή αναζήτησης δεν μπορεί να υποστηρίξει ότι έχει ένα τέλειο αρχείο όλων των πληροφοριών. Βελτιώνοντας τη σχετικότητα των αποτελεσμάτων της αναζήτησης σε σχέση με το ερώτημα του χρήστη, σημαίνει μια εξυπνότερη μηχανή αναζήτησης, και όχι μόνο μια μεγαλύτερη μηχανή. Το πλήθος των πληροφοριών που μπορεί γρήγορα και αποτελεσματικά να αναζητηθεί με τη χρήση των μηχανών αναζήτησης έχει αυξηθεί, αλλά αυτό που έχει μειωθεί είναι το πλήθος που μπορεί να αναζητηθεί έναντι σε αυτό

που ενδεχόμενα θα μπορούσε να αναζητηθεί.

ΚΕΦΑΛΑΙΟ 5

ΟΙ ΜΗΧΑΝΕΣ ΠΟΛΛΑΠΛΗΣ ΑΝΑΖΗΤΗΣΗΣ

Διάφορες χρήσιμες και δημοφιλείς μηχανές αναζήτησης προσπαθούν να διατηρήσουν ένα ευρετήριο με βάση όλο το περιεχόμενο ή το κείμενο των ιστοσελίδων του World Wide Web (π.χ. AltaVista (www.altavista.digital.com), Excite (www.excite.com), HotBot (www.hotbot.com), Infoseek (www.infoseek.com), Lycos (www.lycos.com), και Northern Light (www.nlsearch.com)). Εντούτοις, η αναζήτηση του ιστού είναι ακόμα μια αργή και κουραστική διαδικασία. Οι περιορισμοί των υπηρεσιών αναζήτησης έχουν οδηγήσει στην εισαγωγή των μηχανών πολλαπλής αναζήτησης, όπως π.χ. MetaCrawler και SavvySearch. Μια μηχανή πολλαπλής αναζήτησης ψάχνει τον ιστό υποβάλλοντας τα ερωτήματα στις μηχανές αναζήτησης όπως AltaVista ή Infoseek.

Κατά τη διάρκεια του έτους 2000 (Ιούνιος 2000) υπήρχαν τουλάχιστον 3.500 διαφορετικές διαθέσιμες μηχανές αναζήτησης γενικού περιεχομένου, αλλά και συγκεκριμένων θεμάτων, ή αναζητούσαν συγκεκριμένα στοιχεία του διαδικτύου όπως ιστοσελίδες ή USENET.

Ενώ μερικές από αυτές ήταν ιδιαίτερα αποτελεσματικές και περίπλοκες καμιά από αυτές όμως δεν είναι απόλυτα κατανοητή. Μπορεί να χρησιμοποιήσουν μια μικρή βάση δεδομένων από την οποία να δημιουργήσουν το σύνολο των αποτελεσμάτων στις ερωτήσεις του χρήστη (παραδείγματος χάριν το Yahoo ευρετηριάζει ένα πολύ μικρό ποσοστό σε σχέση με τα 350.000.000 των σελίδων που ευρετηριάζονται από την

Altavista), ή δεν ενημερώνονται ιδιαίτερα γρήγορα (η Altavista ενημερώνεται κάθε 9-10 ημέρες, ενώ η Lycos ενημερώνεται σε ώρες). Το λογισμικό των *spiders* μπορεί να μην είναι πολύ γρήγορο (είναι πιθανόν το *spider* της Excite να χρειαστεί 28 ημέρες για να ολοκληρώσει την εργασία του, σε σύγκριση με το spider της Magellen που χρειάζεται 4 ημέρες), το οποίο σημαίνει ότι η απάντησή τους δεν ανταποκρίνεται στην πραγματική κατάσταση του διαδικτύου.

Τα κυριότερα πλεονεκτήματα των μηχανών πολλαπλής αναζήτησης είναι η δυνατότητα να συνδυαστούν τα αποτελέσματά τους και η δυνατότητα να παρέχουν ένα συνεπές *user interface* για την έρευνα αυτών των μηχανών.

Ακόμα κι αν ένας χρήστης έχει μια ή περισσότερες αγαπημένες μηχανές αναζήτησης, για να εξασφαλίσει μια περιεκτική αναζήτηση ίσως πρέπει να χρησιμοποιήσει αρκετές μηχανές πριν να πεισθεί ότι έχει βρει ότι απαιτείται για ένα συγκεκριμένο θέμα. Μια μηχανή πολλαπλής αναζήτησης μπορεί να του λύσει το πρόβλημα της μετάβασης σε ποικίλα διαφορετικά sites προκειμένου να πραγματοποιηθεί η αναζήτηση, ή μπορεί να προτείνει μια μηχανή αναζήτησης που δεν έχει εξεταστεί, ή ίσως δεν ήξερε ακόμη και την ύπαρξή της.

Σε γενικές γραμμές οι συνδυαστικές μηχανές αναζήτησης έχουν τρεις παραλλαγές στον τρόπο με τον οποίο λειτουργούν, όπως αναφέρθηκε και στην ενότητα 3.4.1 της παρούσας εργασίας:

- **ΣΕΙΡΙΑΚΗ ΑΝΑΖΗΤΗΣΗ (SERIAL SEARCH)**: Μια *meta-search engine* μπορεί να αναζητά τους ζητούμενους όρους ενεργοποιώντας τη μία μηχανή μετά την άλλη
- **ΤΑΥΤΟΧΡΟΝΗ ΑΝΑΖΗΤΗΣΗ (SIMULTANEOUS SEARCH):** Ενεργοποιώντας όλες τις μηχανές ταυτόχρονα και
- **ΣΥΓΚΕΝΤΡΩΣΗ ΣΥΣΤΗΜΑΤΩΝ ΑΝΑΖΗΤΗΣΗΣ**: Παραθέτοντας απλά τη λίστα με όλες τις διαθέσιμες μηχανές και τα εργαλεία αναζήτησης, προκειμένου ο χρήστης να διαμορφώσει το ερώτημά του σε όποια από αυτές τελικά επιλέξει.

Υπάρχουν βασικά δύο προσεγγίσεις για τον προσδιορισμό του προβλήματος των μηχανών πολλαπλής αναζήτησης (Hock, 2001). Η μια είναι *οι ιστοσελίδες πολλαπλής αναζήτησης* ***(meta-search sites)***, στις οποίες η πρόσβαση είναι ελεύθερη. Η άλλη είναι η χρήση του πελάτη ***(client)***, το οποίο είναι ένα *πρόγραμμα πολλαπλής αναζήτησης* ***(meta-search program)*** που βρίσκεται στον υπολογιστή του κάθε χρήστη και στοχεύει στην αναζήτηση των μηχανών πολλαπλής αναζήτησης. Οι σελίδες πολλαπλής αναζήτησης

είναι ελεύθερες και εύκολες στη χρήση τους, αλλά έχουν σημαντικά μειονεκτήματα ως προς την πληρότητα με την οποία επιτελούν την εργασία. Τα *"client-side"* προγράμματα επιτελούν μια πληρέστερη εργασία, αλλά περιλαμβάνουν το downloading (ή την αγορά) ενός προγράμματος και αρκετά ακόμη βήματα για τα τελικά αποτελέσματα. Θα εξετάσουμε κάθε μια από τις δύο προσεγγίσεις.

❖ ***Meta-Search sites***

Οι ελεύθερες ιστοσελίδες έχουν ως κύριο πλεονέκτημα την ευκολία στην χρήση τους, χωρίς την ανάγκη οποιουδήποτε λογισμικού, αλλά υπάρχουν σημαντικά μειονεκτήματα.

Τα μειονεκτήματα είναι εμφανέστερα μέσω ενός παραδείγματος. Συγκεκριμένα, εάν υπάρχουν περισσότερα των σχετικών sites που ανακτώνται από τις μηχανές αναζήτησης του ιστού, οι μηχανές πολλαπλής αναζήτησης δεν εντοπίζουν τα περισσότερα από αυτά. Αυτό οφείλεται σε διάφορους παράγοντες, συμπεριλαμβανομένων των ορίων που επιβάλλονται από την υπηρεσία στον αριθμό των αρχείων που ανακτώνται από την κάθε μηχανή ξεχωριστά. Καθώς και στον τερματισμό της αναζήτησης όπου η υπηρεσία πολλαπλής αναζήτησης σταματά απλά την αναζήτηση όταν είναι πολύ χρονοβόρα. Επίσης, στην αποτυχία να μεταφραστεί επαρκώς η ερώτηση στη συγκεκριμένη σύνταξη που απαιτείται από τη μηχανή αναζήτησης, και τέλος σε άλλους παράγοντες. Υπάρχουν όμως και μερικές μηχανές πολλαπλής αναζήτησης που επιστρέφουν όλα τα αρχεία που είναι πραγματικά στον ιστό (αλλά αυτό έχει άλλα μειονεκτήματα).

Τα τρία σημαντικότερα μειονεκτήματα των μηχανών πολλαπλής αναζήτησης είναι (Hock, 2001):

1. Περιορίζουν συχνά τον αριθμό των αρχείων που θα ανακτηθούν από κάθε μηχανή αναζήτησης (μερικές φορές σε 10)
2. Συχνά δεν μεταφέρουν ακόμη και τις λίγο διαφοροποιημένες ερωτήσεις στις μηχανές αναζήτησης και
3. Στις περισσότερες περιπτώσεις, δεν ευρετηριάζουν περισσότερο από δύο ή τρεις από τις πέντε μεγαλύτερες μηχανές αναζήτησης.

Η δημιουργία ενός τέτοιου site δεν είναι δύσκολη, η οποία βοηθά τον απολογισμό για το μεγάλο αριθμό τους. Πάνω από 100 κατηγορίες του Yahoo! υπάρχουν σε αυτές τις μηχανές. Μερικές είναι βασικά μια συλλογή από *βασικά πεδία αναζήτησης* ***(search***

boxes) που έχουν αντιγραφεί και έχουν επικολληθεί από τις διάφορες μηχανές αναζήτησης. Μερικές ασχολούνται αρκετά περαιτέρω με αυτά και εξετάζουν τουλάχιστον ένα ή δύο από τα προβλήματα που αναφέρθηκαν προηγουμένως.

Συνήθως, οι μηχανές πολλαπλής αναζήτησης διαφέρουν μεταξύ τους στα ακόλουθα σημεία (Hock, 2001):

- Στον αριθμό των μηχανών αναζήτησης που χρησιμοποιούν
- Στον αριθμό των μηχανών αναζήτησης που μπορούν να αναζητήσουν στον ιστό ταυτοχρόνως
- Στη δυνατότητά τους να μεταφέρουν περισσότερο περίπλοκες ερωτήσεις στις μηχανές αναζήτησης, όπως συμπεριλαμβανομένων φράσεων, χρήσεις των τελεστών Boolean, κ.λ.π.
- Στον μέγιστο αριθμό των αρχείων που μπορούν να ανακτούν από κάθε μηχανή (που το ελάχιστο είναι 10)
- Στο χρονικό διάστημα που είναι πρόθυμοι να σπαταλήσουν για την αναζήτηση μιας ερώτησης κάθε μηχανή (πριν από τον τερματισμό)
- Στην εμφάνιση των αποτελεσμάτων, περιέχουν ή όχι τις διπλές εγγραφές της κάθε μηχανής.

Οι μηχανές πολλαπλής αναζήτησης είναι πιο χρήσιμες και αποτελεσματικές όταν αναζητείται κάτι πολύ δύσκολο. Εάν για τη συγκεκριμένη αναζήτηση υποτίθεται ότι υπάρχουν λιγότερα από 10 sites (ή εάν ο χρήστης δεν ενδιαφέρεται να προσδιορίσει περισσότερα από 10 sites), και εάν το ερώτημα αποτελείται μόνο από μια λέξη ή μια φράση. Υπάρχουν πολλά αρχεία που ανακτώνται από μερικές από τις μικρότερες μηχανές και που δεν ανακτώνται από τις τρεις ή τέσσερις μεγαλύτερες μηχανές, και μπορεί να αποβεί χρονοβόρο η αναζήτηση ξεχωριστά για όλες τις σημαντικές μηχανές. Οι μηχανές πολλαπλής αναζήτησης πράγματι επιτρέπουν την πολύ γρήγορη ανίχνεύση ενός μεγάλου αριθμού μηχανών για τέτοιου είδους αναζητήσεις.

- ***Client Meta-Search Programs***

Για την αναζήτηση με μηχανές πολλαπλής αναζήτησης, η εναλλακτική λύση στις *ιστοσελίδες πολλαπλής αναζήτησης* ***(meta-search sites)*** είναι η χρήση του προγράμματος ***client*** πολλαπλής αναζήτησης. Αυτό είναι ένα πρόγραμμα για τον υπολογιστή που επιτελεί την ίδια εργασία με έναν *ευφυή πράκτορα* ***(Intelligent Agent)*** που ειδάλλως θα

γινόταν εάν χρησιμοποιούσαμε τις διάφορες μηχανές.

Η ιδέα της χρήσης αυτών των ευφυών πρακτόρων, που μπορούν να εκτελέσουν διάφορα βήματα παραπάνω από τις μηχανές πολλαπλής αναζήτησης, έχουν μελετηθεί εκτενώς, και διάφορα τέτοια προγράμματα έχουν αναπτυχθεί. Τα πιο ξεχωριστά μεταξύ αυτών των προγραμμάτων είναι το Copernic και το BullsEye, τα οποία με τη βοήθεια ενός προγράμματος που γίνεται *download* στον υπολογιστή του χρήστη, πραγματοποιούν τις αναζητήσεις με διάφορες μηχανές αναζήτησης του ιστού, ταξινομώντας τα αποτελέσματα επιτρέπουν την περαιτέρω τοπική έρευνα και εκτελούν διάφορες σχετικές ενέργειες.

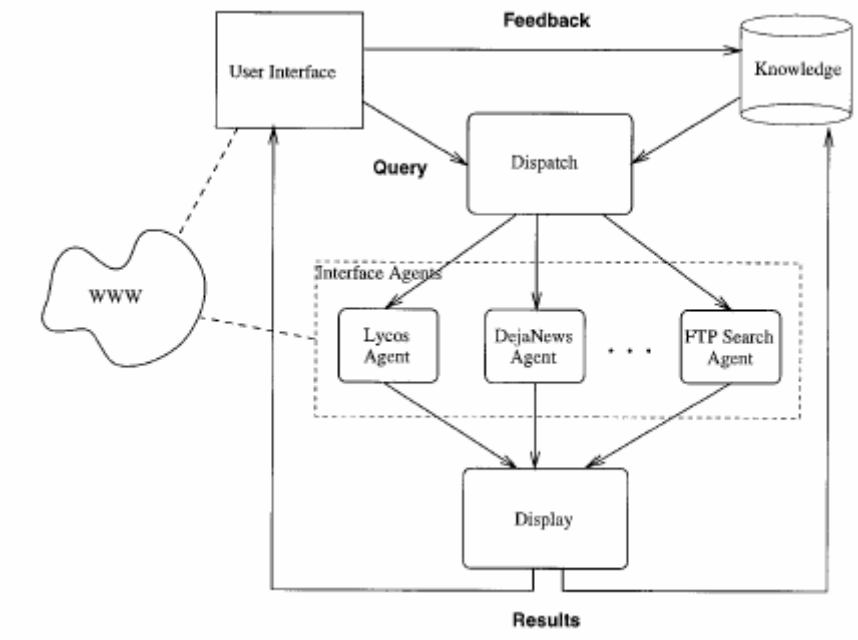
Αυτά τα προγράμματα, σε διαφορετικό βαθμό, έχουν το πλεονέκτημα ότι καλύπτουν περισσότερες μηχανές από οποιαδήποτε από τα τρέχων *sites* πολλαπλής αναζήτησης, επιτρέπουν τον καλύτερο χειρισμό των αποτελεσμάτων και επιτρέπουν την περαιτέρω επεξεργασία, αφού προηγουμένως έχει γίνει η αποσύνδεση του χρήστη από το διαδίκτυο. Είναι βέβαιο ότι η μέγιστη σημασία και η χρησιμότητά τους βρίσκονται στη διαχείριση των στοιχείων αφότου έχουν βρεθεί από τις μηχανές αναζήτησης. Το σημαντικότερο μειονέκτημα είναι ότι παρουσιάζουν ένα παραπάνω βήμα, περιλαμβάνοντας συχνά πολλά περισσότερα *clicks* μεταξύ του αναζητητή και του προϊόντος. Εάν το προϊόν είναι ο εντοπισμός της θέσης των απαραίτητων σχετικών *sites*, αυτά τα εργαλεία μπορούν να είναι κατά ένα μεγάλο μέρος περιττά για πολλούς ειδικευμένους χρήστες μηχανών αναζήτησης. Εάν το προϊόν είναι μια οργανωμένη και κατευθυνόμενη συλλογή των ανακτημένων *sites*, αυτά τα προγράμματα μπορούν να είναι μια πολύτιμη βοήθεια.

Δεν είναι εφικτό να καλυφθούν όλές οι μηχανές πολλαπλής αναζήτησης, που είναι παραπάνω από εκατό σε αυτό το κεφάλαιο. Θα εξετάσουμε τέσσερις από τις δημοφιλέστερες και αντιπροσωπευτικότερες αναλυτικά και θα γίνει και μια περιληπτική αναφορά σε κάποιες άλλες μηχανές πολλαπλής αναζήτησης, ώστε να καλυφθεί όλο το φάσμα των μηχανών πολλαπλής αναζήτησης. Η επόμενη παράγραφος κάνει αναφορά στην αρχιτεκτονική των μηχανών πολλαπλής αναζήτησης που οι περισσότερες μηχανές αυτού του είδους χρησιμοποιούν.

5.1. Αρχιτεκτονική των Μηχανών Πολλαπλής Αναζήτησης

Η αρχιτεκτονική των μηχανών πολλαπλής αναζήτησης εξετάζεται με βάση τους τρεις παρακάτω μηχανισμούς (Dreilinger and Howe, 1997):

- *Μηχανισμός αποστολών* ***(Dispatch Mechanism):*** Αυτός είναι ο αλγόριθμος, ή η προσέγγιση της λήψης αποφάσεων, που καθορίζει σε ποιες μηχανές αναζήτησης μια συγκεκριμένη ερώτηση θα σταλεί.
- *Interface Πράκτορες* ***(Interface Agents)****:* Αυτά τα ανεξάρτητα προγράμματα διαχειρίζονται την αλληλεπίδραση με μια μηχανή αναζήτησης. Οι *interface πράκτορες* προσαρμόζουν- διαμορφώνουν την ερώτηση του χρήστη ώστε να ταιριάζει με το πρότυπο της μηχανής αναζήτησης. Οι interface πράκτορες είναι επίσης αρμόδιοι για την ερμηνεία των διαφορετικών ***formats*** των αποτελεσμάτων.
- *Μηχανισμός Εμφάνισης* ***(Display Mechanism)****:* Τα μη διαμορφωμένα αποτελέσματα από την κάθε μηχανή αναζήτησης πρέπει να ενσωματωθούν ώστε να εμφανιστούν στον χρήστη. Τα αποτελέσματα μπορούν να εμφανισθούν με λίγη πρόσθετη μορφοποίηση και μπορεί να είναι ταξινομημένα. Στα αποτελέσματα μπορούν να διαγραφούν οι διπλές εγγραφές ή να επαληθευτούν οι συνδέσεις.



Σχήμα 5.1.: Αρχιτεκτονική των Μηχανών Πολλαπλής Αναζήτησης (Dreilinger and Howe, 1997)

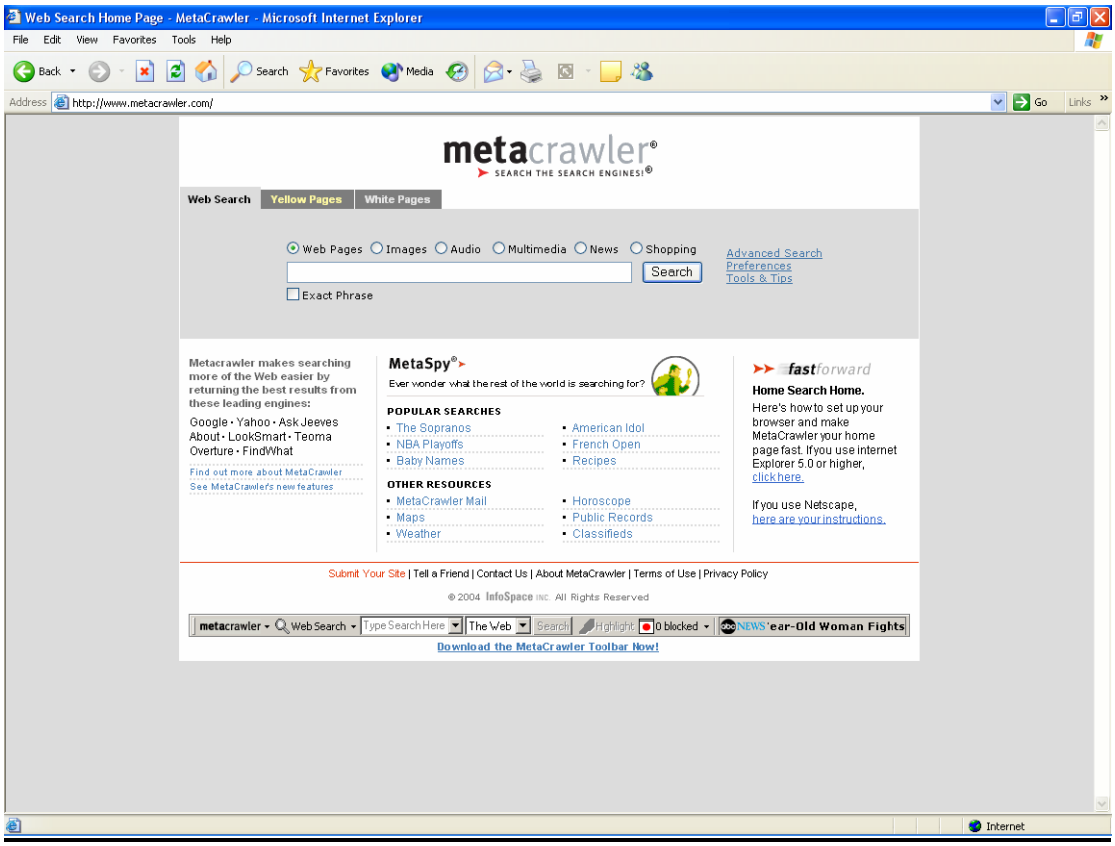
Το παραπάνω Σχήμα 5.1. επεξηγεί την εξιδανικευμένη αρχιτεκτονική των μηχανών πολλαπλής αναζήτησης.

Ένας χρήστης υποβάλλει το ερώτημά του μέσω του *interface* μιας μηχανής πολλαπλής αναζήτησης. Ο μηχανισμός αποστολών αποφασίζει σε ποιες μηχανές αναζήτησης θα στείλει την ερώτηση. Ταυτόχρονα, οι *interface πράκτορες* για τις παραπάνω επιλεγμένες μηχανές αναζήτησης υποβάλλουν την ερώτηση στις αντίστοιχες μηχανές τους. Όταν τα αποτελέσματα επιστρέφονται, οι αντίστοιχοι *interface πράκτορες* τα μετατρέπουν σε ένα ομοιόμορφο σχήμα. Ο μηχανισμός εμφάνισης ενσωματώνει τα αποτελέσματα από τους *interface πράκτορες*, αφαιρώντας τις διπλές εγγραφές και μορφοποιώντας τα για εμφάνιση φυλλομετρητή ***(web browser)*** του χρήστη.

5.2. MetaCrawler (www.metacrawler.com)

5.2.1. Εισαγωγή

Η μηχανή πολλαπλής αναζήτησης MetaCrawler αποτελεί μία από τις σημαντικότερες μηχανές του είδους, αφού καταφέρνει να συγκεντρώνει τις καλύτερες πληροφοριακές πηγές, ενώ ταυτόχρονα προσφέρει ένα ιδιαίτερα φιλικό και εύχρηστο interface αναζήτησης. Η MetaCrawler κατάφερε γρήγορα να καθιερωθεί μεταξύ των μηχανών συνδυαστικής αναζήτησης, αφού πέτυχε να συγκεντρώσει τις μεγαλύτερες και καλύτερες μηχανές και καταλόγους γενικής αναζήτησης, όπως τις AltaVista, WebCrawler, Yahoo, Excite, Lycos, LookSmart, ThunderStone, GoTo, DirectHit, About και Infoseek, καλύπτοντας έτσι ένα τεράστιο εύρος πληροφοριακών πηγών. Η μηχανή MetaCrawler έχει βραβευτεί για δύο συνεχόμενες χρονιές από το «Pc Magazine».



Σχήμα 5.2.: Η Αρχική Σελίδα της Μηχανής Πολλαπλής Αναζήτησης WebCrawler (www.webcrawler.com)

Όπως και σι περισσότερες μηχανές και μηχανές πολλαπλής αναζήτησης έτσι και

η MetaCrawler διαθέτει δύο βασικά interfaces αναζήτησης, ένα για την απλή και ένα για την προχωρημένη αναζήτηση. Στη σελίδα της βασικής αναζήτησης (Σχήμα 5.2.), εκτός του βασικού πεδίου αναζήτησης η μηχανή πολλαπλής αναζήτησης προσφέρει και έναν κατάλογο από θεματικές κατηγορίες, όπου καλύπτονται μια σειρά από γνωστικές περιοχές, σε μια προσπάθεια να βοηθηθεί και ο χρήστης που είναι λιγότερο εξοικειωμένος στην αναζήτηση με τη χρήση *λέξεων-κλειδί* ***(keywords)***.

Σε ό,τι αφορά την αναζήτηση, ο χρήστης μπορεί να χρησιμοποιήσει τους λογικούς τελεστές Boolean, με τη μορφή βέβαια που τους διαθέτει η συγκεκριμένη μηχανή αναζήτησης. Έτσι, αντί του τελεστή OR ο χρήστης θα πρέπει να επιλέξει το πεδίο Any, αντί του τελεστή AND την επιλογή ALL, ενώ ακόμη δίνεται η δυνατότητα να αναζητήσει τους όρους τους οποίους έχει εισαγάγει στο πεδίο αναζήτησης ως φράση επιλέγοντας το πεδίο Phrase. Ακόμη, ο χρήστης έχει τη δυνατότητα να επιλέξει και το χρονικό διάστημα στο οποίο θα επιθυμούσε να ανήκουν τα αποτελέσματα του. Επίσης, ως προς την ταξινόμηση των αποτελεσμάτων, μπορεί να επιλέξει εάν θέλει αυτή να γίνεται με βάση τη *σχετικότητα* ***(relevance)*** του εγγράφου που ανακτήθηκε, το site από το οποίο προέρχεται ή την πηγή που το δημιούργησε.

Η MetaCrawler αναλαμβάνει να απαλείψει τις *διπλές εγγραφές* ***(duplicates ή duplicate entries)*** από τις σελίδες των αποτελεσμάτων, ενώ συνοδεύει κάθε εγγραφή με παραπομπές σε σχετιζόμενα sites, προσφέροντας έτσι στο χρήστη μια πιο ολοκληρωμένη παρουσίαση του θέματος που αναζητά.

Το ερευνητικό πρόγραμμα WebCrawler το ξεκίνησε ο Brian Pinkerton στο τμήμα Computer Science and Engineering στο πανεπιστήμιο του Washington στο Seattle. Αργότερα έγινε κτήμα της America Online. Το Νοέμβριο του 1996, η Excite, απόκτησε τη WebCrawler.

Η WebCrawler έχει τα ακόλουθα χαρακτηριστικά (Bikkannavar, 1999):

- ✓ Χρησιμοποιεί ευρετήριο που βασίζεται στο *περιεχόμενο του κειμένου* ***(content-based indexing)*** καθώς και *ευρετήριο που βασίζεται σε όλο το κείμενο* ***(full-text indexing)*** για την παροχή υψηλής ποιότητας ευρετηρίου. Σε ένα *ρομπότ ιστού* ***(web robot)***, δεν υπάρχει κανένα πρόσθετο δίκτυο που εφαρμόζεται με ευρετηρίαση που βασίζεται σε όλο το κείμενο η εφαρμογή εμφανίζεται μόνο στον κεντρικό υπολογιστή (server).
- ✓ Χρησιμοποιεί μια *πρώτη σε εύρος στρατηγική αναζήτησης* ***(breadth-first search strategy)*** για τη δημιουργία ενός μεγάλου σε εύρος ευρετηρίου, που

μοιράζει το φορτίο μεταξύ των servers και εξασφαλίζοντας ότι κάθε server με χρήσιμο περιεχόμενο αντιπροσωπεύεται με διάφορες σελίδες στο ευρετήριο.

- ✓ Προσπαθεί να περιλάβει όσο το δυνατόν περισσότερους servers διαδικτύου. Το επιτυγχάνει με έναν φιλικό τρόπο, όπως η μη υπερφόρτωση των servers με αιτήματα άμεσης εξυπηρέτησης. Αυτό επίσης είναι σύμφωνο με το ***Robot Exclusion Standard***, το οποία είναι ένας τρόπος για τους Webmasters να επικοινωνούν με ρομπότ που οι τομείς (περιοχές) του ιστού είναι χωρίς όρια.

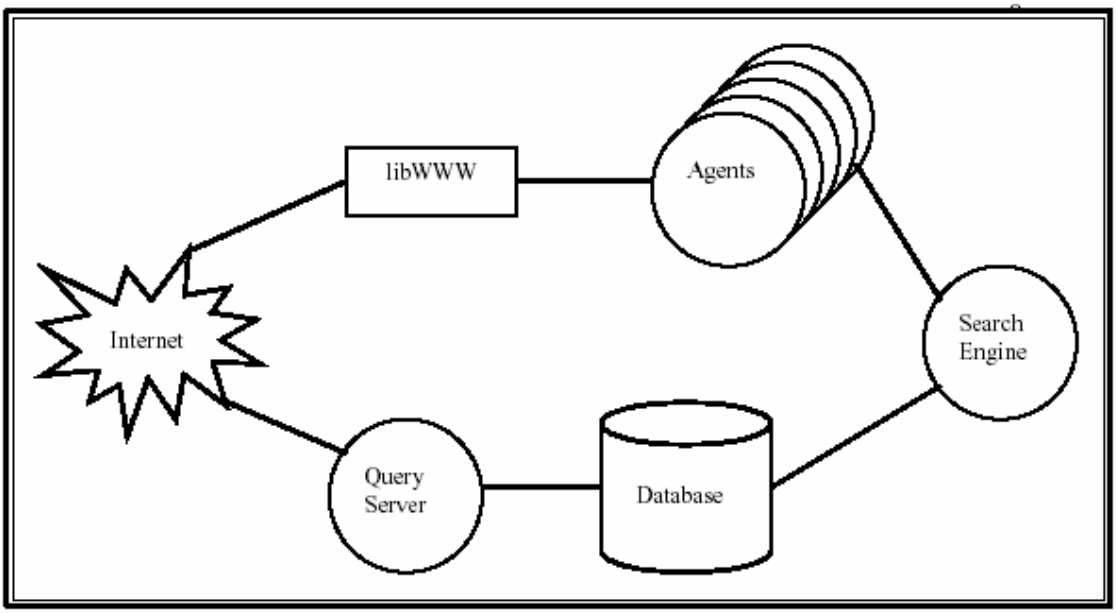
Η WebCrawler ξεκινά με ένα γνωστό σύνολο εγγράφων, εξετάζει τις συνδέσεις που φεύγουν από το έγγραφο, ακολουθεί μια από τις συνδέσεις που οδηγεί σε ένα νέο έγγραφο, και κατόπιν επαναλαμβάνει ολόκληρη τη διαδικασία. Με άλλα λόγια, η WebCrawler εξερευνά το διαδίκτυο ως μεγάλο *κατευθυνόμενο γράφο* ***(directed graph)*** χρησιμοποιώντας έναν graph traversal αλγόριθμο που εκτελεί την ακόλουθη σειρά ενεργειών επανειλημμένως:

1. Ανακαλύπτει ένα νέο έγγραφο
2. Σημαδεύει το έγγραφο ότι έχει ανακτηθεί
3. Αποκρυπτογραφεί οποιεσδήποτε εξερχόμενες συνδέσεις
4. Ευρετηριάζει το περιεχόμενο του εγγράφου.

5.2.2. Αρχιτεκτονική της WebCrawler

Η αρχιτεκτονική της WebCrawler, όπως φαίνεται και από το Σχήμα 5.3. αποτελείται από τα παρακάτω τέσσερα συστατικά:

1. **Μηχανή Αναζήτησης:** Κατευθύνει τις δραστηριότητες της WebCrawler και είναι υπεύθυνη για την απόφαση ποια νέα έγγραφα θα ευρετηριαστούν και για την εκκίνηση της ανάκτησής τους.
2. **Βάση δεδομένων**: Αυτό χειρίζεται την επίμονη αποθήκευση των metadata, τις συνδέσεις μεταξύ των εγγράφων, και το ευρετήριο με βάση το πλήρες κείμενο της ιστοσελίδας.
3. **Πράκτορες**: Αυτοί είναι υπεύθυνοι για την ανάκτηση των εγγράφων από το δίκτυο στη μηχανή αναζήτησης.
4. **Κεντρικός Υπολογιστής Ερωτημάτων**: Αυτό εφαρμόζει την υπηρεσία απάντησης της ερώτησης που παρέχεται από τους χρήστες διαδικτύου.



Σχήμα 5.3.: Αρχιτεκτονική της Μηχανής Πολλαπλής Αναζήτησης (Bikkannavar, 1999)

5.2.3. Η Μηχανή Αναζήτησης της WebCrawler

Η μηχανή αναζήτησης της WebCrawler καθορίζει ποια έγγραφα και ποιοι τύποι εγγράφων θα επισκεφθούν. Μη-ευρετηριάσιμα αρχεία, όπως εικόνες, ήχοι, PostScript, ή δυαδικά στοιχεία, δεν ανακτώνται. Επιπλέον, τα λάθος ανακτημένα αρχεία αγνοούνται κατά τη διάρκεια της σύνταξης ευρετηρίου. Αυτό το είδος της διαχώρισης των τύπων αρχείου εφαρμόζεται και **στην ευρετηρίαση** και **στις σε πραγματικό χρόνο** τρόπους αναζήτησης. Η μηχανή αναζήτησης χρησιμοποιεί διαφορετικές στρατηγικές αναζήτησης κατά όταν χρησιμοποιείται η WebCrawler σε αυτούς τους δύο τρόπους αναζήτησης.

✓ ***Τρόπος ευρετηρίασης***

Με αυτόν τον τρόπο ο στόχος είναι να δημιουργηθεί ένα ευρετήριο με όσο το δυνατόν να καλύπτει το μεγαλύτερο μέρος του ιστού μέσα σε περιορισμένο χώρο αποθήκευσης. Η WebCrawler θεωρεί ότι τα έγγραφα ιστού που χρησιμοποιούνται για τη δημιουργία ευρετηρίου πρέπει να προέρχονται από όσο το δυνατόν περισσότερους διαφορετικούς servers. Χρησιμοποιεί έναν **τροποποιημένο αλγόριθμος breadth-first** για να εξασφαλίσει ότι κάθε server έχει τουλάχιστον ένα έγγραφο που να τον αντιπροσωπεύει στο ευρετήριο. Αυτά τα βήματα δείχνουν πώς ο αλγόριθμος λειτουργεί:

1. Όταν ένα έγγραφο από έναν νέο server εντοπίζεται, αυτός ο server τοποθετείται σε έναν κατάλογο από servers για να επισκεφτεί άμεσα.
2. Ένα έγγραφο από κάθε έναν από τους νέους κεντρικούς υπολογιστές ανακτάται και ευρετηριάζεται πριν επισκεφθούν οποιαδήποτε άλλα έγγραφα.
3. Όταν όλοι οι γνωστοί servers έχουν επισκεφθεί, η διαδικασία σύνταξης ευρετηρίου προχωρά διαδοχικά μέσω του καταλόγου όλων των servers μέχρι να βρεθεί ένας νέος, σε αυτό το σημείο η διαδικασία επαναλαμβάνεται.

✓ ***Τρόπος Αναζήτησης σε Πραγματικό Χρόνο***

Με αυτόν τον τρόπο, ο στόχος είναι να βρεθούν τα έγγραφα τα οποία είναι παρόμοια με την ερώτηση του χρήστη. Η WebCrawler χρησιμοποιεί έναν διαφορετικό αλγόριθμο αναζήτησης. Η διαίσθηση για αυτόν τον αλγόριθμο είναι ότι ακολουθώντας τις συνδέσεις των εγγράφων που είναι παρόμοια με αυτό που ο χρήστης θέλει είναι πιθανότερο να οδηγηθεί σε σχετικά έγγραφα απ' ό,τι ακολουθώντας οποιαδήποτε σύνδεση εγγράφου. Ο αλγόριθμος εκτελείται όπως παρακάτω:

1. Η WebCrawler μεταβιβάζει στο ευρετήριό της το ερώτημα του χρήστη με σκοπό να εντοπίσει έναν αρχικό κατάλογο παρόμοιων εγγράφων.
2. Από τον παραπάνω κατάλογο, τα πιο σχετικά έγγραφα σημειώνονται, και οποιεσδήποτε ανεξερεύνητες συνδέσεις αυτών των εγγράφων ακολουθούνται.
3. Όταν τα νέα έγγραφα ανακτώνται, προστίθενται στο ευρετήριο, και η ερώτηση του χρήστη επαναλαμβάνεται.
4. Τα αποτελέσματα της ερώτησης ταξινομούνται με βάση τη σχετικότητά τους, και τα νέα έγγραφα που βρίσκονται κοντά στην κορυφή είναι υποψήφια για περαιτέρω εξερεύνηση.
5. Η διαδικασία επαναλαμβάνεται είτε μέχρι η WebCrawler να έχει βρει παρόμοια έγγραφα που να ικανοποιήσουν το χρήστη είτε μέχρι ένα χρονικό όριο να εξαντληθεί.

5.2.4. WebCrawlers Πράκτορες

Οι μηχανές αναζήτησης επικαλούνται τους πράκτορες με σκοπό την ανάκτηση των εγγράφων ιστού. Επειδή η αναμονή των servers και το δίκτυο δημιουργεί μια καθυστέρηση στην αναζήτηση, οι πράκτορες τρέχουν σε ξεχωριστές διαδικασίες, και η WebCrawler χρησιμοποιεί παράλληλα μέχρι 15 πράκτορες. Για κάθε νέο έγγραφο του

ιστού που πρέπει να ανακτηθεί, η μηχανή αναζήτησης βρίσκει έναν ελεύθερο πράκτορα, και του ζητά ν' ανακτήσει τη URL που αντιπροσωπεύει το έγγραφο. Ο πράκτορας απαντά στη μηχανή αναζήτησης είτε με ένα αντικείμενο που περιέχει το περιεχόμενο του εγγράφου είτε με μια εξήγηση, γιατί το έγγραφο δεν μπορεί να ανακτηθεί. Αφού έχει απαντήσει ο πράκτορας, γίνεται ελεύθερος πάλι και μπορεί να του δοθεί νέα εργασία να εκτελέσει.

Το πρόγραμμα των πρακτόρων χρησιμοποιεί τη βιβλιοθήκη του CERN WWW Library (libWWW), η οποία υποστηρίζει την πρόσβαση σε διάφορους τύπους περιεχομένων μέσω διαφορετικών πρωτοκόλλων, συμπεριλαμβανομένου του HTTP, FTP, και Gopher. Σαν πρακτικό θέμα, οι τρέχοντες πράκτορες στις χωριστές διαδικασίες βοηθούν στην απομόνωση της κύριας WebCrawler διαδικασίας από διαρροές μνήμης και από λάθη στον πράκτορα και στη libWWW.

5.2.5. Η Βάση Δεδομένων της WebCrawler

Η βάση δεδομένων της WebCrawler κρατά και το *ευρετήριο με βάση όλο το κείμενο* ***(full-text index)*** και την αναπαράσταση του ιστού με *γράφο* ***(Graph)***. Η βάση δεδομένων είναι αποθηκευμένη στο δίσκο και ενημερώνεται καθώς έγγραφα προστίθενται σε αυτή. Για να προστατευθεί η βάση δεδομένων από αστοχίες των συστημάτων, οι ενημερώσεις των περιεχομένων της γίνονται κάτω από το πεδίο των συναλλαγών που είναι δεσμευμένο κάθε λίγες εκατοντάδες έγγραφα.

Η WebCrawler χρησιμοποιεί το NeXTStep IndexingKit για τη δημιουργία του *ευρετηρίου με βάση όλο το κείμενο*, το οποίο καθιστά τις ερωτήσεις γρήγορες: αναζητώντας μια λέξη παράγει έναν *κατάλογο δεικτών* ***(list of pointers)*** στα έγγραφα που περιέχουν αυτή τη λέξη. Οι πιο περίπλοκες ερωτήσεις απαντώνται με συνδυασμό των καταλόγων των εγγράφων για αρκετές λέξεις με συμβατά σύνολα λειτουργιών. Το ευρετήριο χρησιμοποιεί ένα *Vector-Space πρότυπο* για την απάντηση των ερωτήσεων.

Οι λέξεις ενός εγγράφου ελέγχονται μέσω ενός *καταλόγου από stop-words* για να αποτραπεί η ευρετηρίαση των κοινών λέξεων και το βάρος τους ισούται με τη διαίρεση της συχνότητας εμφάνισής τους στο έγγραφο προς τη συχνότητα ενός πεδίου αναφοράς. Λέξεις που εμφανίζονται συχνά στο έγγραφο και σπάνια στο πεδίο αναφοράς έχουν πιο υψηλή σημασία, ενώ οι λέξεις που εμφανίζονται σπάνια και στα δύο τους δίνεται χαμηλότερα βάρη. Αυτός ο τύπος στάθμισης καλείται συνήθως ***peculiarity weighting***.

Το υπόλοιπο της βάσης δεδομένων αποθηκεύει στοιχεία για τους servers, τα

έγγραφα και τις συνδέσεις. Ολόκληρα URLs δεν αποθηκεύονται αλλά χωρίζονται σε αντικείμενα που περιγράφουν τον server και το έγγραφο. Μια αναφορά σε ένα έγγραφο είναι απλά ένας δείκτης σε ένα άλλο έγγραφο. Κάθε αντικείμενο αποθηκεύεται σε ένα χωριστό Btree στο δίσκο: τα έγγραφα στο ένα, οι servers στο άλλο, και οι συνδέσεις στο τελευταίο. Ο χωρισμός των στοιχείων με αυτόν τον τρόπο επιτρέπει στη WebCrawler να ανιχνεύει τον κατάλογο των servers γρήγορα να επιλέγει τους ανεξερεύνητους servers ή τον λιγότερο πρόσφατα χρησιμοποιούμενο server.

5.2.6. Ο Query Server της WebCrawler

Ο server των ερωτήσεων πραγματοποιεί την υπηρεσία αναζήτησης της WebCrawler. Το πρότυπο της ερώτησης που παρουσιάζει είναι ένα απλό ***vector-space query model*** βασισμένο στη ***full-text*** βάση δεδομένων. Οι χρήστες εισάγουν λέξεις κλειδιά ως ερώτηση, και οι τίτλοι, οι ***URLs*** των εγγράφων που περιέχουν μερικές ή όλες αυτές τις λέξεις ανακτώνται από το ευρετήριο και παρουσιάζονται στο χρήστη ως διαταγμένος κατάλογος ταξινομημένος με βάση τη σχετικότητα της ερώτησης. Σε αυτό το πρότυπο, η σχετικότητα είναι το άθροισμα (σε όλες τις λέξεις της ερώτησης) του βάρους της λέξης στο έγγραφο και του βάρους της στην ερώτηση διαιρεμένη με τον αριθμό των λέξεων στην ερώτηση.

5.2.7. Χαρακτηριστικά Αναζήτησης της WebCrawler

1. **Η WebCrawler καθιστά την αναζήτηση εύκολη**: Η WebCrawler υιοθετεί την προχωρημένη αναζήτηση και την τεχνολογία ανάκτησης από το *Personal Library Software* ***(PLS)*** έτσι ώστε ακόμη και οι αρχάριοι χρήστες θα πάρουν τα αποτελέσματα που θέλουν. Η WebCrawler είναι προγραμματισμένη "να κάνει το σωστό πράγμα" ακόμα και όταν οι αναζητήσεις περιγράφονται όχι και με τον ιδανικότερο τρόπο.

2. **Η WebCrawler υποστηρίζει *τη φυσική γλώσσα αναζήτησης***: Η WebCrawler υποστηρίζει τη *φυσική γλώσσα αναζήτησης* ***(natural language searching)*** έτσι ώστε οι χρήστες να μπορούν να διατυπώσουν τις ερωτήσεις τους με σαφή αγγλικά χωρίς να χρειάζεται να μάθουν κάποια σύνθετη σύνταξη αναζήτησης. Οι προηγμένοι χρήστες είναι γνώστες του ότι η WebCrawler υποστηρίζει ένα ευρύ φάσμα Boolean τελεστών αναζήτησης.

3. **Η WebCrawler αντιστοιχεί έναν ή όλους τους όρους αναζήτησης**: Όταν μια σειρά από όρους αναζήτησης δακτυλογραφούνται, η WebCrawler είναι προγραμματισμένη να βρίσκει αποτελέσματα που ταιριάζουν με κάθε έναν ή με όλες εκείνες τις λέξεις.

5.3. Ixquick (www.ixquick.com)

5.3.1. Εισαγωγή

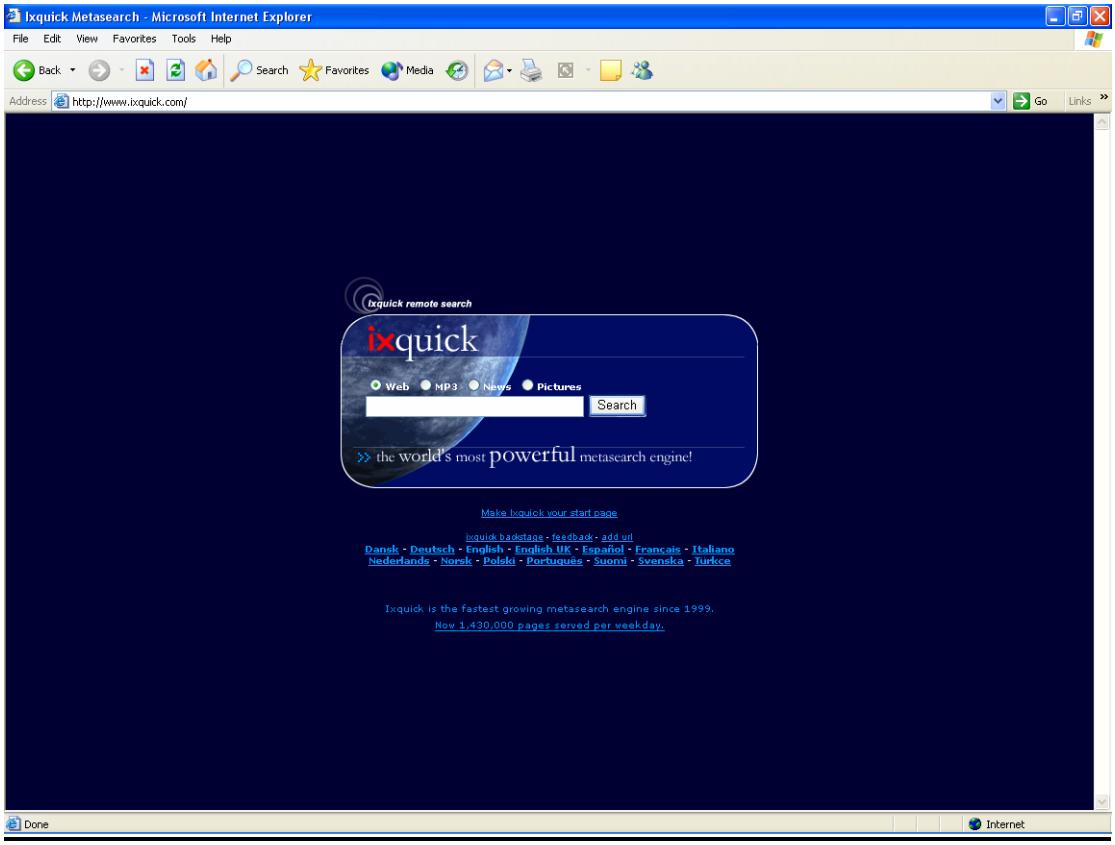
Σε αυτή την ενότητα παρουσιάζεται η μηχανή πολλαπλής αναζήτησης Ixquick. Αρχικά γίνεται αναφορά για το interface της Ixquick. Στη συνέχεια για την πρόσβασή της, για τις μηχανές αναζήτησης που χρησιμοποιεί, για τα αποτελέσματα, για το σύστημα ταξινόμησης, για την εμφάνιση των αποτελεσμάτων, και για τη δυνατότητα της ixquick να χειριστεί ειδικούς τελεστές και τη φυσική γλώσσα καθώς και ερωτήσεις που να περιέχουν Boolean τελεστές.

Σε μια πρόσφατη συνέντευξη, ο Larry Page, ένας από τους δημιουργούς της Google, ίσως της δημοφιλέστερης μηχανής αναζήτησης, είπε ότι η αρχική σελίδα της Google δεν είναι πολύ φορτωμένη με συνδέσεις, επειδή η έρευνα δείχνει ότι ο χρόνος που χρειάζεται για να ολοκληρωθεί μια εργασία είναι ανάλογος με τον αριθμό των διαθέσιμων επιλογών (η υπόθεση είναι ότι είναι περισσότερο χρονοβόρα μια αναζήτηση σε μια αρχική σελίδα με πολλές συνδέσεις από ό,τι σε μια αρχική σελίδα με λίγες). Φαίνεται ότι ο David Bodnick, ο σχεδιαστής της Ixquick και κύριος ιδιοκτήτης της Ixquick.com, συμφωνεί με τον Page.

Στην αρχική σελίδα της ixquick (Σχήμα 5.4.) δεν υπάρχουν αγγελίες, κινούμενα αρχεία GIF, ή ***applets της Java*** που να αποσπούν τους χρήστες. Στην πραγματικότητα, το σκούρο μπλε υπόβαθρο της αρχικής σελίδας καλύπτεται κατά ένα μεγάλο μέρος από το ίδιο το *κουτί αναζήτησης* ***(search box)***, το οποίο συνοδεύεται από τέσσερις επιλογές, τον *ιστό* ***(web)***, *τις ειδήσεις* ***(news)***, τα ***MP3***, και τις *εικόνες* ***(pictures)***. Οι σύνδεσμοι υπερκειμένου σε σελίδες με πληροφορίες για την Ixquick, χαρακτηριστικά αναζήτησης, και σελίδες της ixquick για άλλες γλώσσες βρίσκονται κάτω από το κουτί αναζήτησης, αλλά πολύ λίγο αποσπούν την προσοχή (Σχήμα 5.4.).

Η Ixquick έχει υιοθετήσει αυτήν την μινιμαλιστική μέθοδο. Οι περισσότερες μηχανές πολλαπλής αναζήτησης καθώς και οι μηχανές αναζήτησης παρέχουν ένα *link* στην αρχική σελίδα που οδηγεί στην προχωρημένη αναζήτηση (για ανθρώπους που τους αρέσουν οι πολλές επιλογές), επιτρέποντας στους χρήστες να καθορίσουν τις περιοχές αναζήτησης, τον αριθμό των εγγράφων που ανακτώνται, ή τον αριθμό των σελίδων που απαριθμούν από κάθε περιοχή. Η Ixquick δεν δίνει αυτή την επιλογή. Οι χρήστες πρέπει πρώτα να κάνουν μια αναζήτηση και έπειτα όταν εμφανίζεται η οθόνη των αποτελεσμάτων μπορούν οι χρήστες να τροποποιήσουν την αναζήτηση επιλέγοντας τις μηχανές αναζήτησης που θέλουν να χρησιμοποιήσουν. Επιπλέον, δεν υπάρχει καμία

σύνδεση με τη σελίδα των παραμέτρων αναζήτησης από την οθόνη εμφάνισης των ανακτημένων εγγράφων. Εάν οι χρήστες θέλουν να ξέρουν για τις αναζητήσεις με βάση τους τελεστές Boolean πρέπει να επιστρέψουν στην αρχική σελίδα και να χρησιμοποιήσουν τη σύνδεση που βρίσκεται εκεί.



Σχήμα 5.4.: Η Αρχική Σελίδα της Μηχανής Πολλαπλής Αναζήτησης Ixquick (www.ixquick.com)

5.3.2. Πρόσβαση στην Ixquick

Η δυνατότητα των χρηστών να χρησιμοποιούν συνεχώς μια μηχανή αναζήτησης για την ανάκτηση των αποτελεσμάτων, ανεξάρτητα από τη σχετικότητά τους, αναφέρεται γενικά ως *πρόσβαση* ***(Access)***. Η πρόσβαση καθορίζεται συνήθως από την τεχνολογία που υποστηρίζει τη μηχανή αναζήτησης. Παραδείγματος χάριν, μερικές φορές οι μηχανές αναζήτησης δεν μπορούν να ανακτήσουν τα αποτελέσματα λόγω της κυκλοφοριακής συμφόρησης των δικτύων ή των τυχόν αποτυχιών των *servers*. Η πρόσβαση είναι μια παράμετρος που δεν λαμβάνεται υπόψη στις αξιολογήσεις των μηχανών αναζήτησης. Ωστόσο, η δυνατότητα να εκτελεσθεί και να ολοκληρωθεί μια αναζήτηση είναι σημαντική. Εάν τα αποτελέσματα δεν μπορούν να ανακτηθούν τότε

δεν μπορούν να αξιολογηθούν.

Η Ixquick έχει συνεπή πρόσβαση, που σημαίνει ότι έχει μια ισχυρή τεχνολογία που υποστηρίζει τη μηχανή αναζήτησης. Κατά τη διάρκεια μιας αναζήτησης, εάν η ixquick αποτύχει να ολοκληρώσει την αναζήτηση λόγω περιορισμών των συστημάτων, εμφανίζει το ακόλουθο μήνυμα:

"Ixquick down. We are extremely sorry, but ixquick's servers are temporarily overloaded. Please search again in minute or two"

5.3.3. Οι Μηχανές Αναζήτησης της Ixquick

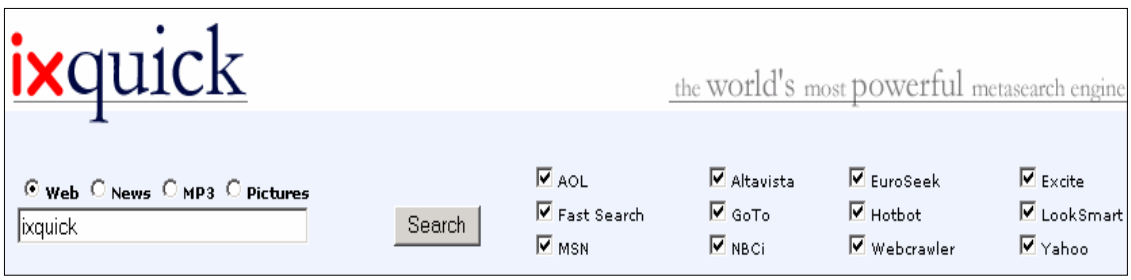
Οι δέκα σημαντικότερες μηχανές πολλαπλής αναζήτησης (συμπεριλαμβανομένου της Ixquick) στηρίζονται γενικά σε 12,2 μηχανές αναζήτησης για να διεξαγάγουν μια αναζήτηση. Οι Τέσσερις χρησιμοποιούν 13 μηχανές αναζήτησης ή περισσότερες. Η Search.com χρησιμοποιεί τις περισσότερες με 19. Η Vivisimo και η Mamma χρησιμοποιούν τις λιγότερες με οκτώ.

Η Ixquick το 2000 χρησιμοποιούσε 14 μηχανές αναζήτησης κάνοντάς την μια από τις περιεκτικότερες μηχανές πολλαπλής αναζήτησης. Από το 2001 μέχρι και σήμερα χρησιμοποιεί 12 για τις αναζητήσεις (Σχήμα 5.5.). Οι 12 μηχανές αναζήτησης είναι:

Οι Μηχανές Αναζήτησης της Ixquick			
AOL	Altavista	Euroseek	Excite
FastSearch	GoTo	Hotbot	LookSmart
MSN	NBCi	WebCrawler	Yahoo

Πίνακας 5.1.: Οι Μηχανές Αναζήτησης της Ixquick

Από τις παραπάνω μηχανές, οι Direct Hit, InfoSeek, LiveDirectory, Lycos, και Open Directory καταργήθηκαν. Οι Euroseek, Hotbot, και LookSmart προστέθηκαν.



Σχήμα 5.5.: Οι Μηχανές Αναζήτησης της Μηχανής Πολλαπλής Αναζήτησης Ixquick

Η Ixquick προτιμά τις μηχανές αναζήτησης που είναι δημοφιλείς και εκείνες με τη μεγάλη κυκλοφορία. Είναι φανερό ότι στην αναζήτηση της Ixquick απουσιάζουν οι ισχυρές μηχανές όπως η Google, η Lycos, και η Northern Light. Λαμβάνοντας υπόψη το μέγεθος των ευρετηρίων και τη φήμη αυτών των μηχανών αναζήτησης, αποκλείοντάς τις από τη μηχανή πολλαπλής αναζήτησης είναι μια μεγάλη απώλεια. Ακόμη και με τις μηχανές πολλαπλής αναζήτησης, η ανεύρεση πληροφοριών στον ιστό είναι δύσκολη διαδικασία, και αποκλείοντας μερικές από τις καλύτερες μηχανές αναζήτησης είναι μια ενέργεια που κάνει μόνο αυτήν την διαδικασία πιο προκλητική.

Όπως αρκετές μηχανές πολλαπλής αναζήτησης, έτσι και η Ixquick επιτρέπει στους χρήστες να ψάξουν για πληροφορίες στις *εξειδικευμένες κατηγορίες* ***(specialized categories)*** χρησιμοποιώντας τις *ειδικές μηχανές αναζήτησης* ***(specialty search engines)***. Οι χρήστες μπορούν να ψάξουν σε επτά πηγές για ειδήσεις: τις Associated Press, CNET, CNN, Los Angeles Times, Reuters, San Francisco Chronicle, και Washington Post. Υπάρχουν τέσσερις πηγές για MP3 αρχεία: AstraWeb MP3, Lycos MP3, MP3.com και Oth.net MP3. Υπάρχουν επίσης οκτώ πηγές για εικόνες: Οι Altavista, Art.com, Fast Search, Go, Lycos, Pictures Now, San Francisco Art Museum, και Yahoo.

5.3.4. Το Σύστημα Ταξινόμησης της Ixquick (Ranking)

Η Ixquick χρησιμοποιεί τον συμβολισμό με τα αστέρια (*) για να ταξινομήσει τα έγγραφα (Σχήμα 5.6.). Αρχικά, η Ixquick εξετάζει τις πρώτες δέκα ιστοσελίδες από κάθε μηχανή αναζήτησης στις οποίες έχει τεθεί το ερώτημα, και έπειτα κάθε ιστοσελίδα παίρνει ένα αστέρι για την εμφάνισή της στις πρώτες δέκα ιστοσελίδες κάθε μηχανής αναζήτησης. Τα αστέρια δείχνουν τη σχετικότητα της σελίδας, και έχοντας ένα αστέρι σημαίνει ότι μια τουλάχιστον μηχανή αναζήτησης θεώρησε αυτή την ιστοσελίδα σχετική. Ωστόσο, μια σελίδα κερδίζει περισσότερα αστέρια για την ταξινόμησή της στις πρώτες δέκα σελίδες του καταλόγου των άλλων μηχανών αναζήτησης. Κατά συνέπεια τα περισσότερα αστέρια σε μια σελίδα, την κατατάσσουν ως περισσότερο σχετική στην ερώτηση του χρήστη. Οι ιστοσελίδες παρουσιάζονται στο χρήστη από εκείνες με τα περισσότερα αστέρια ως εκείνες με τα λιγότερα.

Αυτό το σύστημα ταξινόμησης λειτουργεί καλά εφ' όσον οι χρήστες είναι πρόθυμοι να εξετάσουν όλες τις ανακτημένες σελίδες της Ixquick. Είναι δυνατό μόνο μια μηχανή αναζήτησης να ανακτήσει τη σελίδα που ένας χρήστης θα θεωρήσει ως πιο

σχετική και κατά συνέπεια θα ανακτήσει μόνο ένα αστέρι. Οι χρήστες πρέπει λοιπόν να ταξινομήσουν τα έγγραφα με τα περισσότερα αστέρια πριν την βρουν. Δεδομένου ότι Ixquick συνήθως παρουσιάζει ένα σχετικά μικρό σύνολο εγγράφων, μεταξύ 30-60 ιστοσελίδες, αυτό δεν πρέπει να θεωρηθεί πρόβλημα, αλλά οι χρήστες πρέπει να ξέρουν ότι πρέπει να εξετάσουν όλες τις σελίδες. Επιπλέον, οι χρήστες πρέπει να γνωρίζουν ότι κάθε σελίδα με τον ίδιο αριθμό αστεριών θεωρείται ίδιας σχετικότητας. Έτσι μια σελίδα με δύο αστέρια εμφανίζεται πριν από μια άλλη σελίδα με τα ίδια αστέρια αυτό δε σημαίνει ότι είναι πιο σχετική.

Σχήμα 5.6.: Το σύστημα Ταξινόμησης της Ixquick (Hawkins, 2001)

5.3.5. Παρουσίαση Αποτελεσμάτων

Σαν γενικό κανόνα, η Ixquick ανακτά οπωσδήποτε 30-60 έγγραφα για οποιαδήποτε δεδομένη ερώτηση. Αυτό συμβαίνει επειδή η Ixquick αξιολογεί μόνο τα κορυφαία δέκα έγγραφα από κάθε μηχανή αναζήτησης που χρησιμοποιεί για τη συγκεκριμένη αναζήτηση. Η Ixquick αφαιρεί τις διπλές εγγραφές, αλλά εξετάζει μόνο τη URL. Μερικές φορές οι ιστοσελίδας έχουν διαφορετικές URLs αλλά ίδιο περιεχόμενο, τότε η Ixquick θα λάβει σελίδες με τις ίδιες πληροφορίες.

Όπως όλες οι μηχανές αναζήτησης, η Ixquick ανακτά τις περιγραφές των εγγράφων, και όχι τις πραγματικές ιστοσελίδες. Κατά την ανάκτηση της ιστοσελίδας ή των ειδησεογραφικών άρθρων, η Ixquick επιστρέφει τις περιγραφές των εγγράφων που αποτελούνται από το κείμενο του ιστού που περιέχει τους όρους της ερώτησης, τη *URL* του εγγράφου, μια σύνδεση με την πραγματική ιστοσελίδα, μια σύνδεση με τη μηχανή αναζήτησης ή τις μηχανές αναζήτησης που βρήκαν το έγγραφο και την ταξινόμηση της σελίδας από εκείνη ή εκείνες τις μηχανές αναζήτησης, και τα αστέρια που υποδεικνύουν τη σχετικότητά τους. Κατά την ανάκτηση των MP3 αρχείων, η περιγραφή των εγγράφων αποτελείται από μια σύνδεση με το MP3 αρχείο, τη *URL*, μια σύνδεση με τη μηχανή αναζήτησης ή τις μηχανές αναζήτησης που βρήκαν το MP3

αρχείο, και τα αστέρια. Ωστόσο, κατά την ανάκτηση των εικόνων, η Ixquick περιλαμβάνει εκτός από τη σύνδεση με την εικόνα, το όνομα που δίνεται στην εικόνα, μια σύνδεση με τη μηχανή αναζήτησης ή τις μηχανές αναζήτησης που βρήκαν την εικόνα, και τα αστέρια.

Ένα άλλο στοιχείο εμφάνισης είναι η μέθοδος σύνδεσης της Ixquick με τις ιστοσελίδες. Αντίθετα από τις περισσότερες μηχανές αναζήτησης, οι οποίες ανοίγουν τις ιστοσελίδες στο ίδιο παράθυρο, η Ixquick ανοίγει ένα νέο παράθυρο για τις ιστοσελίδες. Έτσι με αυτό τον τρόπο, οι χρήστες μπορούν να παρακολουθούν τα έγγραφα που έχουν ψάξει χωρίς να πρέπει να χρησιμοποιηθούν οι λειτουργίες πλοήγησης του browser. Το μειονέκτημα αυτής της προσέγγισης είναι ότι, εάν οι χρήστες έχουν παλαιούς υπολογιστές ή πιο αργά modems, οι διαδικασίες των υπολογιστών μπορούν να παγώσουν ως συνέπεια των πολλών ανοικτών παραθύρων.

Τέλος, η Ixquick έχει ένα νέο χαρακτηριστικό εμφάνισης, *το τονισμένο αποτέλεσμα* ***(Highlighted Result)***. Με την επιλογή αυτού του χαρακτηριστικού, η Ixquick ανοίγει την ιστοσελίδα σε ένα ***split-framed window***. Στην κορυφή, το πλαίσιο ελέγχου της Ixquick έχει ένα παράθυρο αναζήτησης με τους όρους της ερώτησης, ένα ***pull-down menu*** για την επιλογή άλλων ιστοσελίδων που ανακτήθηκαν από την αναζήτηση. Κάτω από το πλαίσιο ελέγχου της Ixquick είναι η ιστοσελίδα. Κάθε εμφάνιση ενός όρου της ερώτησης τονίζεται και οι διαφορετικοί όροι της ερώτησης τονίζονται με διαφορετικά χρώματα.

5.3.6. Ειδικές Επιλογές/ Χαρακτηριστικά

- ***Special Operators***

Ο Πίνακας 5.2. δείχνει ότι η Ixquick κάνει σωστή χρήση των ειδικών τελεστών (+) και (-). Τα εισαγωγικά δημιουργούν μερικά προβλήματα

Ερώτηση	Απάντηση
Wom*n	Τα ανακτημένα έγγραφα περιείχαν τους όρους woman, women, και wom@n
sail –boat	Από 43 έγγραφα, κανένα δεν περιείχε τη λέξη boat
River otter -"sea otter"	32 έγγραφα ανακτήθηκαν, κανένα δεν περιείχε τη φράση “sea otter”
+ “diplomatic row” +China	30 έγγραφα ανακτήθηκαν, τα 10 έγγραφα περιείχαν τη φράση “diplomatic row” και China, αλλά κάποια χαμηλότερα ταξινομημένα έγγραφα δεν περιείχαν τη φράση στο κείμενο

Πίνακας 5.2.: Οι Ειδικοί Τελεστές της Ixquick (Hawkins, 2001)

- ***Αναζήτηση με Χρήση της Φυσικής Γλώσσας***

Η Ixquick ισχυρίζεται ότι υποστηρίζει τις ερωτήσεις φυσικής γλώσσας., κάτι το οποίο ισχύει μερικώς. Η πρώτη ένδειξη ότι δεν υποστηρίζει πλήρως την επεξεργασία φυσικής γλώσσας είναι το μικρό κουτί της αναζήτησης. Τα περισσότερα συστήματα ανάκτησης πληροφοριών με ικανότητες φυσικής γλώσσας όπως το Lexis και το Proquest έχουν πολύ μεγαλύτερα κουτιά αναζήτησης για να προσαρμόσουν τις πιο μεγάλες σε μήκος ερωτήσεις που χρησιμοποιούνται στην επεξεργασία της φυσικής γλώσσας. Η δεύτερη ένδειξη ότι η Ixquick δεν υποστηρίζει πλήρως τις ερωτήσεις διατυπωμένες σε φυσική γλώσσα, είναι ότι καμία από τις μηχανές αναζήτησης που υποστηρίζει η Ixquick δεν βεβαιώνει τη χρήση ερωτήσεων φυσικής γλώσσας. Πώς μπορεί έτσι η Ixquick να υποστηρίζει τις ερωτήσεις φυσικής γλώσσας εάν οι μηχανές αναζήτησης που χρησιμοποιεί δεν τις υποστηρίζουν;

Αυτό που η Ixquick πραγματοποιεί είναι να μεταχειρίζεται τις ερωτήσεις διατυπωμένες σε φυσική γλώσσα όπως τις Boolean ερωτήσεις, ενώνοντας κάθε όρο με ένα ***Boolean AND***. Ενώ μερικές μηχανές αναζήτησης προσπαθούν να αναγνωρίσουν το περιεχόμενο ή τις φράσεις μέσα σε μια ερώτηση, πολύ συχνά οι ιστοσελίδες ανακτώνται επειδή περιέχουν όλες ή τους περισσότερους όρους της ερώτησης.

- ***Boolean τελεστές***

Γενικά η Ixquick υποστηρίζει ότι μπορεί και διαχειρίζεται ερωτήματα με Boolean τελεστές. Οι διάφορες μελέτες και έρευνες που έχουν γίνει απέδειξαν ότι οι χρήστες που δεν εμπιστεύονται απολύτως την Ixquick για αυτού του είδους των ερωτήσεων, έχουν δίκιο. Η Ixquick στην πραγματικότητα δεν διαχειρίζεται τις Boolean ερωτήσεις εξίσου καλά με τις άλλες μηχανές αναζήτησης.

5.4. Neci Inquirus

5.4.1. Εισαγωγή

Αρχικά τα κίνητρα που βρίσκονται πίσω από τη μηχανή πολλαπλής αναζήτησης Inquirus ήταν η μικρή ακρίβεια, η περιορισμένη κάλυψη, η περιορισμένη διαθεσιμότητα, τα περιορισμένα *user interfaces*, και οι μη ενημερωμένες βάσεις δεδομένων των σημαντικότερων μηχανών αναζήτησης του ιστού. Παρακάτω γίνεται αναφορά στα σημεία αυτά (Lawrence and Giles, 1998):

- *Μικρή ακρίβεια* ***(Poor Precision)****:* Η διαφορετική φύση του ιστού, και η εστίαση των μηχανών αναζήτησης να χειρίζονται τις σχετικά απλές ερωτήσεις πολύ γρήγορα, οδηγεί σε αποτελέσματα που συχνά έχουν μικρή ακρίβεια. Επιπλέον, η τεχνική του *"spamming"* έχει γίνει δημοφιλής, με την οποία οι χρήστες προσθέτουν ενδεχομένως μη σχετικές λέξεις κλειδιά στις σελίδες τους προκειμένου να αλλάξουν την ταξινόμηση των σελίδων τους. Η εμπειρία δείχνει ότι η σχετικότητα μιας συγκεκριμένης σελίδας είναι συχνά προφανής, μόνο μετά την εμφάνιση της σελίδας και την εύρεση του όρου- όρων στη σελίδα.
- *Περιορισμένη κάλυψη* ***(Limited Coverage)****:* Η εμπειρία της χρήση των διαφορετικών μηχανών αναζήτησης αποδεικνύει ότι η κάλυψη του ιστού από την κάθε μηχανή αναζήτησης ξεχωριστά είναι σχετικά μικρή, δηλαδή πραγματοποιώντας μια δεύτερη αναζήτηση με μια άλλη μηχανή αναζήτησης θα επέστρεφε αρκετά έγγραφα που δεν επέστρεφε η πρώτη μηχανή. Τα αποτελέσματα των Selberg και Etzioni (Selberg και Etzioni, 1995) προτείνουν ότι η κάλυψη του ιστού οποιασδήποτε μηχανής ξεχωριστά είναι περιορισμένη.
- *Περιορισμένη διαθεσιμότητα* ***(Limited Availability)****:* Λόγω των διαφόρων μηχανών αναζήτησης και των δικτυακών προβλημάτων έχει παρατηρηθεί ότι η μηχανή που απαντά γρηγορότερα διαφέρει κάθε φορά.
- *Περιορισμένα User Interfaces* ***(Limited User Interfaces)****:* Είναι δυνατό να προστεθούν διάφορα χαρακτηριστικά που διευκολύνουν τη χρήση των μηχανών αναζήτησης.
- *Μη ενημερωμένες βάσεις δεδομένων* ***(Out of Date Databases)****:* Οι βάσεις δεδομένων των μηχανών αναζήτησης είναι πάντα μη ενημερωμένες. Υπάρχει μια χρονική καθυστέρηση μεταξύ του χρόνου που οι νέες πληροφορίες παρέχονται και του χρόνου που συντάσσεται το ευρετήριο από τις μηχανές αναζήτησης.

5.4.2. Η Μηχανή Αναζήτησης της Inquirus

Ένα από τα βασικά χαρακτηριστικά της μηχανής πολλαπλής αναζήτησης Inquirus είναι ότι αναλύει κάθε έγγραφο και εμφανίζει το κείμενο γύρω από τους όρους της ερώτησης. Το κέρδος της εμφάνισης του παραπάνω κειμένου, παρά μιας περίληψης του εγγράφου, είναι ότι ο χρήστης μπορεί να καθορίσει ευκολότερα εάν το συγκεκριμένο έγγραφο απαντά στην ερώτησή του. Ο χρήστης μπορεί επομένως να βρει έγγραφα υψηλής σχετικότητας γρήγορα ανιχνεύοντας το γύρω κείμενο των όρων της ερώτησης. Αυτή η τεχνική είναι απλή, αλλά μπορεί να αποδειχθεί πολύ αποτελεσματική, ειδικά στην περίπτωση της αναζήτησης του ιστού, όπου η βάση δεδομένων είναι πολύ μεγάλη, διαφορετική, και σε μικρό βαθμό οργανωμένη. Οι χρήστες δείχνουν ότι οι περιλήψεις των σελίδων που δημιουργούνται χρησιμοποιώντας το γύρω κείμενο επιτρέπουν σε αυτούς να αξιολογήσουν τη σχετικότητα των εγγράφων ευκολότερα και γρηγορότερα.

Η εμφάνιση του κειμένου γύρω από τους όρους της ερώτησης δεν απαιτεί τη χρήση των μηχανών πολλαπλής αναζήτησης και μπορεί να είναι πολύ χρήσιμο ακόμα κι αν μόνο μια μηχανή χρησιμοποιηθεί. Ωστόσο, όπως και οι άλλες μηχανές πολλαπλής αναζήτησης, η Inquirus θέτει παράλληλα τα ερωτήματα του χρήστη στις μηχανές πολλαπλής αναζήτησης. Τα σημαντικότερα χαρακτηριστικά της μηχανής πολλαπλής αναζήτησης Inquirus είναι (Lawrence and Giles, 1998):

- Εμφάνιση του κειμένου που περιέχει τους όρους του ερωτήματος,
- Προηγμένη δυνατότητα εντοπισμού των διπλών εγγραφών
- Σταδιακή εμφάνιση των αποτελεσμάτων
- Εμφάνιση των όρων της ερώτησης πιο έντονα στις σελίδες
- Εισαγωγή γρήγορων συνδέσεων για τη μετάβαση στους όρους της ερώτησης στις μεγάλες σελίδες
- Εντυπωσιακά μεγαλύτερη ακρίβεια για ορισμένες ερωτήσεις με τη χρήση συγκεκριμένων εκφραστικών μορφών και
- Καλύτερη ταξινόμηση με βάση τη σχετικότητα της σελίδας.

Ένας πληρέστερος κατάλογος ακολουθεί:

1. Η μηχανή κατεβάζει τις πραγματικές σελίδες που αντιστοιχούν στα *hits* και ψάχνει σε αυτές τους όρους της ερώτησης. Η μηχανή στη συνέχεια εμφανίζει το κείμενο στο οποίο περιέχονται οι όροι της ερώτησης και όχι μια περίληψη της σελίδας. Αυτό γενικά παρέχει μια πολύ καλύτερη ένδειξη της σχετικότητας μιας σελίδας από ότι

παρέχουν οι περιλήψεις που χρησιμοποιούνται από τις άλλες μηχανές αναζήτησης, και επίσης συχνά βοηθά να αποφεύγεται η αναζήτηση των σελίδων που δεν περιέχουν τις απαραίτητες πληροφορίες. Το πλαίσιο κειμένου γύρω από τους όρους της ερώτησης μπορεί να είναι ιδιαίτερα χρήσιμο, όταν μια αναζήτηση περιλαμβάνει όρους που μπορεί να εμφανιστούν σε διαφορετικά πλαίσια από αυτό που απαιτείται. Το μέγεθος του πλαισίου κειμένου καθορίζεται από το χρήστη θέτοντας τον αριθμό των χαρακτήρων που θα εμφανιστούν σε κάθε πλευρά των όρων της ερώτησης. Οι περισσότεροι μη-αλφαριθμητικοί χαρακτήρες φιλτράρονται από το πλαίσιο του κειμένου προκειμένου να παραχθούν τα πιο αναγνώσιμα και πληροφοριακά αποτελέσματα.

2. Τα αποτελέσματα επιστρέφονται σταδιακά αφότου κάθε σελίδα γίνεται ***download*** και αναλύεται, παρά όταν όλες οι σελίδες γίνουν *download*. Το πρώτο αποτέλεσμα επιστρέφεται χαρακτηριστικά γρηγορότερα από το μέσο χρόνο απόκρισης μιας μηχανής αναζήτησης. Όταν πολλές σελίδες παρέχουν τις πληροφορίες που απαιτούνται, η αρχιτεκτονική της μηχανής πολλαπλής αναζήτησης μπορεί να είναι χρήσιμη επειδή οι γρηγορότερες σελίδες είναι οι πρώτες που αναλύονται και που εμφανίζονται.

3. Κατά την εξέταση του πλήρους περιεχομένου των σελίδων που αντιστοιχούν στα *hits*, οι σελίδες φιλτράρονται για εμφανίσουν πιο έντονα τους όρους της ερώτησης, και συνδέσεις εισάγονται στην κορυφή της σελίδας που μεταπηδούν στην πρώτη εμφάνιση του κάθε όρου της ερώτησης. Συνδέσεις σε κάθε εμφάνιση του όρου της ερώτησης μεταπηδούν στην επόμενο του αντίστοιχου όρου. Η πιο έντονη εμφάνιση του όρου της ερώτησης βοηθά στον γρήγορο προσδιορισμό των όρων της ερώτησης και της σχετικότητας των σελίδων.

4. Οι σελίδες που δεν είναι πλέον διαθέσιμες μπορούν να προσδιοριστούν. Αυτές οι σελίδες παρατίθενται στο τέλος της απάντησης. Μερικές άλλες υπηρεσίες πολλαπλής αναζήτησης παρέχουν επίσης την ανίχνευση των ***"dead links"***, ωστόσο το χαρακτηριστικό αυτό δεν διατίθεται συνήθως ως προεπιλογή και κανένα αποτέλεσμα δεν επιστρέφεται έως ότου ελεγχθούν όλες οι σελίδες. Για τη μηχανή πολλαπλής αναζήτησης Inquirus το χαρακτηριστικό αυτό είναι εγγενές στην αρχιτεκτονική της μηχανής, και είναι σε θέση να παραγάγει αποτελέσματα επαυξητικά και γρήγορα.

5. Οι σελίδες που δεν περιέχουν πλέον τους όρους αναζήτησης ή αυτές που δεν ταιριάζουν απολύτως με την ερώτηση μπορούν να προσδιοριστούν. Αυτές οι σελίδες παρατίθενται μετά από τις σελίδες που ταιριάζουν με την ερώτηση. Αυτό μπορεί να είναι πολύ σημαντικό – διαφορετικές μηχανές χρησιμοποιούν διαφορετικές τεχνικές

σχετικότητας, και εάν μια μηχανή επιστρέφει αποτελέσματα μικρής σχετικότητας, αυτό μπορεί να οδηγήσει σε αποτελέσματα μικρής σχετικότητας από τις γενικές τεχνικές πολλαπλής αναζήτησης. Οι όροι αναζήτησης στα *meta tags* αντιμετωπίζονται σαν να ήταν μέρος του κυρίως κειμένου.

6. Προηγμένη δυνατότητα εντοπισμού των διπλών εγγραφών επιτυγχάνεται από την Inquirus. Οι σελίδες θεωρούνται αντίγραφα εάν το σχετικό πλαίσιο κειμένου είναι ίδιο. Αυτό επιτρέπει την ανίχνευση ενός αντιγράφου εάν η σελίδα έχει μια διαφορετική κεφαλίδα ή μια υποσημείωση.

7. Ο Kirsch (Kirsch, 1997) έχει παρουσιάσει μια τεχνική για την ταξινόμηση των σελίδων σύμφωνα με τις τεχνικές πολλαπλής αναζήτησης, όπου οι μηχανές αναζήτησης τροποποιούνται για να επιστρέψουν πρόσθετες πληροφορίες όπως τον αριθμό εμφάνισης κάθε όρου αναζήτησης στα έγγραφα και τον αριθμό εμφάνισής τους σε ολόκληρη τη βάση δεδομένων. Μια τέτοια τεχνική δεν απαιτείται για τη μηχανή πολλαπλής αναζήτησης Inquirus, επειδή οι πραγματικές σελίδες γίνονται *download* και αναλύονται. Είναι επομένως πιθανόν να εφαρμοσθεί μια ομοιόμορφη μέθοδος ταξινόμησης στα έγγραφα που επιστρέφονται από διαφορετικές μηχανές. Η μηχανή εμφανίζει τις σελίδες με φθίνουσα σειρά με βάση τον αριθμό των όρων της ερώτησης που περιέχονται στο έγγραφο (εάν καμιά από τις λίγες πρώτες σελίδες δεν περιέχει όλους τους όρους της ερώτησης τότε η μηχανή εμφανίζει αρχικά τα αποτελέσματα που περιέχουν τον μέγιστο αριθμό των όρων της ερώτησης που βρίσκονται σε μια σελίδα μέχρι εκείνη τη στιγμή). Αφότου έχουν γίνει *download* όλες οι σελίδες, η μηχανή κατόπιν ταξινομεί τις σελίδες σύμφωνα με ένα απλό μέτρο σχετικότητας. Αυτό το μέτρο εξετάζει τον αριθμό των όρων της ερώτησης που βρίσκονται στο έγγραφο, την εγγύτητα μεταξύ των όρων της ερώτησης και της συχνότητας εμφάνισης του όρου (η συνηθισμένη *αντίστροφη συχνότητα των εγγράφων* ***(Inverse Document Frequency)*** μπορεί επίσης να είναι χρήσιμη):

$$R = c_1 N_p + \left(c_2 - \frac{\sum_{i=1}^{N_p-1} \sum_{j=i+1}^{N_p} \min(d(i,j), c_2)}{\sum_{k=1}^{N_p-1} (N_p - k)} \right) / \frac{c_2}{c_1} + \frac{N_t}{c_3}$$

όπου:

- N_P είναι ο αριθμός των όρων της ερώτησης που εμφανίζονται στο έγγραφο (κάθε όρος λαμβάνεται υπόψη μόνο μια φορά),

- N_t είναι ο συνολικός αριθμός των όρων της ερώτησης στο έγγραφο (κάθε όρος λαμβάνεται υπόψη όσες φορές εμφανίζεται),
- $d(i,j)$ είναι η ελάχιστη απόσταση μεταξύ του i και j όρου της ερώτησης που βρίσκονται στο έγγραφο (ο αριθμός των χαρακτήρων),
- c_1 είναι μια σταθερά που ελέγχει το γενικό μέγεθος του R,
- c_2 είναι μια σταθερά που καθορίζει τη μέγιστη απόσταση μεταξύ των όρων της ερώτησης που θεωρείται χρήσιμη, και
- c_3 είναι μια σταθερά που καθορίζει τη σημασία της συχνότητας εμφάνισης του όρου (ισχύει: $c_1 = 100$, $c_2 = 5000$, και $c_3 = 10\ c_1$).

Όταν υπάρχει μόνο ένας όρος ερώτησης τότε χρησιμοποιούμε την απόσταση από την έναρξη της σελίδας μέχρι την πρώτη εμφάνιση του όρου ως δείκτη σχετικότητας. Αυτό το κριτήριο ταξινόμησης μπορεί να είναι ιδιαίτερα χρήσιμο στις αναζητήσεις ιστού. Μια ερώτηση με πολλούς όρους στον ιστό επιστρέφει συχνά έγγραφα που περιέχουν όλους τους όρους, αλλά οι όροι βρίσκονται διασκορπισμένοι, είναι μακριά ο ένας από τον άλλο και μπορούν να βρίσκονται στα ανεξάρτητα τμήματα της σελίδας.

8. Η μηχανή δεν χρησιμοποιεί το χαμηλότερο κοινό παρονομαστή για τη σύνταξη της αναζήτησης. Η μηχανή υποστηρίζει όλους τους κοινούς τρόπους αναζήτησης συμπεριλαμβανομένης της σύνταξης Boolean. Οι ερωτήσεις τροποποιούνται δυναμικά προκειμένου να ταιριάζουν με τη σύνταξη της κάθε μηχανής αναζήτησης (και άλλες μηχανές πολλαπλής αναζήτησης έχουν αυτή τη δυνατότητα).

9. Η Inquirus χρησιμοποιεί μια *συγκεκριμένη εκφραστική μορφή* ***(specific expressive forms)*** τεχνικής αναζήτησης, η οποία μπορεί εντυπωσιακά να βελτιώσει την ακρίβεια για ορισμένες ερωτήσεις. Η τεχνική λειτουργεί με την έρευνα συγκεκριμένων τρόπων ως απάντηση σε μια ερώτηση.

5.4.3. Αρχιτεκτονική της Inquirus

Η μηχανή αποτελείται από δύο κύρια λογικά μέρη:

- ✓ τον κώδικα πολλαπλής αναζήτησης ***(meta search code)*** και
- ✓ μια παράλληλη σελίδα ανάκτησης *daemon* ***(parallel page retrieval daemon)***.

Ο ψευδοκώδικας, για μια πιο απλοποιημένη έκδοση, του κώδικα αναζήτησης είναι ο παρακάτω:

```
Process the request to check syntax and create regular expressions which are used to..
        ..match query terms
Send requests (modified appropriately) to all relevant search engines
Loop for each page retrieved until maximum number of results or all pages retrieved
        If page is from a search engine
                Parse search engine response extracting hits and any link for the next..
                        ..set of results
                Send requests for all of the hits
                Send request for the next set of results if applicable
        Else
                Check page for query terms and create context strings if found
                Print page information and context strings if all query terms are found..
                        ..and duplicate context strings have not been encountered before
        Endif
End loop

Re-rank pages using proximity and term frequency information
```

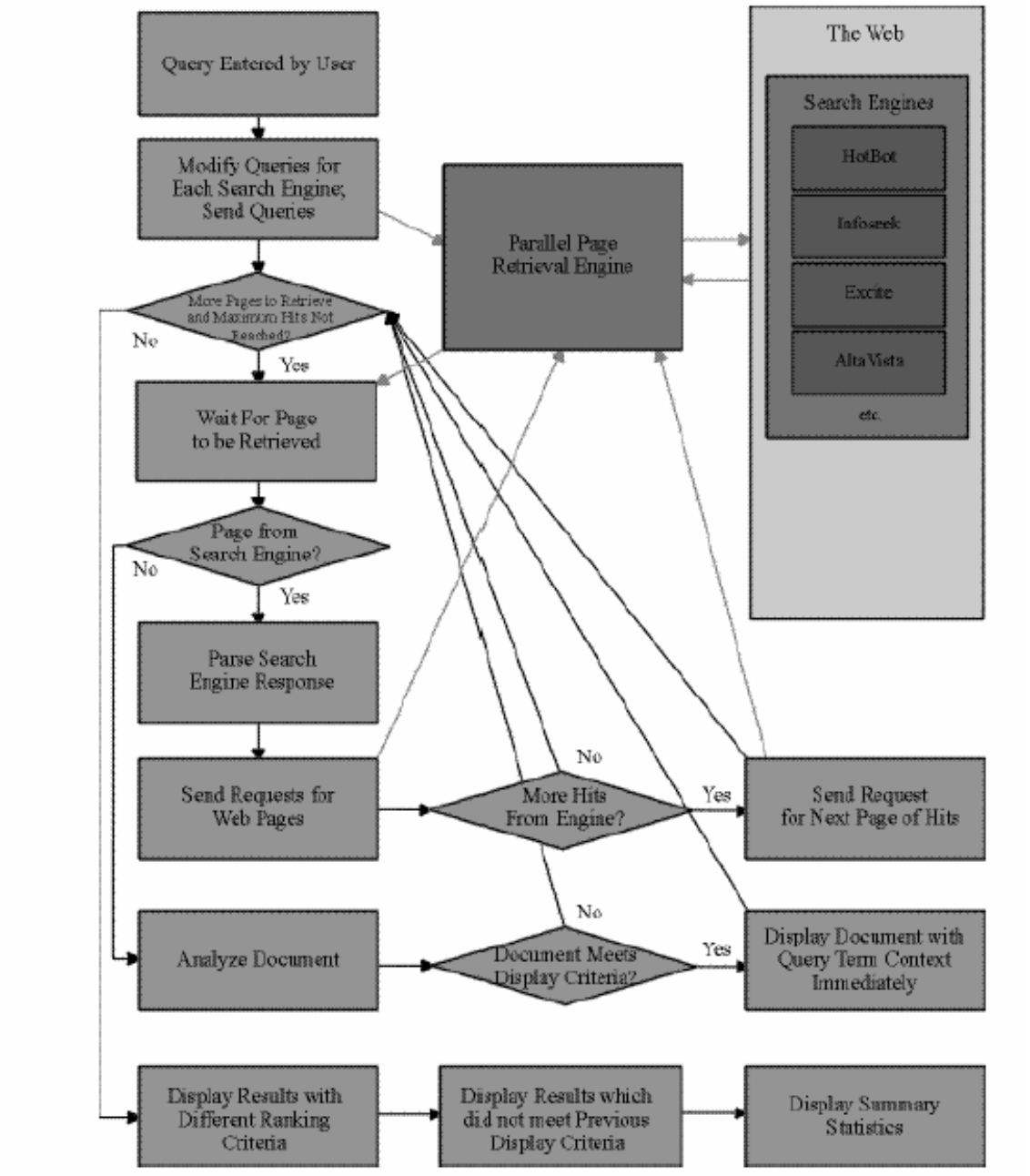
```
Print page information and context strings for pages which contained some but not all..
        ..query terms
Print page information for pages which contained no query terms
Print page information and context strings for pages which contain duplicate context strings
Print page information for pages which could not be downloaded
Print summary statistics
```

Το παρακάτω Σχήμα 5.7. παρουσιάζει το απλουστευμένο διάγραμμα ροής ελέγχου της μηχανής πολλαπλής αναζήτησης. Η μηχανή ανάκτησης των σελίδων είναι σχετικά απλή, αλλά ενσωματώνει αρκετά χαρακτηριστικά όπως αιτήματα αναμονής και εξισορρόπηση του φορτίου από τις πολλαπλές διαδικασίες αναζήτησης, και καθυστερεί αιτήματα για την ίδια περιοχή ώστε να αποτραπεί η υπερφόρτωσή της. Η μηχανή ανάκτησης σελίδων αποτελείται από μια dispatch daemon και από διάφορες *διαδικασίες ανάκτησης πελατών* ***(client retrieval processes)***. Οι διαδικασίες των πελατών ανακτούν απλά τις σχετικές σελίδες, χειρίζονται τα λάθη και τα διαλείμματα, και επιστρέφουν τις σελίδες άμεσα στην κατάλληλη διαδικασία αναζήτησης.

5.4.4. Αποδοτικότητα

Μια απλή ανάλυση των χρόνων ανάκτησης των σελίδων οδηγεί σε μερικά ενδιαφέροντα συμπεράσματα. Ο Πίνακας 5.3. παρουσιάζει το μέσο χρόνο απόκρισης για κάθε μια από τις έξι κύριες μηχανές αναζήτησης, καθώς και το μέσο χρόνο για την πρώτη απόκριση των έξι μηχανών, όταν οι ερωτήσεις τίθενται ταυτόχρονα σε όλες τις μηχανές αναζήτησης, και το μέσο χρόνο απόκρισης για τη μηχανή αναζήτησης Inquirus (εμφάνιση του πρώτου αποτελέσματος). Μπορεί να φανεί ότι, κατά μέσον όρο, η παράλληλη αρχιτεκτονική της Inquirus επιτρέπει την εύρεση, να κάνει download και να αναλύει την πρώτη σελίδα γρηγορότερα από την εμφάνιση του αποτελέσματος στις τυπικές μηχανές αναζήτησης, ακόμα κι αν οι κύριες μηχανές δεν κάνουν download και

δεν αναλύουν τις σελίδες. Η μηχανή Inquirus είναι εκπληκτικά γρήγορη, κάτι το οποίο υποστηρίζουν και οι ίδιοι οι χρήστες. Αυτά τα αποτελέσματα πάρθηκαν από 1.000 ερωτήσεις που τέθηκαν στις μηχανές και σημειώνεται ότι η σχετική ταχύτητα των μηχανών αναζήτησης ποικίλλει σημαντικά μέσα στο χρόνο, εξαρτώμενη και από την τοποθεσία της σελίδας.



Σχήμα 5.7. Το Απλουστευμένο Διάγραμμα Ροής Ελέγχου της Μηχανής Πολλαπλής Αναζήτησης (Lawrence and Giles, 1998)

Ένα μειονέκτημα της μηχανής πολλαπλής αναζήτησης Inquirus είναι το γεγονός

ότι χρησιμοποιεί σχετικά μικρό εύρος σε σχέση με τις άλλες μηχανές αναζήτησης. Οι απαιτήσεις σε πρόσθετο εύρος μπορεί να περιορίσουν τον αριθμό των χρηστών που μπορεί ταυτοχρόνως να χρησιμοποιήσει έναν server ή να παρουσιάσει ένα μειονέκτημα εάν η πρόσβαση στο internet ανάλογα με το μέγεθος των δεδομένων που μεταφέρονται. Σημειώνεται ότι οι απαιτήσεις σε εύρος δεν πρόκειται να είναι σημαντικές στο μέλλον.

ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ	ΜΕΣΟΣ ΧΡΟΝΟΣ ΑΠΟΚΡΙΣΗΣ (SECONDS)
AltaVista	0.9
Infoseek	1.3
HotBot	2.6
Excite	5.2
Lycos	2.8
Northern Light	7.5
Όλων των μηχανών (μέσος όρος)	2.7
Η πρώτη από τις 6 μηχανές	0.8
Πρώτο αποτέλεσμα της Inquirus	1.3

Πίνακας 5.3.: Οι Μέσοι Χρόνοι Απόκρισης των Μηχανών και των Μηχανών Πολλαπλής Αναζήτησης (Lawrence and Giles, 1998)

5.4.5. Συμπεράσματα

Η μηχανή πολλαπλής αναζήτησης Inquirus καταδεικνύει ότι η σε πραγματικό χρόνο ανάλυση των εγγράφων που επιστρέφονται από τις μηχανές αναζήτησης του ιστού είναι εφικτή. Στην πραγματικότητα, η κλήση των μηχανών αναζήτησης ιστού και κάνοντας παράλληλα *download* τις ιστοσελίδες επιτρέπει στη μηχανή πολλαπλής αναζήτησης Inquirus, κατά μέσον όρο, να εμφανίζει το πρώτο αποτέλεσμα πιο γρήγορα από μια τυπική μηχανή αναζήτησης. Η αλληλεπίδραση της μηχανής με τον χρήστη δείχνει ότι η εμφάνιση σε πραγματικό χρόνο του κειμένου γύρω από τους όρους της ερώτησης, και η πιο έντονη εμφάνιση των όρων της ερώτησης στα έγγραφα όταν εμφανίζονται, βελτιώνει σημαντικά την αποδοτικότητα της αναζήτησης του ιστού.

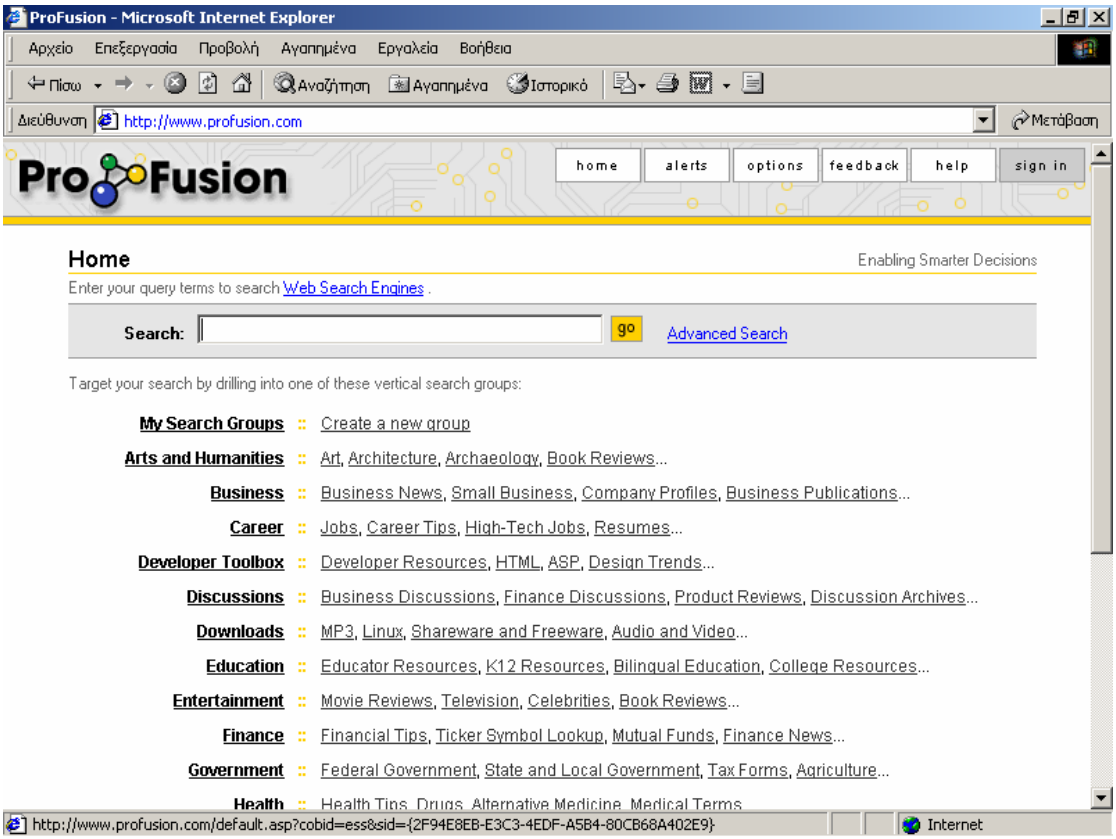
5.5. ProFusion (www.profusion.com)

5.5.1. Εισαγωγή

Δημιουργήθηκε από την Intelliseek το 2000, η ProFusion αναζητά με βάση τις καλύτερες μηχανές του διαδικτύου. Χρησιμοποιεί προχωρημένες τεχνολογίες πληροφοριών. Ο στόχος της Intelliseek είναι να αυξήσει την ισόνομη γνώση, να αυξήσει την παραγωγικότητα των υπαλλήλων, να γίνει περισσότερο ανταγωνιστική κατανοώντας τους ανταγωνιστές της, και επιπλέον, να αυξήσει την ικανοποίηση των πελατών και να μειώσει τις δαπάνες υποστήριξής τους. Η Profusion χρησιμοποιεί τον ***Intelliseek Enterprise Search Server (ESS)***, ο οποίος παρέχει ένα ενοποιημένο *interface* αναζήτησης σε χιλιάδες πηγές πληροφοριών και βάσεις δεδομένων πέρα από το εταιρικό Intranet, Extranet, τις συνδρομητικές πηγές και το διαδίκτυο συμπεριλαμβανομένου των AltaVista, AOL, Britannica, Netscape, και Yahoo!. Ο ESS χαρακτηρίζεται από μια *προσαρμοστική τεχνολογία αναζήτησης* ***(Adaptive Search Technology)*** που μεταβιβάζει τις ερωτήσεις του χρήστη στις καλύτερες πηγές για τα πιο σχετικά αποτελέσματα και τους χρήστες να μένουν πάνω από τις σχετικές πληροφορίες μέσω ενός αυτοματοποιημένου συστήματος ανίχνευσης και προειδοποίησης. Στους προηγούμενους 12 μήνες, η Intelliseek έχει λάβει πάνω από 30 σημαντικά βραβεία από βιομηχανίες και έχει συνάψει πολλές στρατηγικές συνεργασίες, μερικές από τις οποίες με τη ZDNet, InfoSpace και Lycos.

5.5.2. Βασικά Χαρακτηριστικά

Η ProFusion έχει ένα από τα πιο σύγχρονα interfaces (Σχήμα 5.8) που επιτρέπει τον δίκαιο και εκτεταμένο έλεγχο του χρήστη στα αποτελέσματα. Επιτρέπει επίσης τη χρήση μιας πλήρους έκφρασης Boolean (τη μεταβιβάζει στις μηχανές που την κάνουν αποδεκτή). Ωστόσο, όπως ακριβώς συμβαίνει και με τις περισσότερες μηχανές πολλαπλής αναζήτησης κάποιο μέρος αυτών των εξελιγμένων τεχνικών δεν χρησιμοποιείται πλήρως, λόγω του γεγονότος ότι από κάθε μηχανή ανακτώνται 10 με μέγιστο τα 25 αρχεία. Η ProFusion ελέγχει επίσης τις συνδέσεις για να δει αν είναι λειτουργικές. Τέλος όπως όλες οι μηχανές, έτσι και η ProFusion υποστηρίζει την προχωρημένη αναζήτηση.



Σχήμα 5.8.: Η Αρχική Σελίδα της Μηχανής Πολλαπλής Αναζήτησης ProFusion (www.profusion.com)

✓ ***Οι μηχανές / Κατάλογοι Αναζήτησης:***

Κύριες Μηχανές Αναζήτησης:	3
Συνολικός Αριθμός Μηχανών/ Καταλόγων:	9
Επιλογή της Boolean:	Ναι
Διαγραφή των Διπλών Εγγραφών ή συνδυασμός τους:	Ναι
Μέγιστος Αριθμός Εγγράφων από Κάθε Μηχανή:	20
Μέγιστος Αριθμός Ανακτημένων Εγγράφων:	<200

• About.com	• All the Web
• AltaVista	• AOL
• Yahoo!	• Britannica
• Excite	• MSN
• LookSmart	• Netscape

• Lycos	• Adope PDF Online

✓ ***Σύνταξη Αναζήτησης***

Η ProFusion υποστηρίζει όλους τους τελεστές Boolean (AND, OR, NOT) και NEAR. Για τις μηχανές που υποστηρίζουν τους τελεστές Boolean και όχι τον τελεστή NEAR (Excite, WebCrawler), η Profusion αλλάζει τον τελεστή NEAR σε AND.

✓ ***Περιορισμός στον Αριθμό των Αρχείων***

Μπορεί ο χρήστης να επιλέξει τον συνολικό αριθμό των εγγράφων σε 10, 20, 50, 99 ή να εμφανιστούν όλα τα αρχεία, που αυτό μεταφράζεται σε λιγότερα από 200 για κάθε αναζήτηση.

✓ ***Εμφάνιση Αποτελεσμάτων***

Η Profusion εξαλείφει τις διπλές εγγραφές, εμφανίζει ποια μηχανή ανέκτησε το κάθε έγγραφο και ταξινομεί τα έγγραφα με βάση τη σχετικότητά τους.

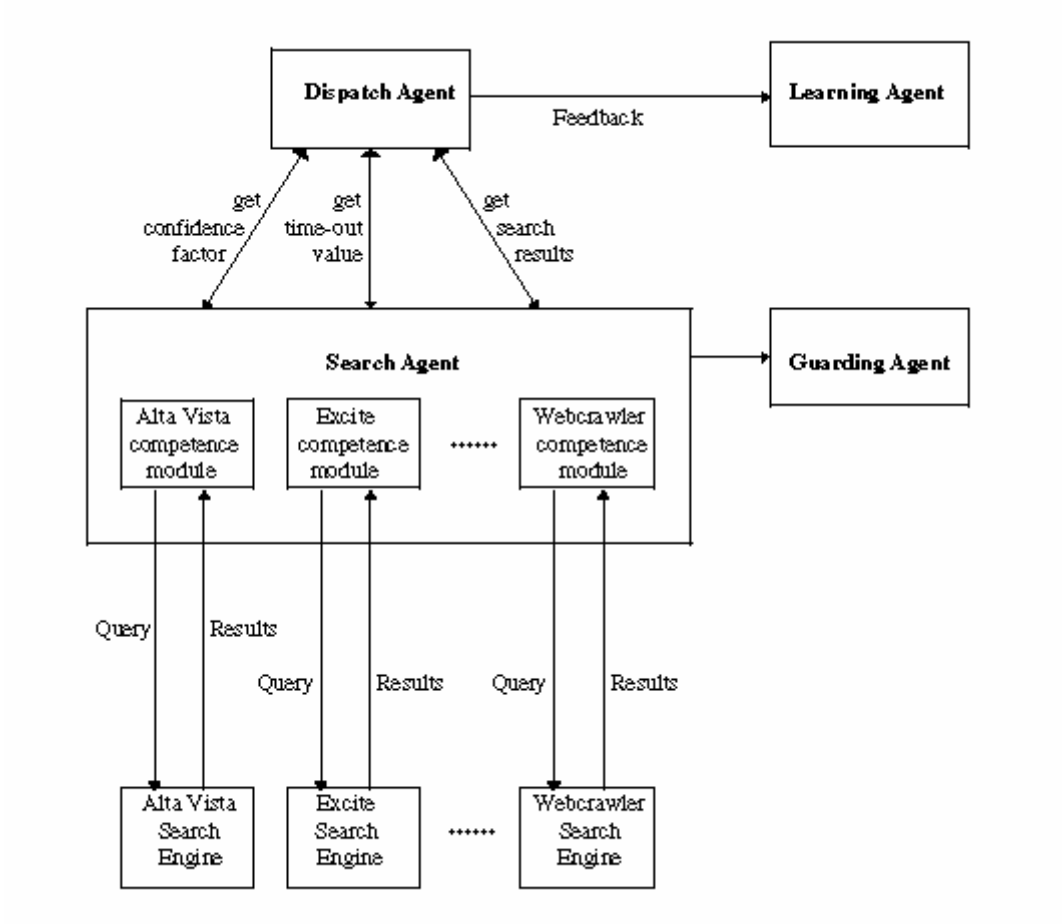
5.5.3. Αρχιτεκτονική της ProFusion

Η ProFusion αναλύει τις εισερχόμενες (υποβαλλόμενες) ερωτήσεις των χρηστών, τις κατηγοριοποιεί, και επιλέγει αυτόματα τις καλύτερες μηχανές αναζήτησης για κάθε ερώτηση, βασισμένη στην ήδη υπάρχουσα γνώση (*παράγοντες εμπιστοσύνης* ***(confidence factors)***), που αντιπροσωπεύει την καταλληλότητα κάθε μηχανής αναζήτησης για κάθε κατηγορία. Χρησιμοποιεί αυτούς τους παράγοντες εμπιστοσύνης για να συγχωνεύσει τα αποτελέσματα της αναζήτησης σε έναν κατάλογο με βάση τα βάρη των επιστρεφόμενων εγγράφων, αφαιρώντας τις διπλές εγγραφές και τις προαιρετικά μη λειτουργικές συνδέσεις και παρουσιάζει τον τελικό ταξινομημένο κατάλογο των αποτελεσμάτων στο χρήστη.

Το *σύστημα των πολυ-πρακτόρων* ***(multi-agent system)*** αποτελείται από τέσσερις διαφορετικούς τύπους πρακτόρων (Fan and Gauch, 1997), δηλαδή:

- ένας *πράκτορας αποστολών* ***(dispatch agent)***,
- ένας *πράκτορας αναζήτησης* ***(search agent)***,
- ένας *πράκτορας εκμάθησης* ***(learning agent)***, και
- ένας *πράκτορας φρουρός* ***(guarding agent)***.

Ο *πράκτορας αποστολών* επικοινωνεί με το χρήστη και αποστέλλει έπειτα τις ερωτήσεις στον *πράκτορα αναζήτησης* και στον *πράκτορα εκμάθησης*. Ο *πράκτορας αναζήτησης* αλληλεπιδρά με τις μηχανές αναζήτησης και είναι αρμόδιος για την υποβολή των αποτελεσμάτων της αναζήτησης, για τους παράγοντες εμπιστοσύνης, και για την απενεργοποίηση των μηχανών αναζήτησης από τον *πράκτορα αποστολών*, καθώς επίσης και για την επίκληση του *πράκτορα φρουρού* όταν χρειάζεται. Ο *πράκτορας εκμάθησης* είναι υπεύθυνος για την εκμάθηση και την ανάπτυξη των μηχανών αναζήτησης, ειδικότερα ρυθμίζει τους παράγοντες εμπιστοσύνης. Ο *πράκτορας φρουρός* επικαλείται όταν μια μηχανή αναζήτησης δεν λειτουργεί και είναι αρμόδιος για την παρεμπόδιση της αποστολής μελλοντικών ερωτήσεων σε μια μηχανή αναζήτησης που δεν ανταποκρίνεται. Επίσης είναι υπεύθυνος και για την ανίχνευση της επαναλειτουργίας της μηχανής αναζήτησης. Το Σχήμα 5.9 παρουσιάζει τη ροή ελέγχου και την ενδοεπικοινωνία μεταξύ των πρακτόρων στο σύστημα της ProFusion.



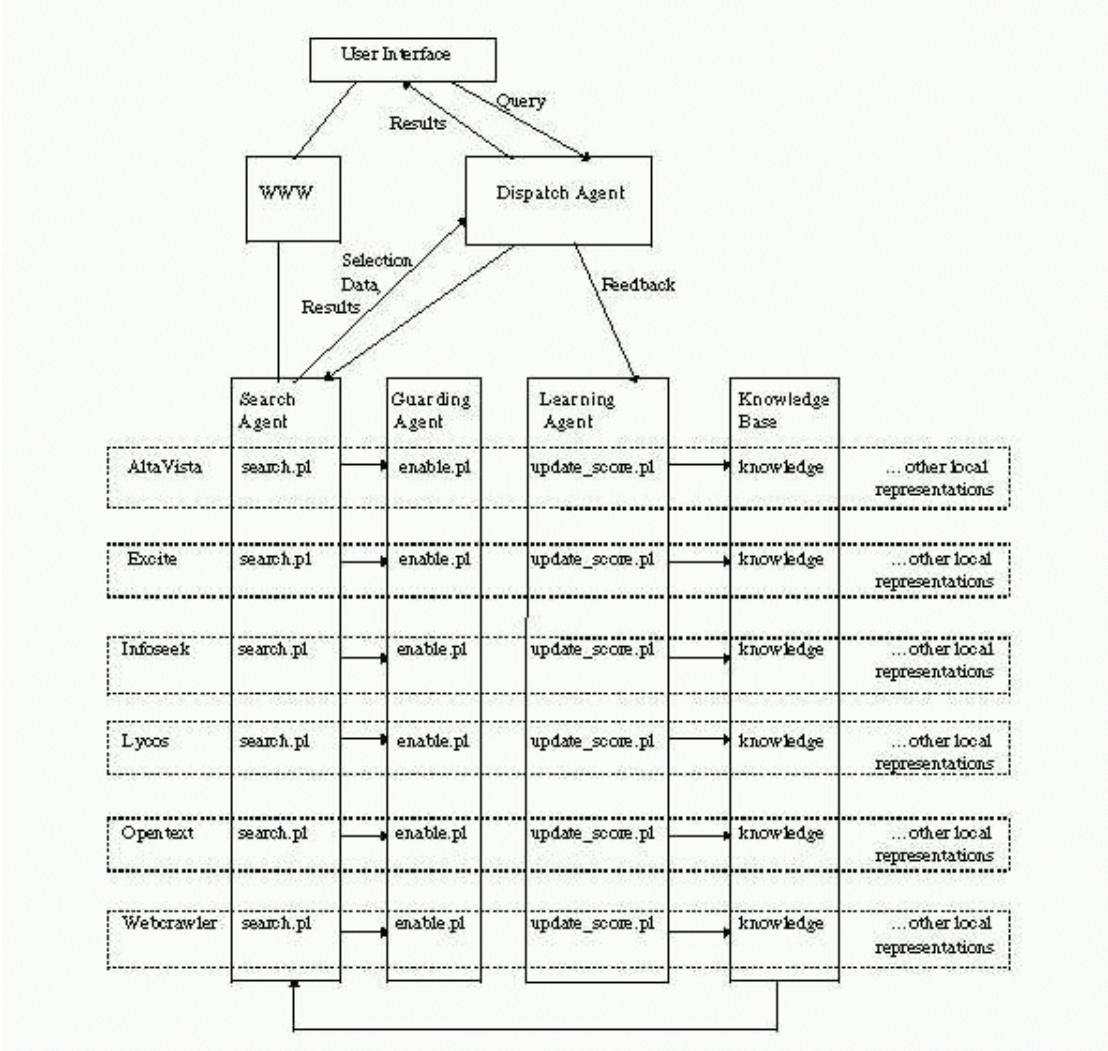
Σχήμα 5.9.: Το Διάγραμμα της Ροής Ελέγχου και της Ενδοεπικοινωνία των Πρακτόρων (Fan and Gauch, 1997)

Η αρχιτεκτονική των πολυ-πρακτόρων καταδεικνύει τα διάφορα επιθυμητά χαρακτηριστικά των πρακτόρων που περιλαμβάνουν: προσανατολισμένες ως προς την εργασία μέρη, λύσεις συγκεκριμένων εργασιών, τις με έμφαση αντιπροσωπεύσεις, την αποκέντρωση της δομής ελέγχου, και την εκμάθηση και την ανάπτυξη. Ο πράκτορας αναζήτησης, ο πράκτορας εκμάθησης, και ο πράκτορας φρουρός, κάθε ένας τους αποτελείται από ένα σύνολο έξι όμοιων ικανοτήτων, κάθε μια από τις οποίες είναι αρμόδια για τη μια από τις έξι μηχανές αναζήτησης (προσανατολισμένα ως προς την εργασία μέρη). Αυτά τα τμήματα ικανότητας είναι ανεξάρτητα μαύρα κουτιά που χειρίζονται όλη την αντιπροσώπευση, τους υπολογισμούς, την αιτιολόγηση, και την εκτέλεση που είναι απαραίτητα για κάθε μηχανή αναζήτησής της. Αν και τα έξι τμήματα ικανότητας για κάθε έναν από τους τρεις πράκτορες, δημιουργούνται χρησιμοποιώντας τον ίδιο κώδικα, κάθε ένας χρησιμοποιεί τα τοπικά δικά του αρχεία διαμόρφωσης και γνώσης για να επιτύχει την ικανότητά του (ικανότητα προσανατολισμένη ως προς την εργασία). Με άλλα λόγια, δεν υπάρχει καμία κεντρική αντιπροσώπευση κοινή για τις διάφορες ενότητες. Αντί αυτού, κάθε ενότητα προσανατολισμένη ως προς την εργασία αντιπροσωπεύει τοπικά οτιδήποτε χρειάζεται για να λειτουργήσει αυτόνομα. Οι τοπικές αντιπροσωπεύσεις των διαφορετικών ενοτήτων δεν συσχετίζονται ***(de-emphasized represantations)***.

Το Σχήμα 5.10 παρουσιάζει την αρχιτεκτονική μορφή του συστήματος πολυ-πρακτόρων της ProFusion (Fan and Gauch, 1997). Αυτή η αρχιτεκτονική είναι ιδιαίτερα διανεμημένη και *αποκεντρικοποιημένη* ***(decentralized)***. Κάθε μηχανή αναζήτησης κρατά τις ενότητες ικανότητας και τις τοπικές αντιπροσωπεύσεις της σε έναν ξεχωριστό κατάλογο. Εκτός από τον πράκτορα αποστολών, όλες οι ενότητες ικανότητας του πράκτορα αναζήτησης, ο πράκτορας εκμάθησης, και ο πράκτορας φρουρός λειτουργούν παράλληλα. Καμία από τις ενότητες δεν ελέγχεται από άλλες ενότητες (αποκεντρωμένη δομή ελέγχου). Λόγω αυτής της διανεμημένης λειτουργίας, το νέο σύστημα είναι σε θέση να αντιδρά γρήγορα στις αλλαγές στο γύρω περιβάλλον και να πραγματοποιεί τις αντίστοιχες προσαρμογές.

Οι ρυθμίσεις γίνονται από τον πράκτορα εκμάθησης που χρησιμοποιεί προσαρμοστικούς αλγορίθμους. Η νέα έκδοση της ProFusion προσαρμόζεται στις αλλαγές της απόδοσης της μηχανής αναζήτησης, στις αλλαγές του χρόνου απόκρισης των μηχανών αναζήτησης και στις αλλαγές των formats του αποτελέσματος των μηχανών αναζήτησης. Η προσαρμογή της απόδοσης επιτυγχάνεται με την παρατήρηση της συμπεριφοράς των χρηστών για την ανατροφοδότηση που αλλάζει δυναμικά την

απόδοση της βάσης γνώσεων, την προσαρμογή στο χρόνο απόκρισης που επιτυγχάνεται με τη χρήση των δυναμικά μεταβαλλόμενων τιμών διαλείμματος, και η προσαρμογή των formats των αποτελεσμάτων επιτυγχάνεται με τη χρήση ενός δυναμικού σχεδίου εξαγωγής, ή με έναν parser.



Σχήμα 5.10.: Η Αρχιτεκτονική της Μηχανής Πολλαπλής Αναζήτησης ProFusion (Fan and Gauch, 1997)

Με αυτήν την προσαρμοστική αρχιτεκτονική των πολυ-πρακτόρων, το σύστημα ProFusion είναι τώρα πιο ανταγωνιστικό στο δυναμικό περιβάλλον του ιστού δεδομένου ότι προσαρμόζεται αυτόματα στις αλλαγές του περιβάλλοντός του. Η ProFusion είναι επίσης πολύ ευκολότερο να διατηρηθεί και να επεκταθεί επειδή δεν απαιτεί πλέον την ήδη υπάρχουσα γνώση παραγόντων εμπιστοσύνης μιας νέας μηχανής αναζήτησης για κάθε κατηγορία (αυτό θα καθοριστεί από το πράκτορα εκμάθησης).

Δεδομένου ότι ο πράκτορας αναζήτησης ενσωματώνει έναν parser, δεν απαιτείται άλλος κώδικας για την εξαγωγή των αποτελεσμάτων αναζήτησης, μόνο μια περιγραφή της γλώσσας που η μηχανή αναζήτησης χρησιμοποιεί.

5.6. Άλλες Μηχανές Πολλαπλής Αναζήτησης

Όπως αναφέρθηκε στην αρχή αυτού του κεφαλαίου, σε αυτή την ενότητα θα περιγράψουμε κάποιες άλλες σημαντικές μηχανές πολλαπλής αναζήτησης, όμως η περιγραφή τους θα είναι περιληπτική. Έτσι καλύπτεται καλύτερα και η ενότητα των μηχανών πολλαπλής αναζήτησης και η σύγκρισή τους, που θα γίνει παρακάτω, θα είναι πιο κατανοητή.

5.6.1. Dogpile (www.dogpile.com)

Η Dogpile εμφανίστηκε πριν μερικά έτη. Ήταν μια από τις πρώτες μηχανές πολλαπλής αναζήτησης που έκανε χρήση της ταυτόχρονης αναζήτησης. Πρόσφατα, η Go2Net (ιδιοκτήτης της Metacrawler.com και βασικός ανταγωνιστής της Dogpile) απόκτησε τη Dogpile. Σήμερα, η Go2Net είναι ιδιοκτήτης των δύο μεγαλύτερων μηχανών πολλαπλής αναζήτησης του ιστού. Η αρχική της σελίδα φαίνεται στο παρακάτω Σχήμα 5.11.

Σχήμα 5.11.: Η Αρχική Σελίδα της Μηχανής Πολλαπλής Αναζήτησης Dogpile (www.dogpile.com)

Η Dogpile είναι μια μηχανή πολλαπλής αναζήτησης που εντοπίζει και ανακτά τόσα (τα ίδια) αρχεία με αυτά που ανακτά μια μηχανή όταν πραγματοποιεί την αναζήτηση μόνη της. Η Dogpile μεταφράζει τις ερωτήσεις στην πλησιέστερη αποδεκτή μορφή σύνταξης για κάθε μηχανή. Αυτό λειτουργεί καλά κάποιες φορές και κάποιες όχι. Συνοπτικά η Dogpile παρουσιάζει τα παρακάτω χαρακτηριστικά:

Κύριες Μηχανές Αναζήτησης:	3
Συνολικός Αριθμός Μηχανών/ Καταλόγων:	14
Επιλογή της Boolean:	Ναι
Διαγραφή των Διπλών Εγγραφών ή συνδυασμός τους:	Όχι
Μέγιστος Αριθμός Εγγράφων από Κάθε Μηχανή:	Απεριόριστος
Μέγιστος Αριθμός Ανακτημένων Εγγράφων:	Απεριόριστος

Η Dogpile επιτρέπει τις αναζητήσεις του ιστού, Usenet, FTP, νέα, εικόνων, πλειστηριασμούς, SmallBiz (Hypermart) και Audio/MP3. Δηλαδή, η Dogpile αναζητά τις παρακάτω υπηρεσίες:

Μηχανές Αναζήτησης και Κατάλογοι:

• Yahoo!	• LookSmart
• Go.com	• Lycos
• Direct Hit	• RealNames
• Sprinks	• Dogpile Directory
• GoTo.com	• Open Directory
• AltaVista	• Google
• AskJeves	• About.com

Usenet:

• AltaVista	• Usenet
• Deja.com	• Fast FTP Search
• FTP	• Deja.com's archival database

News:

- Thunderstone

Auctions:

- GoTo.com

Audio/MP3:

• Astraweb	• AudioGalaxy
• Gigabeat	• MP3Board

Εικόνες:

- Ditto.com

SmallBiz:

- Hypermart

News:

- Thunderstone

- Διάφορες άλλες πηγές (yellow pages, white pages, maps, etc.) παρέχονται από τη συλλογή των *links* στην αρχική σελίδα της Dogpile. Δεν καλύπτονται από την κανονική αναζήτηση της Dogpile

❑ **Meta-Search Χαρακτηριστικά**

Η ίδια η μηχανή αναζήτησης δεν συντάσσει ευρετήριο και δεν ταξινομεί τις ιστοσελίδες. Αντιθέτως, στέλνει ταυτόχρονα το ερώτημα σε πολλαπλές ιστοσελίδες και παραδίδει μια συνδυασμένη σελίδα αποτελεσμάτων των καλύτερων *sites* από κάθε μηχανή αναζήτησης. Όσο ψηλότερα ταξινομείται μια σελίδα σε κάθε μηχανή αναζήτησης, τόσο ψηλότερα ταξινομείται και από τη μηχανή πολλαπλής αναζήτησης.

❑ **Ταξινόμηση των Αποτελεσμάτων**

Η Dogpile έχει μια προκαθορισμένη σειρά αναζήτησης με την οποία επιστρέφονται τα αποτελέσματα. Επομένως, εάν ταξινομηθεί υψηλότερα σε μια από τις πρώτες μηχανές αναζήτησης οι ερωτήσεις της Dogpile, οι πιθανότητες είναι αρκετές για ταξινόμηση των περιοχών υψηλότερα

❑ **Σύνταξη Αναζήτησης**

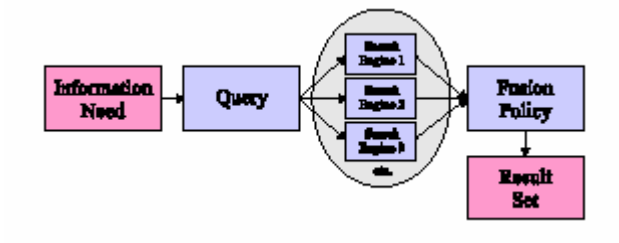
Η Dogpile μεταφράζει τις ερωτήσεις στην πλησιέστερη αποδεκτή μορφή σύνταξης για κάθε μηχανή. Αυτό λειτουργεί καλά κάποιες φορές και κάποιες όχι. Στην AltaVista, σε μια αναζήτηση όπου υπάρχουν πρόσημα του συν (+) και του μείον (-), μετατρέπει το πλην σε NOT και πραγματοποιεί την κύρια αναζήτηση στην AltaVista.

❑ **Εμφάνιση Αποτελεσμάτων**

Η έξοδος καθορίζεται από τη μηχανή αναζήτησης. Η Dogpile εμφανίζει ταυτόχρονα τα αποτελέσματα τριών μηχανών αναζήτησης, εκτός και αν δεν έχουν ανακτήσει 10 έγγραφα μαζί, οπότε θα προστεθούν μιας άλλης μηχανής τα αποτελέσματα. Εμφανίζονται σε δεκάδες τα αποτελέσματα των μηχανών αναζήτησης.

❑ **Αρχιτεκτονική της Dogpile**

Η DogPile, μια χαρακτηριστική μηχανή πολλαπλής αναζήτησης, υποβάλλει το ερώτημα του χρήστη (με τις διάφορες τροποποιήσεις για τη σωστή σύνταξή του) σε ένα σύνολο μηχανών αναζήτησης, και επιστρέφει τα αποτελέσματα με τη σειρά που επιστρέφονται από τις μηχανές αναζήτησης. Αυτό υποδηλώνεται από την αρχιτεκτονική της Dogpile (Σχήμα 5.12.) (Glove, Lawrence, Birmingham, Giles, 1999).

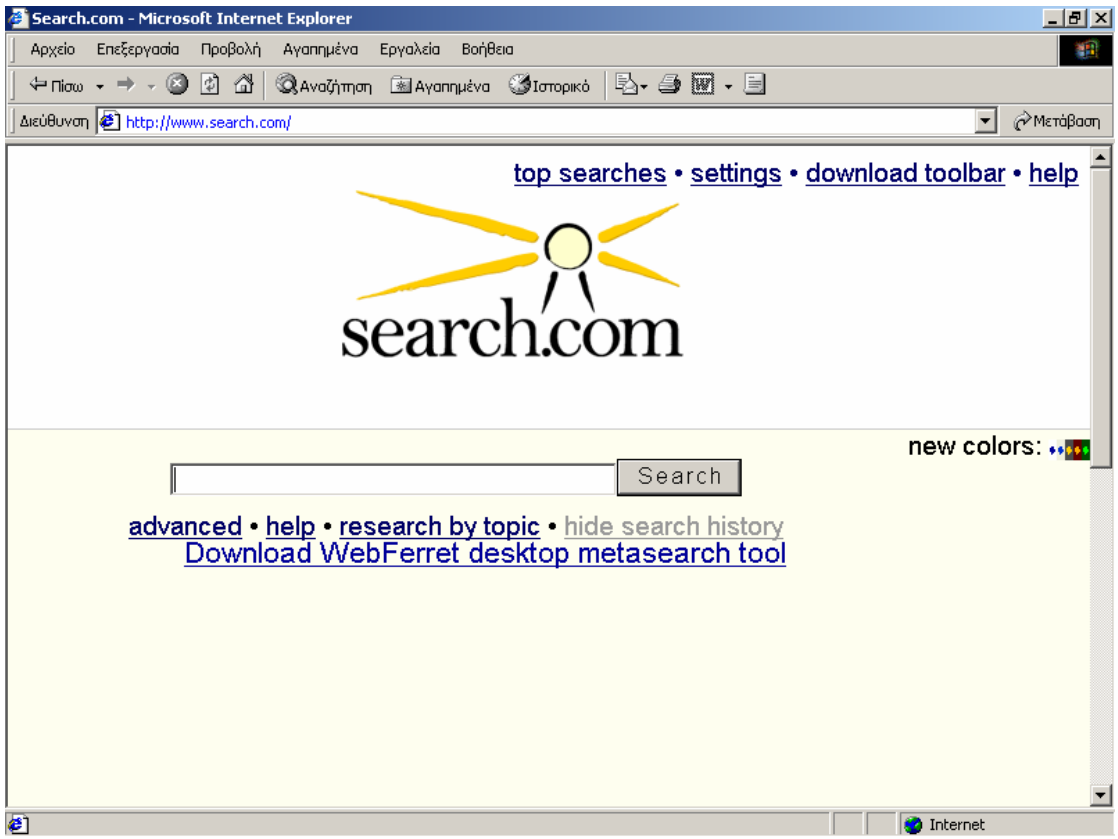


Σχήμα 5.12.:Αρχιτεκτονική της Μηχανής Πολλαπλής Αναζήτησης Dogpile (Glove, Lawrence, Birmingham, Giles, 1999)

5.6.2. Search.com (www.search.com)

Η Search.com είναι μια μηχανή πολλαπλής αναζήτησης που παρέχεται από τη CNET Networks Inc. Είναι μια πηγή πληροφοριών και παροχής εμπορικών υπηρεσιών για τη βιομηχανία της τεχνολογίας. Με τις καθιερωμένες ιστοσελίδες της σε 25 χώρες και το περιεχόμενό τους σε 18 γλώσσες, η CNET Networks συνδέει τους αγοραστές, τους πωλητές και τους προμηθευτές από όλο τον κόσμο μαζί με τους πόρους της

ενημέρωσης, συμπεριλαμβανομένου CNET, ZDNet, mySimon, News.com, Computer Shopper magazine, και CNET Radio, καθώς επίσης και CNET ChannelServices, συμπεριλαμβανομένου CNET DataServices και CNET ChannelOnline. Η CNET προσφέρει όλα τα τεχνολογικά νέα, αποθέματα στα ηλεκτρονικά είδη ευρείας κατανάλωσης, δημοπρασίες, παιχνίδια, free downloads, και βοήθεια, καθώς επίσης και πληροφορίες που επιτρέπουν στις επιχειρήσεις να είναι αποδοτικότερες και να λειτουργούν καλύτερα. Η Search.com συγχρόνως αναζητά τις ιστοσελίδες όπως Yahoo!, Lycos, και Direct Hit για να εμφανίσουν τη σελίδα αποτελεσμάτων της αναζήτησης που να είναι σχετική με τους όρους του χρήστη.



Σχήμα 5.13.: Η Αρχική Σελίδα της Μηχανής Πολλαπλής Αναζήτησης Search.com (www.search.com)

Η Search.com (Σχήμα 5.13) παρέχει πρόσβαση στη συλλογή των 800 και παραπάνω μηχανών αναζήτησης, στους καταλόγους, στα online stores, στις βιβλιοθήκες λογισμικού, στα Usenet archives κ.τ.λ. Πολλά από τα παραπάνω είναι πολύ μικρές συλλογές ή ιδιωτικά εργαλεία αναφοράς. Η πρόσβαση σε ολόκληρη τη συλλογή παρέχεται μέσω της ιστοσελίδας της Search.com. Το βασικό κουτί της αναζήτησης

στην ιστοσελίδα πραγματοποιεί την αναζήτηση με βάση τις μηχανές αναζήτησης, τους καταλόγους και όχι με βάση την πλήρη συλλογή. Το ***Advanced Link*** στην αρχική σελίδα συνδέεται με μια σελίδα που επιτρέπει στο χρήστη να διαλέξει την κατηγορία της πληροφορίας που θα ευρετηριαστεί. Επιτρέπει επίσης την παραμετροποίηση της αναζήτησης σε κάθε κατηγορία, καθορίζοντας παραδείγματος χάριν ποιες μηχανές αναζήτησης θα περιλαμβάνονται στην αναζήτηση. Συνοπτικά η Search.com παρουσιάζει τα παρακάτω χαρακτηριστικά:

Κύριες Μηχανές Αναζήτησης:	3
Συνολικός Αριθμός Μηχανών/ Καταλόγων:	16
Επιλογή της Boolean:	Ναι
Διαγραφή των Διπλών Εγγραφών ή συνδυασμός τους:	Ναι
Μέγιστος Αριθμός Εγγράφων από Κάθε Μηχανή:	10 ή όλα
Μέγιστος Αριθμός Ανακτημένων Εγγράφων:	<50 ή όλα

Η Search.com επιτρέπει την αναζήτηση με τα παρακάτω εργαλεία αναζήτησης (συν τις 700 και παραπάνω ειδικές πηγές).

Οι Μηχανές Αναζήτησης / Κατάλογοι:

• About.com	• LookSmart
• AltaVista	• Lycos
• Britannica.com	• National Directory
• Clearinghouse (Argus)	• NBCi
• Direct Hit	• Open Directory
• Galaxy	• RealNames
• GoTo.com	• Thunderstone
• Hotbot	• Yahoo!

❑ **Σύνταξη Αναζήτησης**

Οι ερωτήσεις μεταβιβάζονται στην μηχανή αναζήτησης όπως τέθηκαν από τον χρήστη. Έτσι, αν ο χρήστης θέσει μια ερώτηση που περιέχει τελεστές Boolean και η μηχανή δεν μπορεί να την διαχειριστεί, τότε τα αποτελέσματα δεν θα είναι τα επιθυμητά.

❑ **Περιορισμός στον Αριθμό των Αρχείων**

Αυτόματα, ταξινομημένα με βάση τη σχετικότητα, τα αποτελέσματα της Search.com θα είναι λιγότερα από 50. Θα χρησιμοποιήσει τις γρηγορότερες μηχανές αναζήτησης και δίνει την επιλογή στο χρήστη να ψάξει όλες τις μηχανές και τους καταλόγους (γενικά επιστρέφει λιγότερα από 100 έγγραφα). Τα πρώτα λίγα αρχεία από κάθε μηχανή εμφανίζονται και υπάρχει ένα *link* στη σελίδα των αποτελεσμάτων της κάθε μηχανής.

❑ **Εμφάνιση Αποτελεσμάτων**

Τα έγγραφα εμφανίζονται με βάση τη σχετικότητά τους (μετρούμενη με τον αριθμό των μηχανών που ανακτήθηκε το αρχείο και την ταξινόμηση που έχει στην κάθε μηχανή).

5.6.3. Mamma (www.mamma.com)

Mamma.com, "Η Μητέρα Όλων των Μηχανών Αναζήτησης", αναγνωρίζεται σήμερα ως μια από τις κορυφαίες μηχανές πολλαπλής αναζήτησης στο διαδίκτυο. Η Mamma.com εμφάνισε την εκρηκτική αύξηση του μεγέθους της το 1996 και εξυπηρετεί πάνω από πέντε εκατομμύρια ιδιώτες χρήστες σε μηνιαία βάση.

Όταν ένας χρήστης εισάγει μια ερώτηση στη Mamma.com, η ισχυρή ιδιόκτητη τεχνολογία της ρωτά ταυτόχρονα 10 μηχανές αναζήτησης, διαμορφώνοντας κατάλληλα τις λέξεις και τη σύνταξη για κάθε μηχανή που τη χρησιμοποιεί. Η Mamma δημιουργεί έπειτα μια εικονική βάση δεδομένων, οργανώνει τα αποτελέσματα ομοιόμορφα και τα παρουσιάζει με βάση τη σχετικότητά τους και την πηγή από την οποία προέκυψαν. Με αυτόν τον τρόπο, η Mamma.com παρέχει στον τελικό χρήστη ένα ιδιαίτερα σχετικό και περιεκτικό σύνολο αποτελεσμάτων αναζήτησης. Η Mamma.com μπορεί να χρησιμοποιηθεί ουσιαστικά σε οποιοδήποτε γεωγραφικό μέρος και γλώσσα. Έχει ένα κεντρικό δίκτυο με πάνω από 100 *servers* σε διάφορες στρατηγικές θέσεις παγκοσμίως και είναι μια από τις γρηγορότερες μηχανές πολλαπλής αναζήτησης.

❑ **Οι Μηχανές Αναζήτησης / Κατάλογοι:**

• AskJeeves	• Mamma Collection
• Business.com	• MSN
• Direct Hit	• Overture
• FindWhat	• Lycos

Σχήμα 5.14.: Η Αρχική Σελίδα της Μηχανής Πολλαπλής Αναζήτησης Mamma.com (www.mamma.com)

5.6.4. SurfWax (www.surfwax.com/servlet/com.surfwax.FrontEnd.home)

Η SurfWax ιδρύθηκε τον Ιούνιο του 2000, παρέχει ***cutting-edge*** αναζήτηση και τεχνολογία ανάκτησης για χρήση με το διαδίκτυο και την επιχείρηση intranets ώστε να παράσχει τα πιο σχετικά αποτελέσματα. Η SurfWax χρησιμοποιεί πιθανώς ένα από τα καλύτερα εργαλεία, την αναγνώριση σημαντικών εναλλακτικών λέξεων, για να βοηθήσει να επικεντρωθεί η αναζήτηση. Η Surfwax επιτρέπει στους χρήστες να περιορίσουν τις αναζητήσεις μέσω του λογισμικού του Word, παγιώνοντας τις διάφορες μηχανές αναζήτησης σε έναν μακρύ κατάλογο αποτελεσμάτων που ανακτώνται από όλες τους, και να παραμετροποιήσει το μοναδικό ύφος και τις προτιμήσεις. Έτσι οι χρήστες να μπορούν να αναθεωρήσουν και να μετρήσουν το περιεχόμενο πριν συνδεθούν με μια σελίδα και να εκμεταλλευτούν τα πρότυπα της αναζήτησής τους τόσο που οι μελλοντικές αναζητήσεις μπορούν να χρησιμοποιήσουν παρόμοιες παραμέτρους.

5.7. Σύγκριση των Μηχανών Πολλαπλής Αναζήτησης

Οι αξιολογήσεις των μηχανών αναζήτησης, όπως αναφέρθηκε στο πρώτο κεφάλαιο είναι δύο ειδών: *αποδεικτικές/ θεωρητικές* ***(testimonials)*** και *πειραματικές* ***(shootouts)*** (Gordon & Pathak, 1999).

Οι *testimonials* συγκρίνουν τις μηχανές αναζήτησης, με βάση την ταχύτητα, την ευκολία χρήσης, τη σχεδίαση και το περιβάλλον τους ή άλλα χαρακτηριστικά, τα οποία είναι εύκολα αντιληπτά στους χρήστες. Ένα άλλο είδος αποδεικτικής αξιολόγησης γίνεται με την έρευνα περισσότερο τεχνικών χαρακτηριστικών των μηχανών αναζήτησης και τη σύγκριση μεταξύ τους σε αυτή τη βάση.

Στις *shootouts*, από την άλλη, χρησιμοποιούνται σε πραγματικό χρόνο διάφορες μηχανές αναζήτησης για την ανάκτηση ιστοσελίδων και συγκρίνεται η αποτελεσματικότητα τους.

Σε αυτή την ενότητα θα γίνει η σύγκριση των μηχανών πολλαπλής αναζήτησης και με τους δύο τρόπους. Βέβαια για κάθε σύγκριση οι μηχανές πολλαπλής αναζήτησης είναι διαφορετικές, έτσι καλύπτεται το μεγαλύτερο φάσμα των μηχανών πολλαπλής αναζήτησης.

5.7.1. Θεωρητική Σύγκριση των Μηχανών Πολλαπλής Αναζήτησης

Οι μηχανές πολλαπλής αναζήτησης επιλέχθηκαν με σκοπό να προσφέρουν όσο το δυνατόν μεγαλύτερη ποικιλία, και για τις μηχανές των σειριακών και των ταυτόχρονων αναζητήσεων. Ο Πίνακας 5.4. συγκεντρώνει τα περισσότερα τεχνικά και μη χαρακτηριστικά των μηχανών πολλαπλής αναζήτησης. Οι μηχανές αναζήτησης που συγκρίνονται είναι οι: Find-it, The Big Hub, Ixquick, Savvy Search, Dogpile, Inference find.

Οι μηχανές σειριακών αναζητήσεων αποδίδουν πολύ καλύτερα. Έγιναν προσπάθειες να ενσωματωθεί η σελίδα τους στις μηχανές αναζήτησης, και έτσι γενικά είναι καλύτερες στην παροχή ενός ευρύτερου φάσματος λειτουργιών, αν και οι οθόνες βοήθειας και οι οδηγίες αναζήτησης ήταν πολύ περιορισμένες. Το σημαντικότερο μειονέκτημα αυτής της προσέγγισης είναι ότι λαμβάνει αρκετό χρόνο για την ολοκλήρωση της αναζήτησης, και η πιο αργή σύνδεση πρόκειται πάντα να είναι και η πιο αργή μηχανή στην οποία αναφέρεται. Εντούτοις, φαίνεται να λειτουργούν πολύ καλά.

Λειτουργία	Find-it	The Big Hub	Ixquick	Savvy Search	Dogpile	Inference find
1. Διαδοχική Αναζήτηση **2.** Ταυτόχρονη Αναζήτηση **3.** Λαμβάνουν υπόψη το προφίλ του χρήστη	1	2	3	3	3	3
Αριθμός των μηχανών που χρησιμοποιούν	10	8	12	10	11	6
Αναζήτηση WWW	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι
Αναζήτηση Usenet	Ναι	Όχι	Όχι	Ναι	Ναι	Όχι
Άνθρωποι	Ναι	Όχι	Όχι	Ναι	Όχι	Ναι
Κάθε Λέξη	Ναι	Ναι	Ναι	Ναι	Ναι	Ποικίλει
Όλες τις Λέξεις	Ποικίλει	Ναι	Ναι	Ναι	Ναι	Ποικίλει
Φράση	Ποικίλει	Όχι	Ναι	Ναι	Ναι	Ναι
Boolean	Ποικίλει	Όχι	Ναι	Όχι	Ποικίλει	Ποικίλει
Αποκοπή	Όχι	Ναι	Ναι	Ναι	Ποικίλει	Ναι
Τελεστές Εγγύτητας	Όχι	Ποικίλει	Ναι	Όχι	Ποικίλει	Ποικίλει
Συγκεκριμένο Τομέα	Όχι	Όχι	Όχι	Όχι	Όχι	Όχι
Τομέα Γεωγραφίας	Όχι	Όχι	Όχι	Όχι	Ναι	Όχι
Συγκεκριμένο Θέμα	Ποικίλει	Ναι	Όχι	Ναι	Ναι	Όχι
Περιορισμός Χρόνου	Όχι	10 secs- 2 minutes	Όχι	Όχι	Όχι	1-30 secs
Περιορισμός των Hits	Όχι	Όχι	Ναι	Όχι	Όχι	Όχι
Επιλογές Εμφάνισης	Όχι	Ναι	Όχι	Ναι	Όχι	Όχι
Συγκέντρωση Αποτελεσμάτων	Όχι	Ναι	Ναι	Όχι	Όχι	Συνηθισμένες Κατηγορίες Αναζήτησης
Βοήθεια	Καμιά	Ναι	Ναι	Ναι	Ναι	Όχι
FAQ	Όχι	Όχι	Ναι	Ναι	Όχι	Όχι
Σχόλια	Γρήγορη Απλή	Άριστη	Γρηγορότερη και η πιο Ευέλικτη	Εντυπωσιακή	Γρήγορη, Εύκολη στη Χρήστη και Αποτελεσματική	Άριστη και η Καλύτερα Προτεινόμενη

Πίνακας 5.4.: Η θεωρητική σύγκριση των Μηχανών Πολλαπλής Αναζήτησης (www.philip.com/msengine.htm)

Οι μηχανές ταυτόχρονης αναζήτησης είναι λιγότερες, αλλά είναι χωρίς καμιά αμφιβολία οι αποτελεσματικότερες. Η Superseek χρησιμοποιεί την προσέγγιση των *πλαισίων* ***(Frames)*** για να ξεπεράσει το πρόβλημα της απόκτησης και της ανάκτησης των αποτελεσμάτων, αλλά αυτή η προσέγγιση απαιτεί να υπάρχει ένας συμβατός

φυλλομετρητής πλαισίων διαθέσιμος, τον οποίο δεν θα τον χρησιμοποιεί ο καθένας.

5.7.2. Πειραματική Σύγκριση των Μηχανών Πολλαπλής Αναζήτησης

Στην αξιολόγηση των μηχανών πολλαπλής αναζήτησης υπάρχουν δύο σημαντικές ερωτήσεις:

1. Πόσο αποτελεσματική είναι μια σχετική μηχανή αναζήτησης στην ανάκτηση μόνο των σχετικών σελίδων;
2. Είναι η σχετική μηχανή αναζήτησης καλή στην εύρεση των περισσότερων ή ενός υψηλού ποσοστού των υπαρχόντων ιστοσελίδων;

Η ακόλουθη αξιολόγηση πραγματοποιήθηκε για να εξετάσει και να συγκρίνει την αποτελεσματικότητα ανάκτησης διάφορων μηχανών πολλαπλής αναζήτησης.

Οι μηχανές αναζήτησης

Οκτώ μηχανές αναζήτησης χρησιμοποιήθηκαν σε αυτή τη σύγκριση. Επιλέχτηκαν επειδή ανήκουν στην ίδια κατηγορία, είναι όλες ελεύθερα διαθέσιμες στο διαδίκτυο και παρουσιάζουν τα αποτελέσματα των άλλων μηχανών αναζήτησης σε ένα ενιαίο ενσωματωμένο έγγραφο. Οι μηχανές που περιλαμβάνονται είναι: C4, Mamma, Metacrawler, Search, Surfwax, Vivvsimo, ProFusion και Ixquick.

Οι ερωτήσεις

Για τη σύγκριση των παραπάνω μηχανών επιλέχτηκαν δέκα ερωτήσεις (Πίνακας 5.5.). Ωστόσο, αν και σε κάθε μηχανή αναζήτησης δόθηκε το ίδιο ερώτημα, η ίδια ερώτηση δεν χρησιμοποιήθηκε (μεταβιβάστηκε) σε όλες τις μηχανές πολλαπλής αναζήτησης. Οι Gordon και Pathak συμβουλεύουν στη μη χρησιμοποίηση της ίδιας ερώτησης κατά την αξιολόγηση των μηχανών αναζήτησης, επειδή μια τέτοια προσέγγιση δεν εκμεταλλεύεται τις αληθινές ικανότητες της κάθε μηχανής αναζήτησης. Επιπλέον, μπορεί να υποστηριχτεί ότι η χρήση της ίδιας ερώτησης περιέχει κινδύνους, επειδή μια μηχανή αναζήτησης μπορεί να χειριστεί αυτή την ερώτηση καλύτερα από την άλλη. Κατά συνέπεια οι ερωτήσεις τέθηκαν για να εκμεταλλευτούν πλήρως τα χαρακτηριστικά των μηχανών αναζήτησης. Στην περίπτωση που κανένα σχετικό έγγραφο δεν ανακτήθηκε, πραγματοποιήθηκε μια νέα αναζήτηση.

Εταιρείες που χρησιμοποιήθηκαν στη σύγκριση των εταιρειών	Ερωτήσεις για τη σύγκριση των μηχανών πολλαπλής αναζήτησης
1. Oracle 2. GE 3. Sun 4. Nike 5. Sunglass Hut 6. Aflac 7. Gap 8. Chevron 9. Shaw Pittman 10. Penguin 11. Dialog 12. Thompson 13. Screaming Media 14. Newsweek 15. Syracuse University 16. Niagara Mohawk 17. Snapple 18. Toyota 19. FannieMae 20. Gannett	1. What is the name of the first Russian astronaut to do a spacewalk? 2. What is Francis Scott Key best known for? 3. What is platinum? 4. What as the name of the famous battle in 1836 between Texas and Mexico? 5. When did the Carolingian Period begin? 6. When was the slinky invented? 7. How much in miles is a ten K run? 8. When was the Triangle Shirtwaist fire? 9. What were the names of the ships used by Columbus? 10. What do river otters eat?
	Αριθμός σχετικών εγγράφων για κάθε ερώτηση
	1. 39 Σχετικά Έγγραφα 2. 41 Σχετικά Έγγραφα 3. 30 Σχετικά Έγγραφα 4. 59 Σχετικά Έγγραφα 5. 34 Σχετικά Έγγραφα 6. 47 Σχετικά Έγγραφα 7. 37 Σχετικά Έγγραφα 8. 80 Σχετικά Έγγραφα 9. 76 Σχετικά Έγγραφα 10. 84 Σχετικά Έγγραφα

Πίνακας 5.5.: Η πειραματική σύγκριση των Μηχανών Πολλαπλής Αναζήτησης (Hawkins, 2001)

Προκειμένου να υπολογιστούν η ακρίβεια και η ανάκληση, τέθηκαν για αξιολόγηση τα πρώτα 50 αποτελέσματα για κάθε ερώτηση, αποφασίζοντας ποιες σελίδες ήταν σχετικές. Εάν οι μηχανές ανάκτησαν λιγότερες από 50 ιστοσελίδες, χρησιμοποιήθηκαν τα διαθέσιμα αποτελέσματα.

Για τη σχετικότητα χρησιμοποιήθηκε η ίδια τεχνική. Ένας (1) πόντος δινόταν εάν η ερώτηση απαντήθηκε, μισός (0,5) πόντος δινόταν εάν η ερώτηση απαντήθηκε μερικώς ή εάν η ανακτημένη ιστοσελίδα είχε μια σύνδεση με μια άλλη ιστοσελίδα με

τις απαιτούμενες πληροφορίες, μηδέν (0) πόντοι δινόταν εάν η σελίδα δεν ήταν σχετική καθόλου.

Ο αριθμός των σχετικών εγγράφων υπολογίστηκε με την πρόσθεση του αριθμού των μοναδικών ιστοσελίδων που ανακτώνται από την κάθε μηχανή πολλαπλής αναζήτησης για τις ίδιες ερωτήσεις.

Το μέτρο της ακρίβειας που χρησιμοποιήθηκε υπολογίστηκε ως εξής: ***ο αριθμός των σχετικών εγγράφων που ανακτήθηκαν διαιρώντας τα με το συνολικό αριθμό των εγγράφων που ανακτήθηκαν***.

Δεδομένου ότι είναι αδύνατο να είναι γνωστό πόσα σχετικά έγγραφα είναι διαθέσιμα στον ιστό για ένα συγκεκριμένο ερώτημα, υπολογίστηκε το μέτρο της ***σχετικής ανάκλησης*** αντί της πραγματικής ανάκλησης, η οποία χρησιμοποιήθηκε όταν ήταν γνωστός ο αριθμός των σχετικών εγγράφων σε ένα σύστημα.

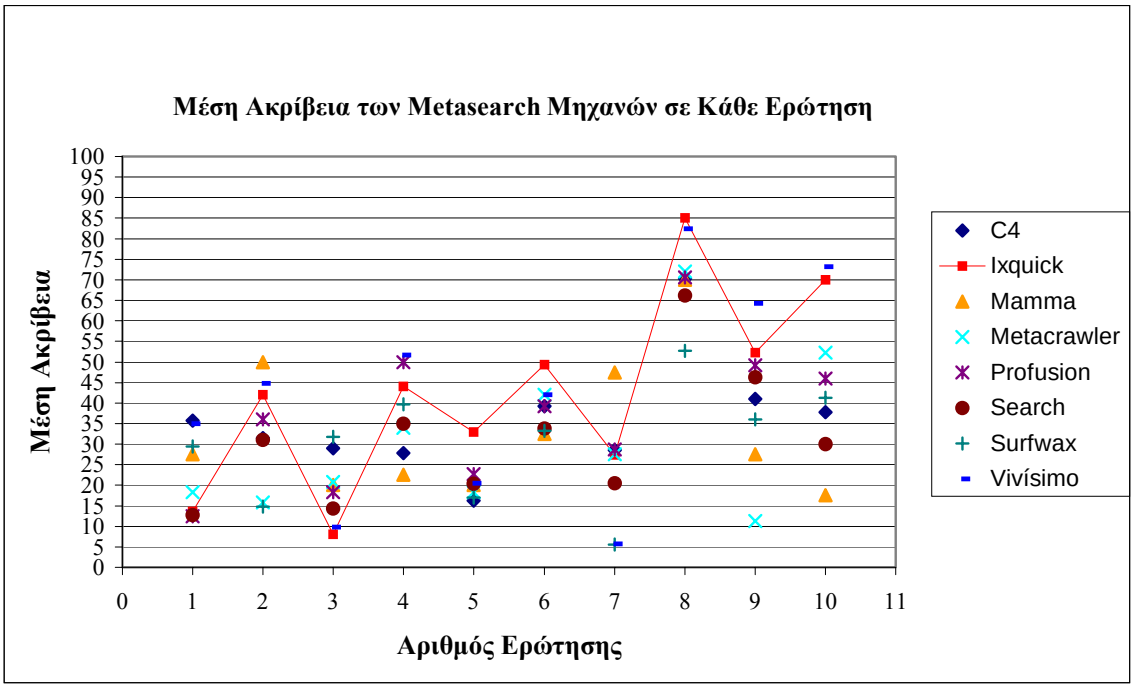
Η σχετική ανάκληση υπολογίστηκε ως ποσοστό των εγγράφων που βρέθηκαν. Παραδείγματος χάριν, η μεγαλύτερη ακρίβεια ανάκλησης υπολογίστηκε στα 50 έγγραφα, 100 έγγραφα, κ.λπ. Οι περισσότερες από τις μηχανές αναζήτησης δεν καταφέρνουν να ανακτήσουν τόσα πολλά έγγραφα. Επιπλέον, μόνο τα πρώτα 50 έγγραφα εξετάστηκαν. Κατά συνέπεια η ανάκληση υπολογίστηκε ως ποσοστό των ανακτημένων εγγράφων. Παραδείγματος χάριν, για την ερώτηση 3 ανακτήθηκαν 30 σχετικά έγγραφα. Κατά συνέπεια η σχετική ανάκληση ήταν τρία έγγραφα στο 10%, έξι έγγραφα στο 20%, εννέα έγγραφα στο 30%, κ.λπ. Αυτό επέτρεψε την μερική αξιολόγηση της ανάκλησης ακόμα και στην περίπτωση που πολύ λίγα έγγραφα ανακτήθηκαν.

Η ακρίβεια υπολογίστηκε σε διαφορετικά διαστήματα- μετά από το πρώτο, το δέκατο, το εικοστό, κ.λ.π. έγγραφο που ανακτάται. Για αυτό το πείραμα, η ακρίβεια υπολογίστηκε σε τέσσερα διαστήματα -πέντε, δέκα, εικοσιπέντε, και πενήντα (50).

Το Σχήμα 5.15 παρουσιάζει τη μέση ακρίβεια για όλες τις μηχανές πολλαπλής αναζήτησης σε κάθε ερώτηση. Η μέση ακρίβεια υπολογίστηκε με την προσθήκη των αποτελεσμάτων ακρίβειας για κάθε ερώτηση στις πέντε, 10, 25, και 50, και διαιρέθηκε με το τέσσερα, τον αριθμό των διαστημάτων υπολογισμού της ακρίβειας.

Όπως και για τις άλλες μηχανές πολλαπλής αναζήτησης, η απόδοση της Ixquick κυμάνθηκε από πολύ μικρή ως πολύ καλή. Για παράδειγμα, στην ερώτηση 1, η Ixquick είχε τη δεύτερη χαμηλότερη μέση ακρίβεια με 13,67%. Στην ερώτηση 3, η Ixquick είχε τη χαμηλότερη μέση ακρίβεια με 8%. Εντούτοις, στις ερωτήσεις 6 και 8, η Ixquick είχε την υψηλότερη μέση ακρίβεια με 49,3% και 85% αντίστοιχα. Είναι ενδιαφέρον, ότι

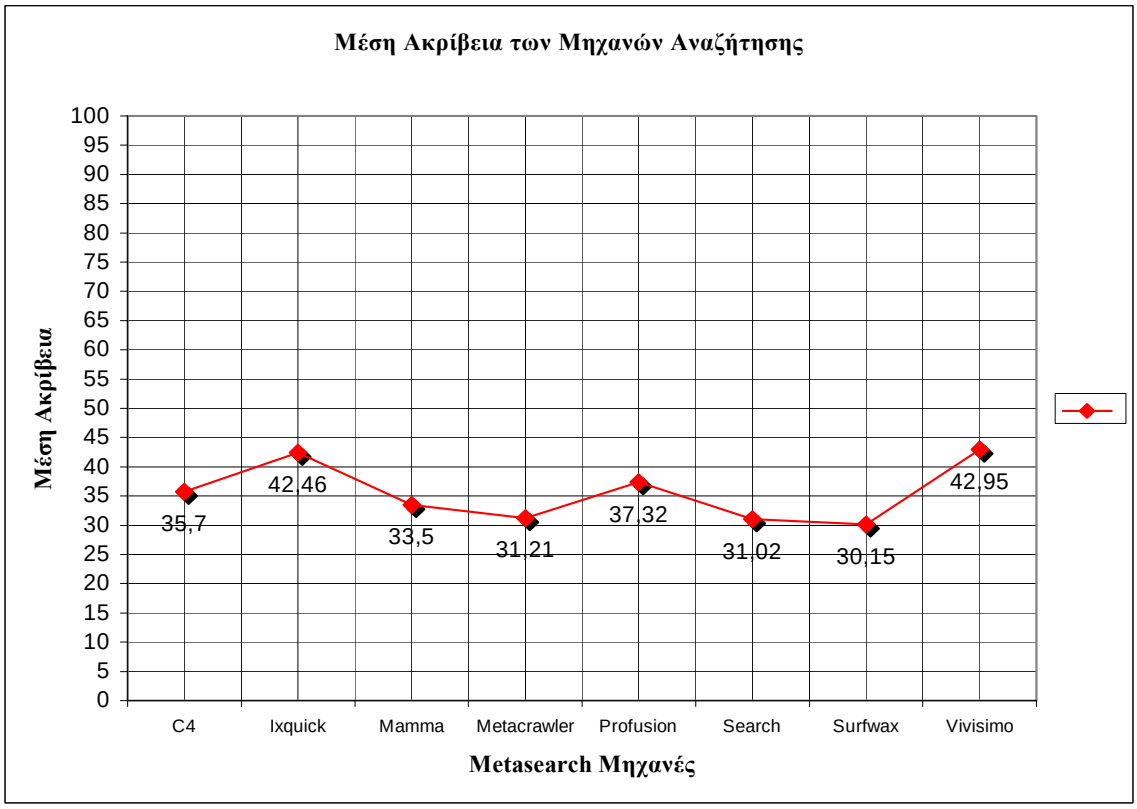
στην ερώτηση 8, όλες οι μηχανές είχαν μεγάλη ακρίβεια.



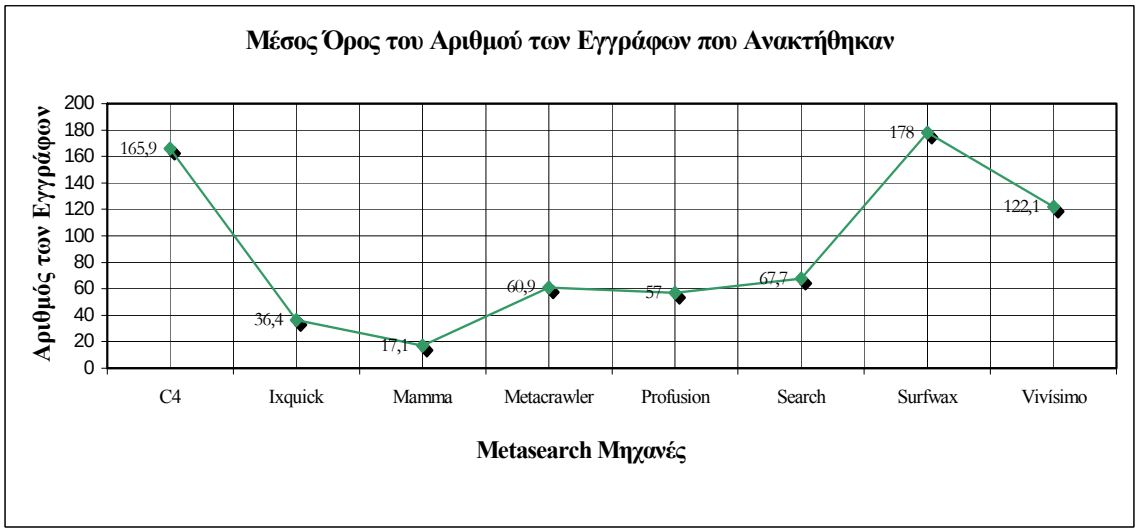
Σχήμα 5.15.: Μέση Ακρίβεια για τις Μηχανές Πολλαπλής Αναζήτησης σε Κάθε Ερώτηση (Hawkins, 2001)

Μια σαφέστερη εικόνα προκύπτει όταν οι μέσοι όροι ακρίβειας εξετάζονται ως προς το σύνολο- το γενικό μέσο όρο όλων των ερωτήσεων και όχι για κάθε ερώτηση χωριστά. Το Σχήμα 5.16. παρουσιάζει αυτά τα αποτελέσματα. Η Vivisimo είχε το μεγαλύτερο μέσο όρο υψηλότερης ακρίβειας με 42,95%. Η ixquick ήταν δεύτερη με μικρή διαφορά με 42,6%. Η Surfwax είχε το χαμηλότερο μέσο όρο με 30,15%. Η Metacrawler είχε το δεύτερο χαμηλότερο μέσο όρο με 31,21.

Η ανάκληση και η σχετική ανάκληση εξαρτώνται από τον αριθμό των εγγράφων που ανακτώνται. Ωστόσο, οι μηχανές πολλαπλής αναζήτησης, όπως η ixquick θυσιάζουν συχνά την ανάκληση για την ακρίβεια. Το Σχήμα 5.17 παρουσιάζει το μέσο όρο του αριθμού των εγγράφων που ανακτήθηκαν από κάθε μηχανή για όλες τις ερωτήσεις. Τα αποτελέσματα κυμάνθηκαν από το 17,1 έως το 17,8. Η Ixquick είχε το δεύτερο χαμηλότερο μέσο όρο σε 36,4 έγγραφα ανά ερώτηση.

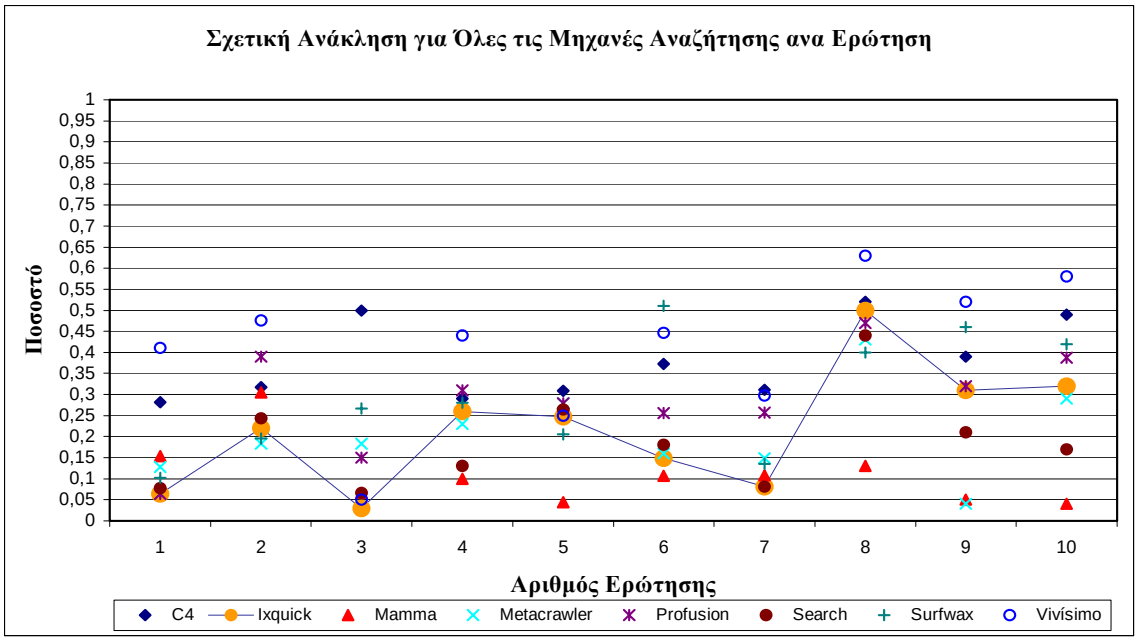


Σχήμα 5.16.: Μέσοι Όροι Ακρίβειας για κάθε Μηχανή Πολλαπλής Αναζήτησης (Hawkins, 2001)



Σχήμα 5.17.: Μέσος Όρος του Αριθμού των Εγγράφων που Ανακτήθηκαν (Hawkins, 2001)

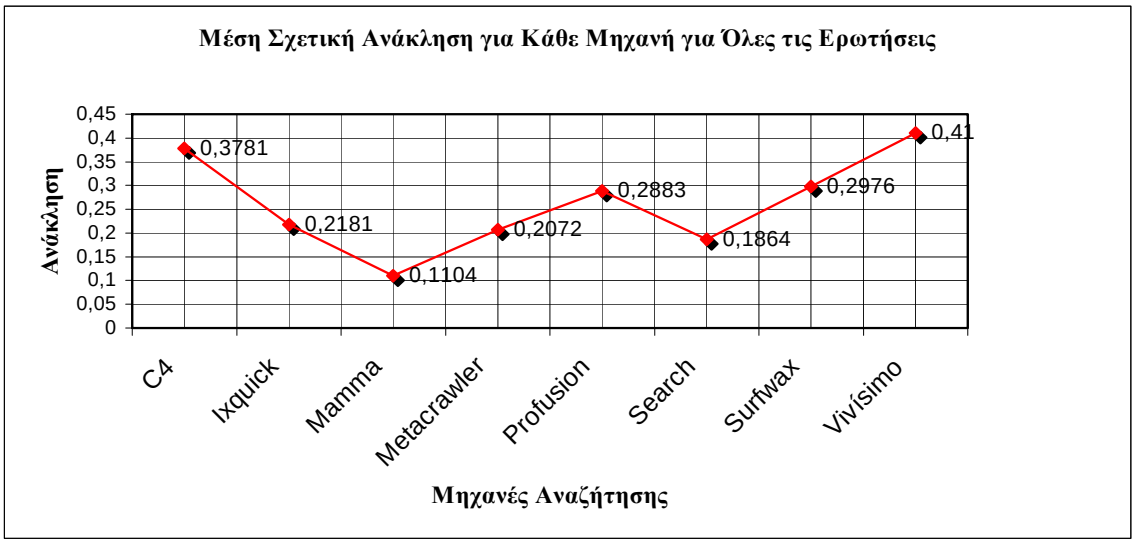
Το Σχήμα 5.18. παρουσιάζει τη σχετική ανάκληση για όλες τις μηχανές αναζήτησης για κάθε ερώτηση.



Σχήμα 5.18.: Σχετική Ανάκληση για τις Μηχανές Αναζήτησης σε κάθε ερώτηση (Hawkins, 2001)

Τα αποτελέσματα δείχνουν ότι η Ixquick δεν επιτυγχάνει μια ιδιαίτερα υψηλή σχετική ανάκληση σε κάποια ερώτηση. Στην ερώτηση 10, η Vivisimo πέτυχε σχετική ανάκληση με 58%, το υψηλότερο ποσοστό μεταξύ όλων των μηχανών αναζήτησης σε κάθε ερώτηση. Στην ερώτηση 3, η Ixquick είχε σχετική ανάκληση 3%, την χαμηλότερη ανάκληση μεταξύ όλων των μηχανών σε κάθε ερώτηση. Αντιθέτως, η Ixquick επιτυγχάνει μια αρκετά υψηλή σχετική ανάκληση με 50% στην ερώτηση 8. Ωστόσο, η Vivvsimo και η C4 είχαν καλύτερη απόδοση, με σχετικές ανακλήσεις 63% και 52% αντίστοιχα. Πάλι, είναι ενδιαφέρον να σημειωθεί ότι στην ερώτηση 8, όλες οι μηχανές πέτυχαν υψηλή σχετική ανάκληση.

Το Σχήμα 5.19 παρουσιάζει τη μέση σχετική ανάκληση για κάθε μηχανή για όλες τις ερωτήσεις. Η Vivisimo έχει τον υψηλότερο σχετικό μέσο όρο ανάκλησης με 41%. Η C4 είναι δεύτερη με 37,8%, και η Surfwax είναι τρίτη με 29,7%. Η Mamma είχε το χαμηλότερο με 11%, και η Ixquick είχε τον τρίτο χαμηλότερο μέσο όρο με 21,8%.



Σχήμα 5.19.: Μέση Σχετική Ανάκληση για Κάθε Μηχανή για Όλες τις Ερωτήσεις (Hawkins, 2001)

ΚΕΦΑΛΑΙΟ 6

ΘΕΜΑΤΙΚΟΙ ΚΑΤΑΛΟΓΟΙ-ΘΕΜΑΤΙΚΑ ΕΥΡΕΤΗΡΙΑ

Οι θεματικοί κατάλογοι είναι μια μέθοδος βελτίωσης της αποδοτικότητας της αναζήτησης ενός μεγάλου χώρου πληροφοριών με τον χωρισμό του σε ευδιάκριτες ευρείες κατηγορίες, οι οποίες όμως έχουν σημασία για το χρήστη. Η κατηγοριοποίηση και η ταξινόμηση των θεμάτων εφαρμόζονται συχνά στις βιβλιοθήκες και στην επιστήμη των πληροφοριών (π.χ. ***INSPEC*** -βάση δεδομένων για το πεδίο του computer engineering, ***ERIC*** -βάση δεδομένων για την κοινωνιολογία κ.λ.π.). Ο χωρισμός των πληροφοριών σε θεματικές κατηγορίες δημιουργεί μικρότερες βάσεις δεδομένων που μπορούν να εξερευνηθούν- περιηγηθούν αποτελεσματικότερα. Επιπλέον, οι ευρείες αυτές κατηγορίες επιτρέπουν στους εξερευνητές την *περιήγηση των θεματικών αυτών καταλόγων* ***(Directory Browsing)***. Η περιήγηση των θεματικών καταλόγων στο διαδίκτυο είναι μια καθοδηγούμενη από το χρήστη αναζήτηση πληροφορίας, που μοιάζει με τους παλαιότερους δημοφιλείς *servers* πληροφοριών, τα ***Gophers***. Οι χρήστες του *Gopher* συνδέονταν μέσω αυτού με τις περιοχές ενδιαφέροντός τους στο διαδίκτυο και αναζητούσαν τους διαθέσιμους καταλόγους αυτών των περιοχών. Εάν ένας θεματικός κατάλογος φαινόταν να είναι ενδιαφέρων, θα μπορούσε να εξερευνηθεί εκτενέστερα. Αυτό είναι ένα παράδειγμα ενός ιεραρχικά οργανωμένου καταλόγου.

Όπως αναφέρθηκε στην ενότητα 3.5 της παρούσας εργασίας, τα *Θεματικά Ευρετήρια* ***(Subject Catalogues- Subject Directories)*** είναι τα εργαλεία εκείνα που παρέχουν μια ιεραρχικά οργανωμένη ταξινόμηση της ανθρώπινης γνώσης σε

κατηγορίες, προκειμένου η αναζήτηση στα περιεχόμενα τους να γίνεται με βάση το θέμα. Συχνά ονομάζονται και *δέντρα* ***(trees)*** λόγω της δομής τους, η οποία ξεκινά από ένα βασικό κορμό-κατηγορία και σιγά-σιγά διακλαδώνεται σε υποκατηγορίες.

Το Yahoo ήταν η πρώτη διαδικτυακή υπηρεσία αναζήτησης/ περιήγησης προσφέροντας έναν θεματικό κατάλογο που ταξινομεί το χώρο των πληροφοριών του διαδικτύου σε κάποιες κατηγορίες που έχουν νόημα για το χρήστη (π.χ., επιστήμη, ψυχαγωγία, επιχειρήσεις, κ.λ.π.). Πιο πρόσφατα, η Lycos, ένας από τους μεγαλύτερους φορείς παροχής υπηρεσιών στο διαδίκτυο, πρόσθεσε μια υπηρεσία αναζήτησης θεματικού καταλόγου στην παραδοσιακή της αναζήτηση με βάση τις λέξεις κλειδιά. Χωρίζοντας το πεδίο των πληροφοριών με βάση το περιεχόμενο ή τις θεματικές κατηγορίες μπορεί να βελτιωθεί η αναζήτηση καθιστώντας τη παράλληλα αποδοτικότερη. Αυτός ο τρόπος αναζήτησης βέβαια, παρουσιάζει και κάποια προβλήματα. Τα πιο κοινά είναι:

- Οι κατηγορίες είναι περιορισμένες ως προς την ενημέρωσή τους, και
- Η διαδικασία της δημιουργίας των κατηγοριών και της σύνδεσης των αρχικών σελίδων σε αυτές είναι χειρωνακτική, αργή, και δυσκίνητη. Ίσως αυτή η μέθοδος μπορεί να βελτιωθεί με την ενσωμάτωση ενός *ευφυούς, αυτόματου αλγορίθμου κατηγοριοποίησης* ***(intelligent automatic categorization algorithm)***, ως τμήμα της διαδικασίας δημιουργίας του θεματικού καταλόγου.

Όταν αρχίζει μια αναζήτηση με βάση τις λέξεις κλειδιά του περιεχομένου ενός θεματικού καταλόγου, ο κατάλογος προσπαθεί να *ταιριάξει* ***(matches)*** τις λέξεις κλειδιά και τις φράσεις με εκείνες στις γραπτές περιγραφές του. Οι θεματικοί κατάλογοι περιέχουν πολλές κατηγορίες. Υπάρχουν *γενικοί κατάλογοι* ***(general directories)***, *ακαδημαϊκοί κατάλογοι* ***(academic directories)***, *εμπορικοί κατάλογοι* ***(commercial directories)***, *πύλες* ***(portals)*** και ***vortals***. Οι πύλες είναι κατάλογοι που έχουν δημιουργηθεί ή ασχολούνται με εμπορικά ενδιαφέροντα και έχουν μετατραπεί να λειτουργούν ως πύλες του ιστού. Αυτές οι *portal* περιοχές όχι μόνο συνδέονται με τις δημοφιλείς θεματικές κατηγορίες, προσφέρουν επίσης πρόσθετες υπηρεσίες όπως e-mail, τρέχοντα νέα, πληροφορίες ταξιδιών και χάρτες. Οι *Vortals, ή κάθετες πύλες* ***(Vertical Portals)***, είναι συγκεκριμένων θεμάτων κατάλογοι, εν αντιθέσει με το ευρύτερο, το γενικότερο των θεμάτων και των άλλων συνδέσεων που βρίσκονται συνήθως στις πύλες.

Σήμερα, η διαχωριστική γραμμή μεταξύ των θεματικών καταλόγων και των

μηχανών αναζήτησης δεν είναι ευδιάκριτη. Οι περισσότεροι θεματικοί κατάλογοι συνεργάζονται με τις μηχανές αναζήτησης για την αναζήτηση των βάσεων δεδομένων τους και να για την αναζήτηση του ιστού για πρόσθετες πηγές, ενώ οι μηχανές αναζήτησης αποκτούν τους θεματικούς καταλόγους ή δημιουργούν δικούς τους.

Όπως οι κίτρινες σελίδες ενός τηλεφωνικού καταλόγου, οι θεματικοί κατάλογοι είναι καλύτεροι για την περιήγηση και για τις αναζητήσεις θεμάτων γενικότερης φύσης. Είναι καλές πηγές για πληροφορίες που σχετίζονται με δημοφιλή θέματα, οργανισμούς, εμπορικά *sites* και προϊόντα. Όταν κάποιος επιθυμεί να δει το είδος των πληροφοριών που είναι διαθέσιμες στον ιστό για ένα ιδιαίτερο πεδίο ή για έναν τομέα ενδιαφέροντος, πηγαίνει σε έναν κατάλογο και κάνει μια περιήγηση μέσω των θεματικών κατηγοριών. Παραδείγματα των θεματικών ευρετηρίων και των *Portals* βρίσκονται παρακάτω:

- **Θεματικοί Κατάλογοι (Subject Directories)**
  - Beaucoup
  - CompletePlanet
  - LookSmart
  - Lycos
  - Open Directory Project
  - Yahoo!

- **Portals**
  - Excite
  - MSN
  - Netscape
  - Yahoo!

6.1. Yahoo! (www.yahoo.com)

6.1.1. Εισαγωγή

Το θεματικό ευρετήριο Yahoo! *(**Y**et **A**nother **H**ierarchical **O**fficious **O**racle)* δείχνει να είναι ο αδιαφιλονίκητος ηγέτης του χώρου, αφού καταφέρνει εδώ και πολλά χρόνια να παραμένει πρώτο στις προτιμήσεις των χρηστών (Μακριδάκης, 2003). Το Yahoo!, εκτός της ιεραρχικής δομής που διαθέτει, προσφέρει εξειδικευμένες αναζητήσεις για ειδήσεις, μετοχές, εταιρείες, φυσικά πρόσωπα, μικρές αγγελίες και μια σειρά θεμάτων που διευκολύνουν το χρήστη να αποκτήσει ένα σημείο αναφοράς από την πρώτη κιόλας στιγμή.

Υπάρχουν ευρωπαϊκές σελίδες για το Yahoo! στη Δανία, στη Γαλλία, στη Γερμανία, στην Ιταλία, στη Νορβηγία, στην Ισπανία, στη Σουηδία και στην Αγγλία. Γενικά το *site* του Yahoo! Παρέχει τις υπηρεσίες του σε 274 εκατομμύρια ιδιώτες κάθε μήνα (www.submitcorner.com/Guide/SE/yahoo.shtml, Μάρτιος, 2004).

Ξεκινώντας στα τέλη του 1994, το Yahoo! αποτέλεσε ίσως το πρώτο θεματικό ευρετήριο με τόσο μεγάλη απήχηση που διαρκεί μέχρι σήμερα. Το μεγάλο πλεονέκτημα του Yahoo! είναι ο ανθρώπινος παράγοντας που βρίσκεται πίσω από την ευρετηρίαση όλων των *sites* τα οποία καταγράφει. Περισσότεροι από εκατό συντάκτες συγκεντρώνουν και ταξινομούν τα περιεχόμενα του διαδικτύου, προσφέροντας έτσι στους χρήστες του Yahoo! μια θεματική ταξινόμηση που ανταποκρίνεται σε μεγάλο βαθμό στην ταξινόμηση που εφαρμόζουν παγκοσμίως αποδεκτά ταξινομικά συστήματα για την ανθρώπινη γνώση. Είναι προφανές ότι σε τέτοιου είδους ευρετήρια υπάρχει η δυνατότητα για την εισαγωγή και νέων κατηγοριών και υποκατηγοριών, προκειμένου να καλυφθούν νέοι θεματικοί τομείς χωρίς να αλλοιώνεται η αρχική δομή.

Στο Yahoo! πιστώνονται μια σειρά από άλλα χαρακτηριστικά, τα οποία σιγά- σιγά καθιερώθηκαν και στα υπόλοιπα εργαλεία και τις μηχανές αναζήτησης και πλέον θεωρούνται απαραίτητα συμπληρώματα. Τέτοιες καινοτομίες ήταν η *παραμετροποίηση **(customization)*** του Yahoo! προκειμένου να καλύπτει τις προσωπικές ανάγκες κάθε χρήστη, εμφανίζοντας, για παράδειγμα, μόνο τις ειδήσεις που εκείνος είχε επιλέξει να εμφανίζονται, ενώ ταυτόχρονα εμφάνιζε στην αρχική οθόνη του χρήστη μια σειρά πληροφοριών που ο ίδιος είχε ζητήσει να λαμβάνει σε τακτική βάση. Το ***MyYahoo!***, δηλαδή η παραμετροποίηση της μηχανής-ευρετηρίου σύμφωνα με τις ανάγκες κάθε χρήστη αποτέλεσε πραγματική επανάσταση στο χώρο και σύντομα καθιερώθηκε από το σύνολο των άλλων εργαλείων-μηχανών αναζήτησης, προκειμένου να προσφέρουν όσο

το δυνατόν περισσότερες υπηρεσίες και διευκολύνσεις στους εν δυνάμει χρήστες τους.

Στην κεντρική σελίδα του Yahoo! εκτός του βασικού *query box* ο χρήστης θα συναντήσει τις γενικές θεματικές κατηγορίες που έχει στη διάθεσή του προκειμένου να ξεκινήσει την αναζήτηση. Το θεματικό ευρετήριο του Yahoo!, που θεωρείται από τα πιο πλήρη, συμπληρώνεται από μια σειρά επιλογών που σχετίζονται με την αναζήτηση μετοχών, φυσικών προσώπων, χαρτών, εταιρειών, ειδήσεων, μικρών αγγελιών, κλπ. Στο Σχήμα 6.1. παρουσιάζεται η κεντρική σελίδα του θεματικού καταλόγου Yahoo!. Στο σχήμα αυτό δεν φαίνονται όλες οι θεματικές κατηγορίες του.

Σχήμα 6.1.: Η Κεντρική Σελίδα του Θεματικού Καταλόγου Yahoo!

Ο κατάλογος του Yahoo απαριθμεί πολλές χιλιάδες πόρους *(sites)* ιεραρχημένα σε κατηγορίες θεμάτων. Στην κορυφή της ιεραρχίας βρίσκονται δεκατέσσερις (14) ευρείες κατηγορίες οι οποίες είναι:

Business&Economy	***Regional***	***Computer&Internet***
Society&Culture	***News&Media***	***Education***
Entertainment	***Arts&Humanities***	***Recreation&Sports***

Science	***Health***	***Social Science***
Goverment	***Reference***	

Κάτω από κάθε μια από αυτές τις κατηγορίες υπάρχει μια ιεραρχία από υποκαταλόγους, και μέσα σε αυτούς είναι καταχωρημένοι οι πόροι, ταξινομημένοι κάτω από το κάθε θέμα. Οι *υπερ-συνδέσεις των καταχωρήσεων* ***(hyperlinked entries)*** μπορεί να αποτελούνται μόνο από τον τίτλο, ή από τον τίτλο και μια μικρή περιγραφή.

Στην κεντρική σελίδα υπάρχει ακόμη η επιλογή της παραμετροποίησης (My Yahoo!), η οποία, όπως αναφέραμε, αποτέλεσε καινοτομία στο χώρο την εποχή που εμφανίστηκε.

6.1.2. Ανάκτηση Πληροφορίας

Οι δημιουργοί των *sites* πρέπει να παρέχουν πληροφορίες σχετικά με τους πόρους όταν υποβάλλουν υλικό προς καταχώρηση στο Yahoo!. Κατά την αναζήτηση στο Yahoo!, το εργαλείο αναζητά τους όρους της αναζήτησης στον τίτλο του *Website* και στην περιγραφή του, καθώς και στις επικεφαλίδες των κατηγοριών. Είναι η ίδια προσέγγιση που χρησιμοποιείται και στις *gateways* και *virtual libraries*, αλλά όχι και στις μηχανές αναζήτησης (αυτές αναζητούν τους όρους σε κάθε λέξη κάθε σελίδας που βρίσκεται στη βάση δεδομένων τους). Το Yahoo! εμφανίζει τα αποτελέσματα βάσει της συχνότητάς τους, η οποία αξιολογείται από τον αριθμό των φορών που οι όροι της αναζήτησης εμφανίζονταν στο *site* ή στην περιγραφή της κατηγορίας. Γενικά τα *sites* ή οι κατηγορίες με τα περισσότερα *matches* (ταιριάσματα) κατατάσσονται υψηλότερα. Τα *sites* και οι κατηγορίες με *matches* που βρίσκονται στον τίτλο κατατάσσονται υψηλότερα από εκείνα και εκείνες με *matches* που βρίσκονται στην περιγραφή.

6.1.3. Ευρετηρίαση και Ταξινόμηση (Indexing and Ranking)

Μια μηχανή αναζήτησης επισκέπτεται ένα *web site*, τοποθετεί σε λίστα κάθε σελίδα από ένα *site*, τα οποία είναι ανεξάρτητα μεταξύ τους. Η ταξινόμηση εξαρτάται από το περιεχόμενο της σελίδας. Αντιθέτως, το Yahoo! δέχεται υποβολές που περιγράφουν ολόκληρο ένα *web site* και όχι κάθε σελίδα χωριστά. Μια καρτέλα (φόρμα) συμπληρώνεται με πληροφορίες για το *site*. Αυτό επιθεωρείται από τον συντάκτη του Yahoo! και, εάν εγκρίνεται, προστίθεται στον οδηγό. Εάν ένα *site* δεν έχει υποβληθεί, οι πιθανές αλλαγές που μπορεί να υπάρξουν στη σελίδα δεν θα

σημειωθούν.

Επίσης, το πόσο σημαντικά οι μηχανές αναζήτησης θα θεωρήσουν τα *meta tags*, τους τίτλους των σελίδων, και το κύριο μέρος των σελίδων δεν παίζει σημαντικό ρόλο στο Yahoo!. Αυτό συμβαίνει γιατί το Yahoo δεν χρησιμοποιεί τα *robots*, αλλά η ευρετηρίαση του Yahoo γίνεται από ανθρώπους. Έτσι τα *META tags*, τα *ALT image tags* ή τα σχόλια σε *HTML* δεν θα βοηθήσουν στην καλύτερη ταξινόμηση της σελίδας. Η καλύτερη περίπτωση καταχώρησης μιας σελίδας είναι να περιγραφεί όσο το δυνατόν ακριβέστερα. Η τελική απόφαση θα παρθεί από τον συντάκτη που αναθεωρεί την ιστοσελίδα, έτσι το περιεχόμενό της θα πρέπει να τον εντυπωσιάσει ώστε να είναι άξιο προσθήκης στη βάση δεδομένων του Yahoo. Η Google τώρα παρέχει δεύτερες λίστες για το Yahoo! στην περίπτωση που δεν επιστραφεί κανένα αποτέλεσμα από το ευρετήριο του Yahoo!. Όλα περιστρέφονται γύρω από τις πληροφορίες και τη μορφή που υποβάλλονται, αν και είναι σημαντικό ότι η ιστοσελίδα είναι καλής ποιότητας και το περιεχόμενο απεικονίζει την περιγραφή που στέλνεται στο Yahoo!.

Όταν οι χρήστες εισάγουν τους όρους αναζήτησης, το Yahoo! ελέγχει τον κατάλογο των ιστοσελίδων του και επιστρέφει τις λίστες με αυτήν τη σειρά:

- Οι κατηγορίες Yahoo! που περιέχουν τους όρους
- Τα sites με τους όρους στους τίτλους τους
- Τα sites με τους όρους στις περιγραφές τους.

Οι αναζητήσεις του Yahoo! διευρύνονται αυτόματα με σύνδεση με τη μηχανή αναζήτησης Google. Επίσης, αν δεν υπάρχουν sites ή κατηγορίες σε απόκριση κάποιας αναζήτησης στο Yahoo!, η αναζήτηση διεξάγεται αυτόματα από την Google.

Γενικά, υπάρχουν μερικοί παράγοντες που επηρεάζουν τα αποτελέσματα του Yahoo!: ο τίτλος, το όνομα των πεδίων, η περιγραφή. Οι λέξεις κλειδιά που βρίσκονται στον τίτλο θα βοηθήσουν στην υψηλότερη ταξινόμηση από τις λέξεις κλειδιά που βρίσκονται μόνο στην περιγραφή και ομοίως οι λέξεις κλειδιά μέσα στο όνομα των πεδίων θα ταξινομηθούν υψηλότερα.

6.1.4. Η Αναζήτηση με Βάση το Yahoo!

Η δυνατότητα της γενικής αναζήτησης (είναι διαθέσιμη στην αρχική σελίδα) προσφέρει αναζητήσεις στις *θεματικές κατηγορίες* ***(subject categories)*** του Yahoo!, στα ***Web sites*** συν μια *γενική αναζήτηση του ιστού* ***(general web search)***. Η αναζήτηση

παρέχει μια περίληψη των καλύτερων *αντιστοιχιών* ***(matches)*** από κάθε περιοχή. Τα αποτελέσματα ταξινομούνται με βάση τη σχετικότητα που υπολογίζεται από τον αριθμό των λέξεων αναζήτησης, την ακριβή σειρά των λέξεων, και της θέσης των λέξεων αναζήτησης στο έγγραφο (η θέση στον τίτλο την ταξινομεί υψηλότερα). Το πρώτο σύνολο των αποτελεσμάτων προέρχεται από τις κατηγορίες του Yahoo!. Μια κατηγορία μπορεί να περιέχει εκατοντάδες ή ακόμα και χιλιάδες σχετικές ιστοσελίδες. Εάν καμία κατηγορία δεν βρέθηκε, το Yahoo! παρουσιάζει το επόμενο σύνολο των αποτελεσμάτων *(Web sites)*. Αυτές είναι περιοχές που βρίσκονται στους καταλόγους του Yahoo! που καταχωρούνται με τις κατηγορίες που τις περιέχουν. Εάν κανένα *Web site* του Yahoo! δεν ανακτηθεί πηγαίνει στο επόμενο σύνολο των αποτελεσμάτων *(Web pages)* που παράγονται από μια ολοκληρωμένη αναζήτηση με βάση το κείμενο του ιστού από τη μηχανή αναζήτησης Google. Μια προχωρημένη επιλογή αναζήτησης είναι επίσης διαθέσιμη που προσφέρει πρόσθετες δυνατότητες αναζήτησης, όπως η έρευνα των Usenet News (DejaNews) ή το Yahoo!. Οι αναζητήσεις μπορούν να περιοριστούν στην τελευταία ημέρα, εβδομάδα, μήνα, έτος ή στα τρία έτη. Οι όροι αναζήτησης μπορούν να ενωθούν με AND ή OR, και να θεωρηθούν ως ***substring*** ή ως πλήρη λέξη.

6.1.5. Υποβολή των Σελίδων στο Yahoo!

Όταν υποβάλλεται μια σελίδα στο Yahoo!, υπάρχει η δυνατότητα επιλογής δύο κατηγοριών: Μια *αρχική κατηγορία* ***(primary category)*** και μια *πρόσθετη σχετική κατηγορία* ***(additional related category)*** με βάση την αρχική επιλογή. Ο καλύτερος τρόπος να καθοριστεί σε ποια κατηγορία ανήκει η σελίδα που καταχωρήθηκε είναι να πραγματοποιηθεί μια αναζήτηση με βάση μια λέξη κλειδί που περιγράφει καλύτερα το περιεχόμενο της ιστοσελίδας. Μόλις γίνει αυτό, ελέγχεται αν ταιριάζει η προτεινόμενη κατηγορία με την επιστρεφόμενη και εξερευνώνται οι διαθέσιμες λίστες κάθε κατηγορίας. Οι δευτεροβάθμιες κατηγορίες είναι καταλληλότερες για τις περιφερειακές λίστες.

Οι συντάκτες του Yahoo! περικόπτουν τον τίτλο αφήνοντας μόνο το όνομα της επιχείρησης. Ο καλύτερος τρόπος αποφυγής του παραπάνω είναι η προσπάθεια ενσωμάτωσης λέξεων κλειδιών στο όνομα της επιχείρησης. Πρακτικά το Yahoo! μικραίνει τον τίτλο όσο το δυνατό περισσότερο.

Παρόμοια με τους τίτλους, οι συντάκτες του Yahoo! θέλουν να ελαχιστοποιήσουν και την περιγραφή της σελίδας. Το Yahoo! θέτει ως όριο στις περιγραφές τις 25 λέξεις,

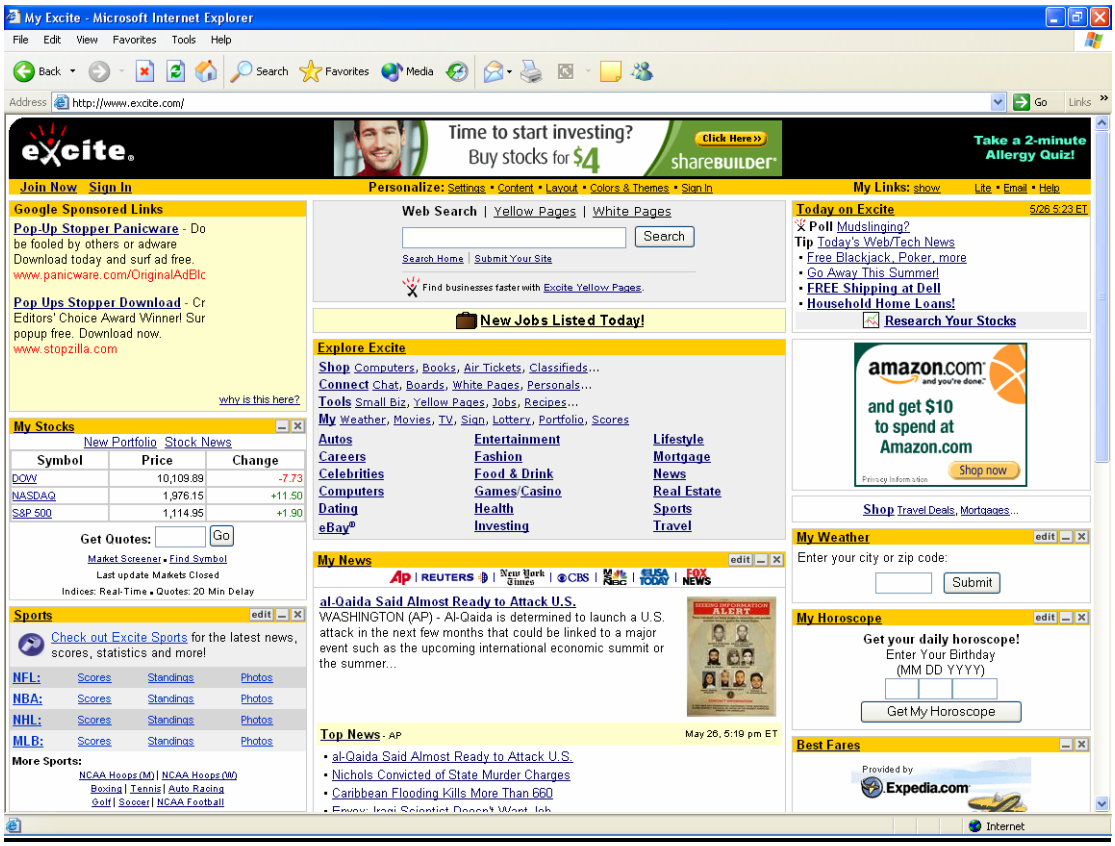
αλλά στην πράξη αυτό είναι πολύ μικρότερο. Πρέπει να αποφεύγονται τα κόμματα και η περιγραφή να γραφεί σε μια ενιαία πρόταση. Εάν χωριστεί σε δύο ή περισσότερες προτάσεις πιθανότατα το Yahoo! θα αφαιρέσει αυτόματα τις τελευταίες προτάσεις

Για να καταχωρηθεί στο Yahoo! μια σελίδα μπορεί να χρειαστεί μεταξύ 4-8 εβδομάδων. Όλες οι καταχωρήσεις πρέπει να γίνουν με το χέρι και μέσω της κατάλληλης κατηγορίας. Συχνά, διαρκεί πολλές εβδομάδες εάν όχι μήνες η καταχώρηση μιας ιστοσελίδας στο Yahoo!. Ένας τρόπος να επιταχυνθεί αυτή η διαδικασία είναι η υποβολή της σελίδας στο Yahoo! του Καναδά (http://www.yahoo.ca). Το Yahoo! του Καναδά έχει σημαντικά λιγότερες υποβολές καταχώρησης και επομένως θα επισπευσθεί η παραπάνω διαδικασία. Επιπλέον, εάν η περιοχή δεν είναι καναδική, συχνά θα διαβιβάσουν την περιοχή στον αρμόδιο συντάκτη της ίδιας κατηγορίας στον αμερικανικό κατάλογο. Αυτό έχει το πλεονέκτημα της αποφυγής της αναμονής και της πολύ γρηγορότερης καταχώρησης. Εάν η περιοχή είναι καναδική, υπάρχει το πλεονέκτημα της καταχώρησης και στον καναδικό και στον αμερικάνικο κατάλογο αυτόματα.

6.2. Excite (www.excite.com)

6.2.1. Εισαγωγή

Η Excite ισχυρίζεται ότι έχει ευρετήριο περίπου 125 εκατομμύρια ιστοσελίδες, το οποίο είναι περίπου ίσο με το 15,6% του συνολικού ιστού. Στη βασική οθόνη της Excite (Σχήμα 6.2.) κυριαρχεί το *πεδίο αναζήτησης* ***(query box)*** στο μέσο της σελίδας, ενώ αριστερά και δεξιά βρίσκονται οι άλλες επιλογές που παρέχει η μηχανή. Η δυνατότητα της αναζήτησης με λέξεις κλειδιά συμπληρώνεται από την επιλογή της γλώσσας στην οποία θα αναζητηθούν και ανακτηθούν τα αποτελέσματα από τη μηχανή. Θα πρέπει να ομολογήσουμε ότι τόσο τα χρώματα όσο και η διάταξη όλων αυτών των επιλογών δεν κάνουν ξεκάθαρο το σκοπό της μηχανής: την αναζήτηση και ανεύρεση πληροφοριών. Η αρχική σελίδα μοιάζει περισσότερο με σελίδα περιοδικού στην οποία έχουν συσσωρευτεί δεκάδες πληροφοριών.



Σχήμα 6.2.: Η Κεντρική Σελίδα του Θεματικού Καταλόγου Excite

6.2.2. Βασικά Χαρακτηριστικά

Σε ό,τι αφορά τα γενικά χαρακτηριστικά της μηχανής, η Excite εκτός του world

wide web καλύπτει το Usenet, τις υπηρεσίες listservs, όπως επίσης άρθρα και ειδήσεις σε ένα σύνολο από 450 και πλέον δημοσιεύματα. Η Excite ευρετηριάζει το *πλήρες κείμενο* ***(full-text indexing)*** των σελίδων που συγκεντρώνει, ενώ επίσης διαθέτει και το μοναδικό χαρακτηριστικό της *νοηματικής ευρετηρίασης* ***(context-based indexing)*****,** το οποίο την ξεχωρίζει από τις υπόλοιπες μηχανές γενικής αναζήτησης.

Η Excite υποστηρίζει τόσο τη θεματική αναζήτηση, με βάση τα *θεματικά ευρετήρια/ καταλόγους* που διαθέτει ***(subject search)***, όσο και την αναζήτηση με βάση *λέξεις-κλειδί* ***(keyword search)***. Ακόμη, ευρετηριάζει την *ηλεκτρονική διεύθυνση* ***(URL)*** όπως επίσης τα ***image maps***. Η Excite υποστηρίζει τόσο απλή όσο και προχωρημένη αναζήτηση και για το λόγο αυτό διαθέτει δύο διαφορετικές σελίδες. Τόσο στην απλή όσο και στην προχωρημένη αναζήτηση η Excite υποστηρίζει τη χρήση των λογικών τελεστών Boolean, αν και στην προχωρημένη αναζήτηση η σύνταξη που θα πρέπει να χρησιμοποιήσουμε είναι η ίδια με αυτή της Infoseek, δηλαδή αντιστοιχίζονται οι τελεστές AND, NOT και OR με τις λέξεις Must, Must Not και Can αντίστοιχα.

Η εμφάνιση της Excite είναι μάλλον ξεχωριστή σε σχέση με τις άλλες μηχανές που συναντήσαμε ως τώρα, αφού η σχεδίαση της σελίδας ίσως μπερδέψει το χρήστη με τις πολλαπλές επιλογές και *παραπομπές* ***(links)*** αριστερά και δεξιά της φόρμας με το πεδίο αναζήτησης, το οποίο βρίσκεται στο άνω μέρος της σελίδας. Ακριβώς κάτω από αυτό ακολουθούν οι παραπομπές σε πηγές άμεσης πληροφόρησης, όπως Yellow Pages, NewsTracker, People Find.

Τόσο για την απλή όσο και για την προχωρημένη αναζήτηση η Excite παρέχει στο χρήστη αναλυτικές και καλογραμμένες *οδηγίες χρήσης* ***(help files)*** και παραδείγματα, προκειμένου να γίνει όσο το δυνατόν καλύτερη αξιοποίηση των δυνατοτήτων της μηχανής και των ιδιαίτερων χαρακτηριστικών που αυτή διαθέτει. Ένα τέτοιο μοναδικό χαρακτηριστικό είναι η *αναζήτηση-ευρετηρίαση του ευρύτερου νοήματος* ***(concept-based indexing or searching)***, όπως αναφέρθηκε και στην ενότητα 3.3.2.1. Σε αντίθεση με άλλα εργαλεία και μηχανές αναζήτησης, τα συστήματα που υποστηρίζουν την αναζήτηση αυτού του τύπου δεν αναζητούν μόνο τους όρους που έχει εισαγάγει ο χρήστης αλλά, προχωρώντας ένα βήμα παραπέρα, προσπαθούν να προσδιορίσουν το ευρύτερο νόημα των όρων αυτών, προκειμένου να ανακτήσουν όσο το δυνατόν πιο συναφείς πληροφορίες. Σε πολλές περιπτώσεις μια αναζήτηση τέτοιου είδους θα καταφέρει να προσδιορίσει με αρκετή επιτυχία το ζητούμενο θέμα με βάση τους όρους που έχει εισαγάγει ο χρήστης και να ανακτήσει τις σχετικές εγγραφές-web pages, άσχετα με το αν αυτές εμπεριέχουν τους συγκεκριμένους όρους. Στο σημείο

αυτό θα πρέπει να αναφερθεί ότι η Excite κατέχει την πρώτη θέση ανάμεσα στα συστήματα που υποστηρίζουν την αναζήτηση τέτοιου είδους.

Σε γενικές γραμμές, τα συστήματα τέτοιου είδους βασίζονται σε πολύπλοκους αλγόριθμους και τεχνικές βασισμένες σε γλωσσολογικά μοντέλα προκειμένου να προσδιορίσουν το θέμα της αναζήτησης. Η Excite χρησιμοποιεί τη γλώσσα των αριθμών για να προσδιορίσει τη συχνότητα με την οποία εμφανίζονται κάποιοι κρίσιμοι όροι μέσα στις σελίδες τις οποίες ευρετηριάζει και τους συσχετισμούς μεταξύ τους. Με βάση τη στατιστική ανάλυση, η μηχανή προσδιορίζει με αρκετή επιτυχία το θέμα για το οποίο τελικά αναζητούνται πληροφορίες. Βέβαια, σε καμία περίπτωση δε θα πρέπει να θεωρηθούν τα εργαλεία ***concept-based indexing or searching*** ως πανάκεια για κάθε είδους ερώτημα. Τα ποσοστά επιτυχίας μπορεί να είναι αρκετά υψηλά, απέχουν όμως σημαντικά από το ιδανικό. Οι ίδιες οι λέξεις, εξάλλου έχοντας πολλαπλές μεταφορικές σημασίες, δυσκολεύουν αυτού του είδους την αναζήτηση και μειώνουν τα ποσοστά επιτυχίας σε μεγάλο βαθμό. Κατά γενική ομολογία, ωστόσο, η ανάπτυξη συστημάτων τέτοιου είδους έχει αρκετό δρόμο να διανύσει προτού τελειοποιηθεί και σίγουρα θα αποτελέσει πεδίο έντονου ανταγωνισμού μεταξύ των μηχανών αναζήτησης τα επόμενα χρόνια.

6.2.3. Ειδικά Χαρακτηριστικά

- **Λογική αναζήτησης και σύνταξης**

Η Excite υποστηρίζει τη χρήση όλων των τελεστών *Boolean*, AND, NOT και OR, κάνοντας χρήση των συμβόλων + και - σε ό,τι αφορά την απλή αναζήτηση και ειδικής ορολογίας για την προχωρημένη αναζήτηση. Η λέξη Must χρησιμοποιείται αντί του τελεστή AND, η λέξη Can χρησιμοποιείται αντί του τελεστή OR και, τέλος, η φράση Must Not αντί του τελεστή NOT.

Η μηχανή χρησιμοποιεί τον τελεστή OR ως *default*, δηλαδή αυτόματα θα αναζητήσει και θα ανακτήσει όποιον από τους όρους βρει κατά την πορεία της αναζήτησης.

- **Αναζήτηση φράσεων (Phrase Searching)**

Η μηχανή υποστηρίζει την αναζήτηση φράσεων. Η Excite θεωρεί τους όρους που περικλείονται μέσα σε *ανωφερή εισαγωγικά* ***(quotation marks, “”)*** ως φράση, την οποία

και αναζητά αυτούσια. Ακόμη, η χρήση της παύλας (-) ανάμεσα σε δύο όρους θα καθοδηγήσει την μηχανή στο να αναζητήσει τους όρους αυτούς ως φράση.

- **Τελεστές περιορισμού αναζήτησης (Limit Operators)**

Η Excite μπορεί να εστιάσει την αναζήτηση σε συγκεκριμένες περιοχές της σελίδας, για παράδειγμα στον τίτλο ή στην ***ηλεκτρονική διεύθυνση (URL)***. Επίσης έχει τη δυνατότητα να αναζητήσει *ειδήσεις* ***(news)***, προσωπικές πληροφορίες, όπως ***e- mail*** και διευθύνσεις άλλων χρήσεων, και γεωγραφικούς χάρτες. Τα χαρακτηριστικά αυτά είναι διαθέσιμα στην προχωρημένη αναζήτηση, αφού στην απλή αναζήτηση ο χρήστης δεν έχει στη διάθεση του κάποια *πεδία-επιλογές περιορισμού* ***(limit options)***.

- **Αποκοπή (Truncation)**

Η Excite δεν υποστηρίζει την αποκοπή των Λέξεων.

- **Αναγνώριση πεζών/ κεφαλαίων (Case sensitivity)**

Η μηχανή αναζήτησης δεν υποστηρίζει την *αναγνώριση πεζών κα κεφαλαίων χαρακτήρων* ***(case sensitivity)***. Έτσι, στην περίπτωση που εισαγάγουμε τον αναζητήσιμο όρο με πεζούς χαρακτήρες, η μηχανή θα τον ανακτήσει μόνο όπου είναι γραμμένος με πεζούς, ενώ αν τον εισαγάγουμε με κεφαλαίους χαρακτήρες, η μηχανή θα αναζητήσει και θα ανακτήσει τον όρο μόνο όπου τον συναντήσει με κεφαλαίους χαρακτήρες.

6.2.4. Crawling

Το ειδικό λογισμικό της Excite, το ***mega-spider*** ανιχνεύει τον ιστό κάθε 21 ημέρες. Αυτό είναι το ειδικό λογισμικό που θα επισκεφτεί τις ιστοσελίδες των περισσότερων ανθρώπων μετά από μια αρχική υποβολή και που θα τις ξαναεπισκεφτεί σε τακτό χρονικό διάστημα. Συλλέγει το μέγιστο δυνατό από την υποβληθείσα σελίδα μέσα σε ένα χρονικό διάστημα 30 δευτερολέπτων. Μετά το πέρας αυτού του χρόνου θα επιστρέψει αργότερα. Η Excite επίσης τρέχει ένα ***Fresh-Spider,*** το οποίο κάθε εβδομάδα διερευνά δύο εκατομμύρια σημαντικές ιστοσελίδες, όπως ορίζεται από την Excite.

6.2.5. Ταξινόμηση (Ranking)

Όπως οι περισσότερες μηχανές, η Excite χρησιμοποιεί:

- Τις σελίδες με λέξεις-κλειδιά στον τίτλο, οι οποίες επαναλαμβάνονται συχνά σε σχέση με το υπόλοιπο του εγγράφου.
- Τις λέξεις που περιλαμβάνονται σε πλήρεις προτάσεις.
- **Το πλήθος συνδέσεων:** Οι σελίδες που έχουν πολλές συνδέσεις που καταλήγουν σε αυτές, ειδικά εκείνες που συνδέονται από τις δημοφιλείς ιστοσελίδες.

Η Excite προσπαθεί να *κανονικοποιήσει* ***(normalize)*** τα έγγραφα όταν τα ταξινομεί. Παραδείγματος χάριν, μια σελίδα με μερικές λέξεις μόνο, συμπεριλαμβανομένων των όρων αναζήτησης, μπορεί να θεωρηθεί πιο σχετική έναντι ενός μεγαλύτερου εγγράφου όπου οι πρόσθετες λέξεις "αραιώνουν" τις εμφανίσεις των όρων αναζήτησης. Η Excite προσπαθεί να ρυθμίσει αυτές τις ακρότητες. Επίσης στα μέσα του 1998, η Excite άρχισε να επεξεργάζεται τα αρχικά αποτελέσματα της αναζήτησής της εισάγοντάς τα και ταιριάζοντάς τα με τις κατηγορίες του θεματικού καταλόγου της.

6.2.6. Ευρετήριο (Indexing)

- Η Excite *ευρετηριάζει το πλήρες κείμενο* ***(full-text index)***.
- Παρέχει επίσης *την έρευνα με βάση τη νοηματική ευρετηρίαση* ***(concept searching)***. Αυτό σημαίνει ότι η Excite καταλαβαίνει ότι οι λέξεις έχουν *συνώνυμα* ***(synonyms)***.
- Η Excite δεν ευρετηριάζει τα σχόλια.
- Η Excite θεωρεί τις βασικές αντωνυμίες και τα κοινά επίθετα ως ***stop words***.
- Η Excite υποστηρίζει τα *metatags*. Αυτή η δυνατότητα ενσωματώθηκε στα μέσα του 1998. Εάν δεν υπάρχουν tags, η Excite δημιουργεί μια περιγραφή του κειμένου επιλέγοντας προτάσεις που θεωρεί ότι καλύτερα αντιπροσωπεύουν το περιεχόμενο της σελίδας. Σε άλλη περίπτωση περίπου 395 χαρακτήρες του κείμενο εμφανίζονται.
- Η Excite θα εμφανίσει 70 χαρακτήρες του τίτλου.

6.2.7. Spamming

Η Excite προσπαθεί να εντοπίσει και να εμποδίσει την εμφάνιση του *spsmming* πριν προσθέσει στο ευρετήριό της την ιστοσελίδα. Παραδείγματος χάριν εάν εμφανιστεί μια συμβολοσειρά λέξεων όπως:

money money money money money money money money

θα αντικατασταθεί η υπερβολική επανάληψη, έτσι ώστε ουσιαστικά, η συμβολοσειρά να γίνει:

money xxxxx xxxxx xxxxx xxxxx xxxxx xxxxx xxxxx

Η Excite εάν δεν μπορεί να παραβλέψει και να διαχωρίσει το *spamming*, τότε τιμωρεί τη σελίδα. Ειδικότερα όσο περισσότερο ανιχνεύει την ασυνήθιστη επανάληψη, τόσο περισσότερο θα τιμωρήσει τη σελίδα. Η Excite δεν τιμωρεί τη χρήση του *κρυμμένου κειμένου* ***(hidden text)***, αλλά οι περιορισμοί θα ισχύσουν εάν το κρυμμένο κείμενο χρησιμοποιείται για συγκάλυψη *spam* περιεχομένου.

6.2.8. Παρουσίαση Αποτελεσμάτων

- **<u>Εμφάνιση και διάταξη των αποτελεσμάτων (Display & order results)</u>**

Κατά την παρουσίαση των αποτελεσμάτων περιλαμβάνονται ο τίτλος της *web page* ή του εγγράφου που ανακτήθηκε, η *ηλεκτρονική διεύθυνση* ***(URL)*** όπως επίσης μια σύντομη περίληψη, προκειμένου ο χρήστης να έχει μια πρώτη ιδέα σχετικά με τα περιεχόμενα της σελίδας, καθώς και ένας *δείκτης σχετικότητας* ***(relevance)*** με τη μορφή ποσοστού επί τοις εκατό (%). Επίσης, υπάρχει δυνατότητα να εμφανίζεται στη σελίδα των αποτελεσμάτων μόνο ο τίτλος του εγγράφου- *web page*. Σε ό,τι αφορά τη διάταξη των αποτελεσμάτων, στην πρώτη σελίδα παρουσιάζεται μια λίστα με *προτεινόμενα sites* ***(recommended sites)***, τα οποία η μηχανή θεωρεί ότι περιέχουν συναφείς πληροφορίες, ενώ στο τέλος κάθε τέτοιας παραπομπής υπάρχει το ***Link More Like This***, το οποίο ουσιαστικά εκτελεί ένα νέο ερώτημα με βάση τη συγκεκριμένη εγγραφή.

Το ερώτημα αυτό ονομάζεται ***query-by-example***, δηλαδή πρόκειται για ερώτημα που βασίζεται σε μια ήδη έτοιμη εγγραφή (ως παράδειγμα) για να αναζητήσει συναφείς με αυτή εγγραφές. Τα αποτελέσματα παρουσιάζονται με σειρά αντιστρόφως ανάλογη της σχετικότητας, δηλαδή η πιο σχετική *web page* να εμφανίζεται στην αρχή της

σελίδας, συνοδευόμενη, όπως αναφέραμε, από το δείκτη σχετικότητας. Τέλος, ο χρήστης μπορεί και πάλι να επέμβει προκειμένου να μορφοποιήσει τη σελίδα των αποτελεσμάτων, επιλέγοντας να τα ταξινομήσει με βάση κάποιο site, αφού στις περισσότερες περιπτώσεις κάποια από τα sites που ανακτήθηκαν από τη μηχανή εμφανίζονται περισσότερες από μία φορές. Με την ταξινόμηση αυτή ο χρήστης μπορεί άμεσα να εξακριβώσει ποια *sites* συγκεντρώνουν τα περισσότερα *links*, προκειμένου να θέσει μια προτεραιότητα σε αυτά.

- **Βελτίωση αποτελεσμάτων**

Η Excite διαθέτει δύο τρόπους με τους οποίους ο χρήστης μπορεί να βελτιώσει τα αποτελέσματα της αναζήτησης του. Ο πρώτος τρόπος είναι η επιλογή ***Search Wizard,*** που ουσιαστικά επιτρέπει στο χρήστη να εισαγάγει εκ νέου μια σειρά όρων, τους οποίους προτείνει η μηχανή, προκειμένου να εξειδικεύσει το ερώτημα του. Ο δεύτερος τρόπος είναι ιδιαίτερα χρήσιμος, καθώς επιτρέπει ατό χρήστη ακολουθώντας την επιλογή ***More Like This*** να αναζητήσει συναφείς όρους, προκειμένου να βελτιώσει τα διαθέσιμα αποτελέσματα.

- **Ειδικά χαρακτηριστικά**

Η Excite εκτός της απλής και της προχωρημένης αναζήτησης προσφέρει μια σειρά διευκολύνσεων προς το χρήστη, προκειμένου η αναζήτηση του να γίνει πιο εύκολη και αποδοτική. Ειδικότερα, παρέχεται η δυνατότητα αναζήτησης μέσω των λεγόμενων channels, που στην ουσία δεν είναι τίποτα άλλο από θεματικοί κατάλογοι-κατηγορίες, ενώ επίσης δίνεται η δυνατότητα για αναζήτηση εταιρικών πληροφοριών, newsgroups και επίκαιρων ειδήσεων. Τέλος, η Excite διαθέτει μια σειρά παραμέτρων όπως ειδικό *link* για την αναζήτηση μετοχών, εταιρειών, ταξιδιωτικών πληροφοριών, μετεωρολογικών πληροφοριών όπως επίσης ειδήσεων με βάση την υπηρεσία News Tracker.

6.2.9. Περιφερειακές εκδόσεις *(Regional Editions)*

Η Excite λειτουργεί σε μια σειρά περιφερειακών εκδόσεων. Αυτές μοιράζονται μια κοινή βάση δεδομένων, που βασίζεται στις ΗΠΑ. Όταν κάποιος εκτελεί μια αναζήτηση σε μια συγκεκριμένη χώρα, η Excite θα εφαρμόσει *το φιλτράρισμα περιοχών- πεδίων* ***(Domain Filtering)*** για την παροχή των σχετικών αποτελεσμάτων.

Παραδείγματος χάριν, κατά τη διάρκεια μιας συγκεκριμένης αναζήτησης στη Βρετανία, η Excite θα φιλτράρει και θα παραβλέψει τις περιοχές που δεν υπάγονται στην περιοχή .uk.

Προφανώς υπάρχουν μερικές ιστοσελίδες συγκεκριμένων περιοχών που χρησιμοποιούν μη τοπικές περιοχές, όπως εκείνες που τελειώνουν σε .com ή .net. Η Excite τις αντισταθμίζει αυτές με την προσθήκη τους σε έναν κατάλογο συνυπολογισμού. Κάθε περιφερειακή έκδοση πρέπει επίσης να έχει το δικό της έντυπο (φόρμα) πρόσθεσης *URL*. Αυτές διαβιβάζουν τις πληροφορίες *URL* στο ***Excite spider***, το οποία θα προσθέσει τελικά την περιοχή αυτή στην κύρια βάση δεδομένων. Επειδή οι πληροφορίες πηγαίνουν στην ίδια θέση, δεν είναι ανάγκη να υποβληθεί σε κάθε έναν από τους περιφερειακούς οδηγούς χωριστά.

6.2.10. Branded Μηχανές Αναζήτησης

Η Excite ισχυροποιεί τις υπηρεσίες της ***AOL Netfind*** στις ΗΠΑ και τον Καναδά και της Netscape μηχανής αναζήτησης. Και στις δύο περιπτώσεις, τα ακατέργαστα (πρώτα) αποτελέσματα των μηχανών αναζήτησης πρέπει να είναι ίδια με τα αποτελέσματα της Excite. Περιστασιακά μικρές διαφορές μπορούν να εμφανιστούν, αλλά αυτό είναι σπάνιο. Δεδομένου ότι και οι δύο υπηρεσίες βασίζονται στην κύρια βάση δεδομένων της Excite, δεν είναι αναγκαίο να υποβληθεί σε αυτές χωριστά. Εάν κάποια ιστοσελίδα προστεθεί στην Excite, κατόπιν θα προστεθεί και σε αυτές.

Η Excite επίσης παράγει το περιεχόμενο των καναλιών της Netscape, το οποίο περιλαμβάνει τις λίστες καταλόγου. Η βάση δεδομένων της Excite παρέχει επίσης τα αποτελέσματα της αναζήτησης της Magellan, αν και οι λίστες μπορούν να είναι διαφορετικές από την κύρια υπηρεσία της Excite, καθώς μπορεί να χρησιμοποιηθεί ένας λίγο διαφορετικός αλγόριθμος ταξινόμησης. Αντίθετα η υπηρεσία WebCrawler της Excite χρησιμοποιεί το δικό της ξεχωριστό ευρετήριο.

ΚΕΦΑΛΑΙΟ 7

ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ ΜΕ ΒΑΣΗ ΤΟ ΠΡΟΦΙΛ ΤΟΥ ΧΡΗΣΤΗ

Σε αυτό το κεφάλαιο της εργασίας θα γίνει αναφορά στις μηχανές αναζήτησης που βασίζουν την αναζήτησή τους στο προφίλ του χρήστη. Σήμερα υπάρχουν πάρα πολλά sites που δίνουν τη δυνατότητα στους χρήστες να διαμορφώνουν το προφίλ τους στο σύστημα της κάθε μηχανής και να αναζητούν πληροφορίες σύμφωνα με τα ενδιαφέροντά τους. Όπως αναφέρθηκε και στο δεύτερο κεφάλαιο της παρούσας εργασίας, το ***User profiling*** μπορεί να οριστεί ως η προσπάθεια να δημιουργηθεί ένα *προφίλ* ***(profile)*** του χρήστη με βάση τα ενδιαφέροντά του και τις συνήθειές του με σκοπό τη βελτίωση της *αλληλεπίδρασης του ανθρώπου με τον υπολογιστή* ***(human computer interaction)***. Τα συστήματα *User Modeling* διαφέρουν στους τρόπους που αποκτούν, χρησιμοποιούν και παρουσιάζουν το προφίλ του χρήστη.

Σε αυτή την ενότητα θα περιγράψουμε δύο μηχανές αναζήτησης από τις πάρα πολλές που έχουν αναπτυχθεί στο διαδίκτυο, που χρησιμοποιούν το *User profiling*. Το ένα σύστημα ονομάζεται Amalthaea, το οποίο είναι γενικής αναζήτησης και το άλλο είναι το LawBot, το οποίο χρησιμοποιείται για νομικές έρευνες από επαγγελματίες.

7.1. Amalthaea

7.1.1. Εισαγωγή

Ο πυρήνας των διαδικασιών-λειτουργιών του συστήματος βρίσκεται συγκεντρωμένος ***(centralized)*** στον ***Amalthaea server***. Ο server περιέχει τους διάφορους πράκτορες *φιλτραρίσματος* ***(Filtering Agents)*** και *αναζήτησης* ***(Discovery Agents)*** για κάθε χρήστη, τις προτιμήσεις του και πληροφορίες για τα *sites* που ο χρήστης έχει επισκεφτεί ήδη. Από την πλευρά του χρήστη, η Amalthaea ελέγχεται μέσω ενός *γραφικού interface* ***(Amalthaea User Interface - AUI)***, το οποίο τρέχει στον υπολογιστή του χρήστη.

Το AUI δημιουργείται με χρήση της γλώσσας ***java*** της Sun. Όταν ένας χρήστης συνδέεται με τον *Amalthaea server*, το AUI παρουσιάζεται στην οθόνη του χρήστη ως παράθυρο της *java* χωριστά από τον *browser*. Το *Amalthaea Interface* τρέχει συνεχώς όταν ο χρήστης είναι συνδεδεμένος με το σύστημα. Μέσω του AUI ο χρήστης μπορεί να λάβει λίστες από *sites* του διαδικτύου που τον ενδιαφέρουν, να ανατροφοδοτήσει, να διαμορφώσει και να απεικονίσει την κατάσταση του συστήματος. Το AUI αποτελείται από διαφορετικά μέρη που αντιστοιχούν σε διαφορετικές λειτουργίες του συστήματος.

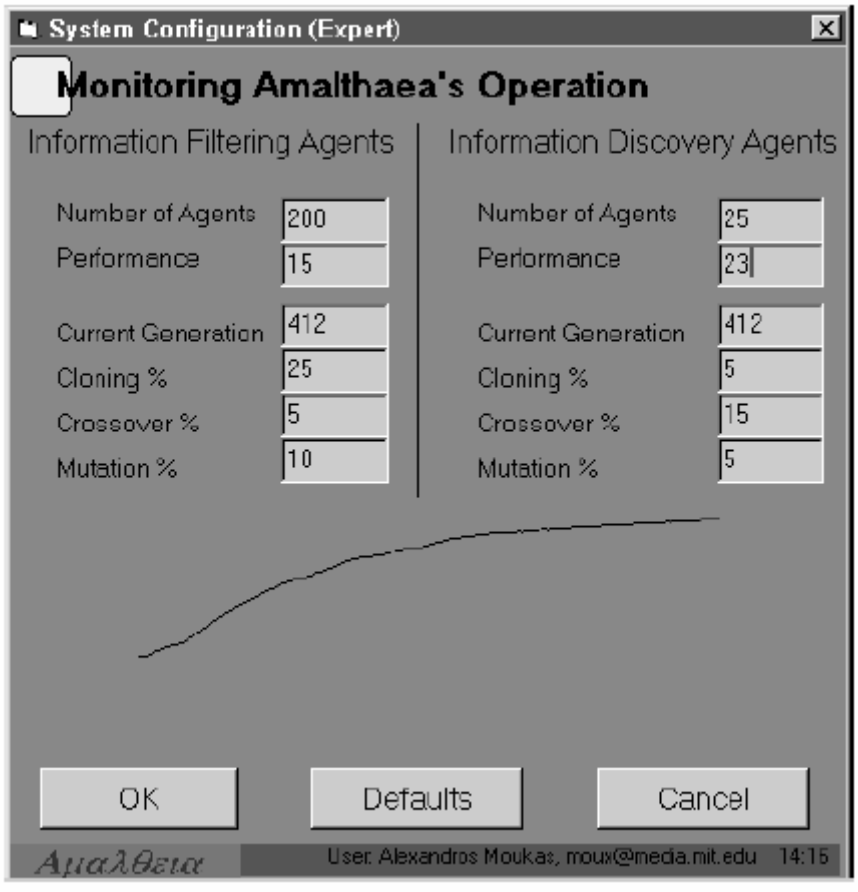
7.1.2. Παραμετροποίηση του Συστήματος

Ο χρήστης έχει τη δυνατότητα να αλλάξει τις βασικές παραμέτρους της Amalthaea και να διαμορφώσει την κατάστασή της. Οι βασικές παράμετροι μπορούν να διαμορφωθούν με δύο τρόπους:

- Ο *τρόπος χαμηλού επιπέδου* ***(low level mode)*** (ειδικός) επιτρέπει το χειρισμό των μεταβλητών όπως την εξέλιξη, τη μεταλλαγή, και τα ποσοστά διασταύρωσης καθώς επίσης και τον αριθμό των πρακτόρων φιλτραρίσματος και αναζήτησης στο σύστημα και τον τρόπο που αυτοί συνεργάζονται.
- Ο *τρόπος υψηλού επιπέδου* ***(higher level mode)*** είναι καταλληλότερος για τους *τελικούς χρήστες* ***(end users)***: παρέχει υψηλότερου επιπέδου θεωρήσεις για τις παραπάνω low level παραμέτρους. Αυτές οι θεωρήσεις περιλαμβάνουν τη διαμόρφωση των χαρακτηριστικών όπως "Quick Learning" έναντι "Slow Learning", "Fast Adaptation", έναντι "Stable Interests" κ.λ.π., τα οποία συντονίζουν τις χαμηλού επιπέδου μεταβλητές κατά τέτοιο τρόπο ώστε το σύστημα είτε να μαθαίνει γρηγορότερα, αλλά με τη μικρότερη ακρίβεια και εύρος, είτε πιο αργά αλλά ακριβέστερα.

Παραδείγματος χάριν το *ποσοστό εξέλιξης* ***(evolution rate)*** του συστήματος είναι μικρότερο στο "Slow Learning" από αυτό του "Quick Learning". Οι προκαθορισμένες τιμές στις παραπάνω παραμέτρους ορίστηκαν στο σύστημα μετά τη διεξαγωγή ενός συνόλου πειραμάτων μικρής κλίμακας στις *παραμέτρους εξέλιξης* ***(evolution parameters)***.

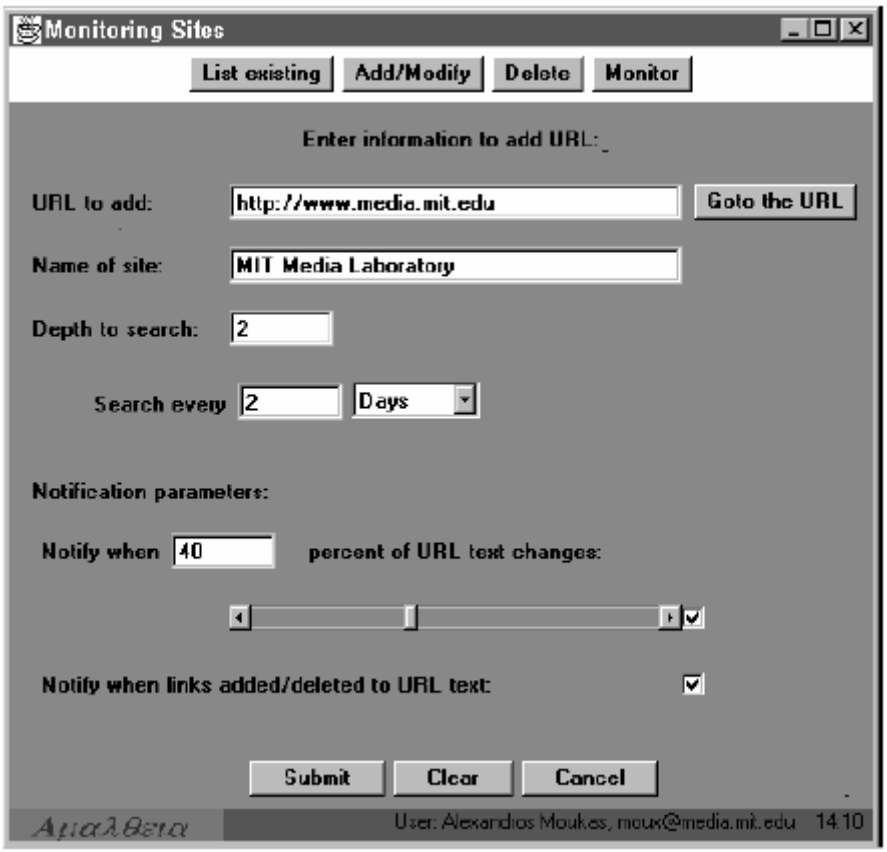
Κατά τη διάρκεια της παραμετροποίησης του AUI, ο χρήστης μπορεί να γνωρίζει την τρέχων κατάσταση του συστήματος από την πλευρά της ικανότητας των πρακτόρων και της εξέλιξής τους στο χρόνο (με τη μορφή σχεδιαγράμματος). Κάθε χρήστης μπορεί περαιτέρω να προσαρμόσει την Amalthaea στις ανάγκες του, τροποποιώντας τις παραμέτρους όπως τον αριθμό εγγράφων που το σύστημα πρέπει να ανακτήσει και να τα παρουσιάσει στους IFAs για το φιλτράρισμα και στη συνέχεια τον αριθμό των εγγράφων των IFA που πρέπει να παρουσιαστούν στο χρήστη. Ένα παραθύρου παραμετροποίησης του AUI φαίνεται στο παρακάτω Σχήμα 7.1.



Σχήμα 7.1.: Το Παράθυρο Παραμετροποίησης της Amalthaea

Ο χρήστης μπορεί να καθορίσει ότι ορισμένες περιοχές που ενημερώνονται σε

καθορισμένα χρονικά διαστήματα πρέπει να ελέγχονται από την Amalthaea. Ο χρήστης μπορεί να εισαγάγει πληροφορίες σχετικά με το πόσο συχνά θα ελέγχεται η περιοχή, και εάν θέλει ο έλεγχος να περιλαμβάνει URLs που βρίσκονται σε αυτή την περιοχή (και σε πόσο βάθος). Άλλες επιλογές περιλαμβάνουν παραμέτρους όπως, πότε ο χρήστης πρέπει να ενημερώνεται (δηλ. όταν κάτι στην περιοχή άλλαξε, όταν νέα URL προστέθηκε ή διαγράφηκε από αυτή την περιοχή, όταν η περιοχή άλλαξε πάνω από ένα καθορισμένο ποσοστό).

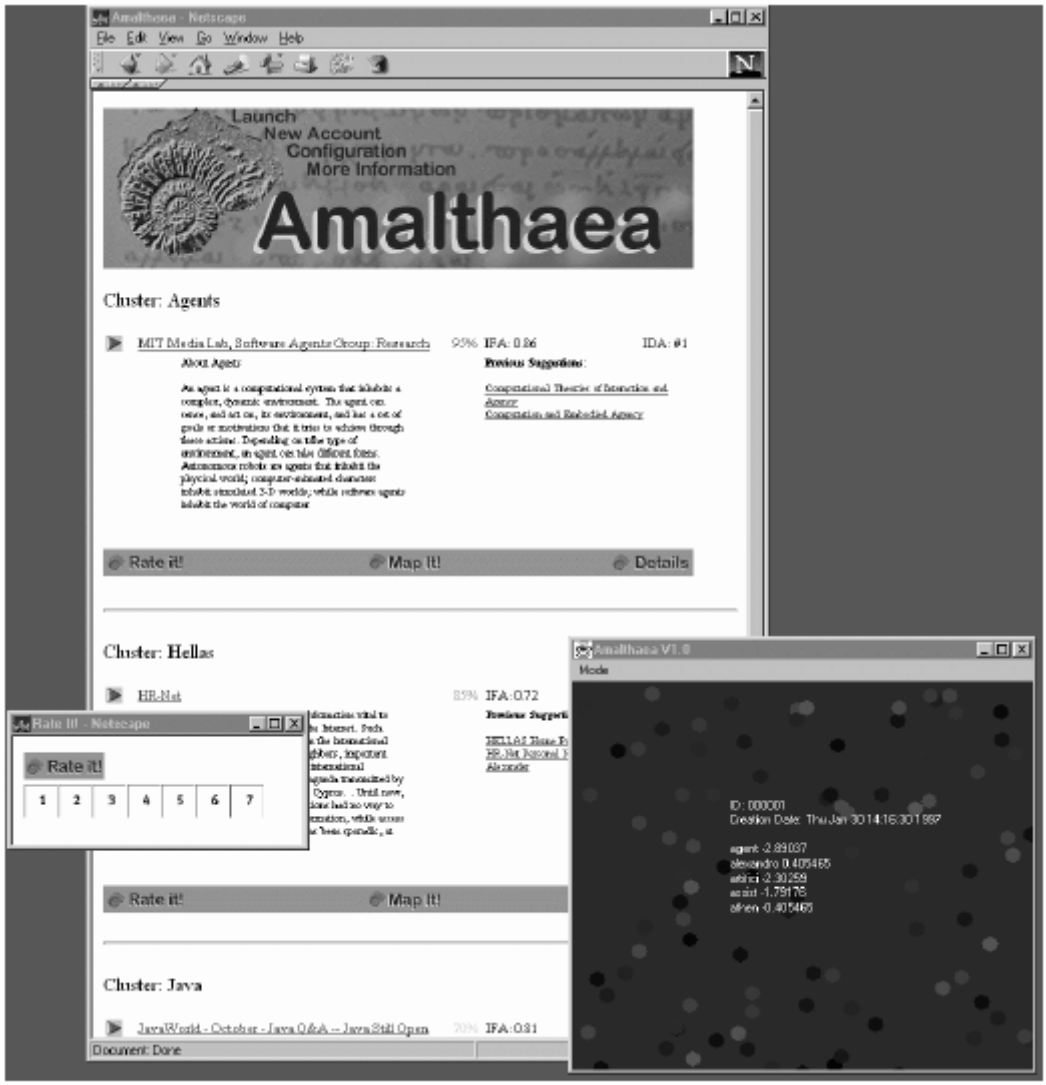


Σχήμα 7.2.: Το Παράθυρο Ελέγχου της Amalthaea

7.1.3. Μελέτη του Συστήματος

Αυτό είναι το τμήμα του AUI όπου το σύστημα παρουσιάζει τις προτάσεις του στο χρήστη και περιμένει την απάντησή του. Οι προτεινόμενες URLs οργανώνονται σε τμήματα (όπου κάθε τμήμα περιέχει τα έγγραφα που ανακτήθηκαν από μια ορισμένη ομάδα πρακτόρων φιλτραρίσματος πληροφοριών), και εμφανίζονται μαζί με ένα μικρό μέρος του πραγματικού περιεχομένου της περιοχής και του επιπέδου εμπιστοσύνης που το σύστημα διαθέτει για κάθε URL. Ο χρήστης μπορεί να κάνει *click* σε αυτά τα URLs και το AUI ανοίγει τον browser σε αυτή τη συγκεκριμένη URL. Μπορεί να ελέγξει το

site και να επιλέξει έπειτα ένα μέτρο για αυτή την πρόταση (το AUI τρέχει πάντα σε ξεχωριστό παράθυρο έξω από τον browser). Αυτή η εκτίμηση πρόκειται να είναι η ανατροφοδότηση του συστήματος.



Σχήμα 7.3.: Μια Έρευνα του Χρήστη με τα Προτεινόμενα Sites

Στο παραπάνω σχήμα (Σχήμα 7.3.), στην κάτω αριστερή γωνία υπάρχει το παράθυρο της ταξινόμησης (όσο μεγαλύτερος είναι ο αριθμός, τόσο πιο ενδιαφέρουσα ήταν η πρόταση) και στην κάτω δεξιά γωνία υπάρχει μια σχετική απεικόνιση των πρακτόρων φιλτραρίσματος πληροφοριών. Κάθε πιθανή είσοδος ενός *site* στο σύστημα περιέχει την URL της συνιστώμενης περιοχής, τον τίτλο της, τα πρώτα 300 bytes του περιεχομένου του καθώς επίσης και τους IFA και IDA πράκτορες που προτείνουν αυτό το *site* και την εμπιστοσύνη της πρότασης. Στο δεξί μισό τμήμα του browser, το

σύστημα επιδεικνύει άλλες περιοχές (σελίδες WWW) που προτάθηκαν από τον ίδιο πράκτορα IFA σε προηγούμενες αναζητήσεις του.

Η ανατροφοδότηση χρησιμοποιείται από το σύστημα για να δώσει την πίστωση στους πράκτορες για το φιλτράρισμα των πληροφοριών που ήταν αρμόδιοι για την επιλογή αυτού του εγγράφου. Μια επταβάθμια κλίμακα του μηχανισμού ανατροφοδότησης είναι διαθέσιμη στο χρήστη. Μια εκτίμηση με 1 σημαίνει ότι προτεινόμενη περιοχή ήταν πολύ κακή ενώ μια εκτίμηση με 7 σημαίνει ότι η περιοχή ήταν άριστη. Εάν ο χρήστης επιλέξει να μην χρησιμοποιήσει τη ρητή μορφή της ανατροφοδότησης, το σύστημα προσπαθεί να συμπεράνει, το πόσο καλή ήταν για το χρήστη μια σελίδα, με έναν έμμεσο τρόπο. Μια διαδικασία *JavaScript* που τρέχει στον *browser* ελέγχει τον χρόνο που ο χρήστης ξοδεύει πραγματικά ακολουθώντας μια σύνδεση, το μέγεθος των περιεχομένων της σύνδεσης, τον χρόνος που η μηχανή του χρήστη ήταν ανενεργή, τον χρόνο που το παράθυρο του *browser* ήταν σε πρώτο πλάνο, και υπολογίζει έμμεσα μια εκτίμηση ανατροφοδότησης. Η βασική ιδέα που κρύβεται πίσω από αυτόν τον μηχανισμό είναι ότι εάν ο χρήστης δεν βρίσκει ενδιαφέρον το συγκεκριμένο site, δεν θα μείνει πολύ σε αυτό. Οι υπόλοιπες παράμετροι (όπως ο χρόνος μη απασχόλησης του υπολογιστή) είναι ένας τρόπος ελέγχου πιθανών προειδοποιήσεων. Τελικά ο χρήστης μπορεί να επιλέξει διάφορες λέξεις κλειδιά από το διάνυσμα των εγγράφων που περιγράφουν καλύτερα το συγκεκριμένο έγγραφο και τα βάρη αυτών των λέξεων κλειδιών θα ενισχυθούν.

Ο χρήστης έχει την επιλογή να συλλέγει περισσότερες πληροφορίες για μια συγκεκριμένη πρόταση κάνοντας *click* στην επιλογή ***"Details"*** στην αφομοίωση.

7.1.4. Η Πρώτη Επαφή με την Amalthaea

Η Amalthaea δημιουργήθηκε από την παραγωγή διάφορων πρακτόρων φιλτραρίσματος πληροφοριών και από τον ίδιο αριθμό πρακτόρων αναζήτησης. Αυτή η πρώτη γενεά των πρακτόρων που φιλτράρουν τις πληροφορίες πρέπει να σχετίζονται με τα ενδιαφέροντα των χρηστών. Αυτό μπορεί να πραγματοποιηθεί με έναν από τους ακόλουθους τρόπους:

- Ο χρήστης υποβάλλει έναν κατάλογο από αγαπημένα bookmarks ή έγγραφά του. Αυτός είναι συνήθως ο κατάλογος των bookmarks. Κάθε ένα από τα sites του καταλόγου εξετάζεται και για κάθε site δημιουργείται ένας πράκτορας φιλτραρίσματος πληροφοριών.

- Η Amalthaea ελέγχει τα αρχεία που έχει επισκεφθεί ο browser του χρήστη. Πολλοί browsers διατηρούν ιστορικά αρχεία (συχνά έχουν μέγεθος αρκετά MBytes) που περιέχουν όλες τις URLs που ο χρήστης έχει επισκεφθεί. Το ιστορικό αρχείο είναι επίσης χρήσιμο στην περίπτωση που η Amalthaea προτείνει κάποια sites: εάν ο χρήστης έχει ήδη επισκεφτεί ένα συγκεκριμένου site, η Amalthaea δεν το προτείνει ποτέ ξανά, εκτός και αν έχει αλλάξει από την τελευταία επίσκεψη του χρήστη. Το σύστημα αναλύει αυτές τις πληροφορίες και προσπαθεί να δημιουργήσει πρότυπα προκειμένου να αποφασίσει εάν πρέπει να ελέγξει κάποια sites.
- Οι χρήστες μπορούν να οδηγήσουν την Amalthaea σε μια συγκεκριμένη σελίδα (πιθανώς όταν κάνουν browsing) και να ζητήσουν την παραγωγή περισσότερων πρακτόρων που θα αναζητήσουν παρόμοιες πληροφορίες. Σε αυτή την περίπτωση, το περιεχόμενο της σελίδας ανακτάται και ένας νέος πράκτορας φιλτραρίσματος πληροφοριών δημιουργείται βασισμένος σε αυτό.
- Τέλος οι χρήστες μπορούν να επιλέξουν προδιαμορφωμένες ρυθμίσεις ομάδων πρακτόρων (κάθε ομάδα εστιάζει σε ένα θέμα). Αυτή η μέθοδος επιταχύνει σημαντικά την εκμάθηση του συστήματος.

Οι παραπάνω μέθοδοι εξηγούν την παραγωγή των πρακτόρων φιλτραρίσματος πληροφοριών. Οι πράκτορες αναζήτησης πληροφοριών παράγονται από την τυχαία ανάθεση σύνταξης ευρετήριο των μηχανών του ιστού σε κάθε πράκτορα. Κάθε πράκτορας αναζήτησης πληροφοριών περιέχει επίσης πληροφορίες για τον αριθμό των λέξεων κλειδιών που παρέχονται στους πράκτορες αναζήτησης από τους πράκτορες φιλτραρίσματος που πραγματικά θα χρησιμοποιηθούν στις ερωτήσεις και πώς αυτές οι ερωτήσεις θα διατυπωθούν από την άποψη των λογικών τελεστών μεταξύ των λέξεων κλειδιών. Οι διαφορετικές μηχανές αναζήτησης έχουν διαφορετικούς τρόπους διαμόρφωσης των ερωτήσεων και επιτρέπουν διαφορετικούς τελεστές σε αυτές (AND, OR, NEARTO κ.λ.π.). Εξαρτάται στη μηχανή που ορίζεται σε κάθε πράκτορα αναζήτησης, να καθορίσει τους σωστούς τελεστές που θα χρησιμοποιηθούν.

7.1.5. Επικοινωνία Client- Server

Το AUI χρησιμοποιεί την τυποποιημένη ***socket επικοινωνία*** με τον Amalthaea server. Όταν το βασισμένο σε java AUI αρχίζει να τρέχει δημιουργεί μια *socket* σύνδεση με τον *server* και στέλνει (σε κρυπτογραφημένη μορφή) την ταυτότητα και τον

κωδικό πρόσβασης του χρήστη. Ο server αποκρίνεται στέλνοντας τις πληροφορίες που έχει συλλέξει για τον συγκεκριμένο χρήστη από την τελευταία φορά που συνδέθηκε, και το AUI τα εμφανίζει.

Η Amalthaea υποστηρίζει τους χρήστες της και με συνεχείς και με *dial-up συνδέσεις* στο δίκτυο. Όλά τα τμήματα του συστήματος γνωρίζουν αυτή τη διάκριση και ενεργούν αναλόγως. Για παράδειγμα εάν ένα site που έχει ελεγχθεί έχει αλλάξει, όταν ο χρήστης συνδέεται στο σύστημα αυτές οι πληροφορίες παρουσιάζονται σε ένα παράθυρο. Εάν ο χρήστης δεν συνδεθεί, οι πληροφορίες είτε αποθηκεύονται ώστε να του τις εμφανίσουν την πρώτη φορά που θα συνδεθεί με το σύστημα, είτε στέλνονται στον χρήστη μέσω του ηλεκτρονικού ταχυδρομείου. Το ίδιο συμβαίνει και με την παρουσίαση των προτεινόμενων *sites*. Μπορούν να εμφανίζονται και να ενημερώνονται συνεχώς ή να παρουσιάζονται στο χρήστη κάθε φορά που εισέρχεται στο σύστημα. Για τους χρήστες που έχουν άμεση και συνεχή σύνδεση στο διαδίκτυο, η Amalthaea ακολουθεί μια ***"background running approach"***. Δηλαδή το AUI τρέχει στην κάτω γωνία της οθόνης του χρήστη και συνεχώς εμφανίζει πληροφορίες και ενημερώνεται. Εάν ο χρήστης ενδιαφέρεται για αυτές τις πληροφορίες που εμφανίζει, ενεργεί αναλόγως εάν όχι αγνοεί το σύστημα.

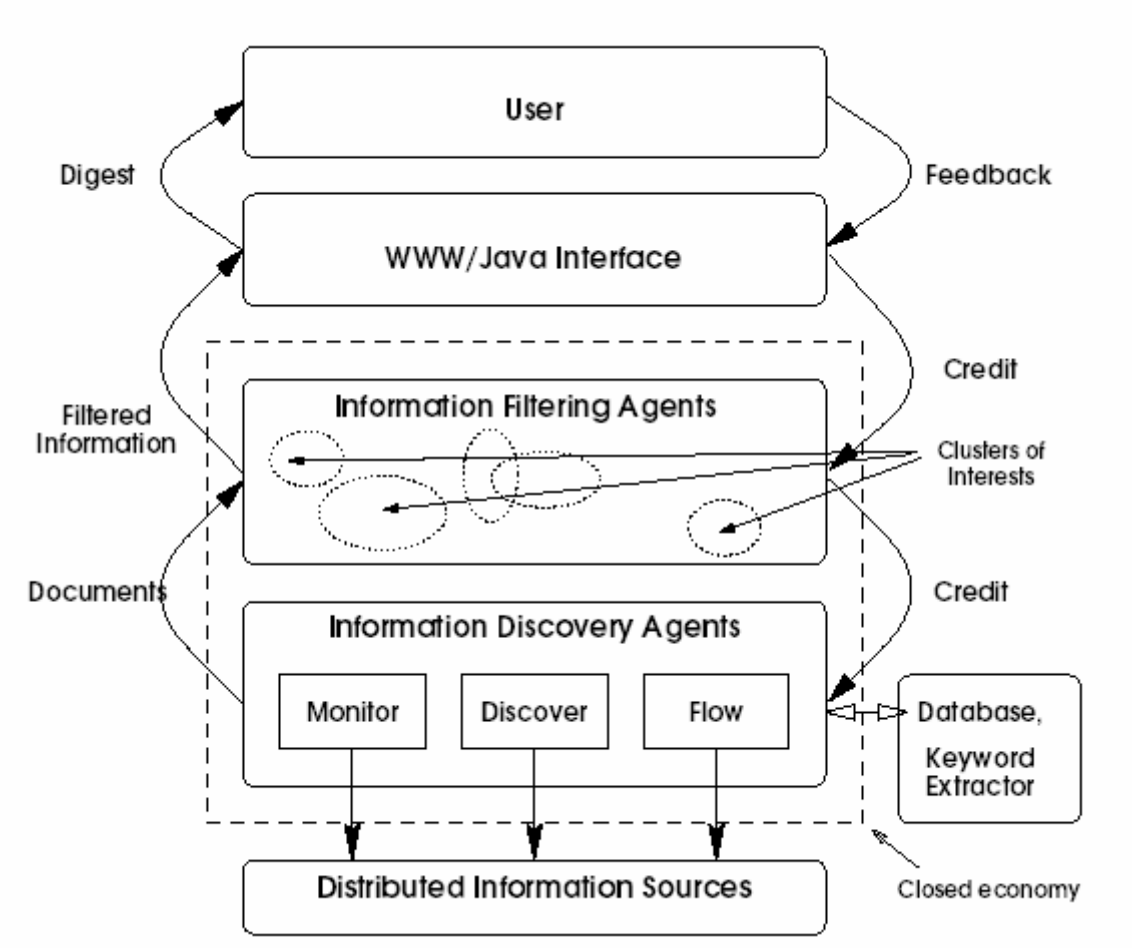
7.1.6. Αρχιτεκτονική της Amalthaea

Η αρχιτεκτονική της Amalthaea ορίζει για κάθε χρήστη τους δικούς του πράκτορες φιλτραρίσματος πληροφοριών και αναζήτησης πληροφοριών δημιουργώντας ένα κλειστό σύστημα. Περισσότεροι από ένα άτομα μπορούν να χρησιμοποιήσουν το σύστημα, αλλά όλα τα αρχεία τους είναι χωριστά. Καμία αλληλεπίδραση μεταξύ των διαφορετικών χρηστών δεν πραγματοποιείται στο σύστημα. Όλα τα στοιχεία του συστήματος λειτουργούν σε έναν χρήστη. Για το χειρισμό πολλαπλών χρηστών το σύστημα κάνει χρήση των πολλαπλών υποδειγμάτων των στοιχείων.

Η Amalthaea αποτελείται από τα ακόλουθα μέρη (Σχήμα 7.4.):

- ✓ Το User Interface, όπου εμφανίζονται στο χρήστη οι ανακτημένες πληροφορίες και δίνει ανατροφοδότηση της σχετικότητάς του.
- ✓ Δύο ευδιάκριτοι τύποι πρακτόρων: Οι *πράκτορες φιλτραρίσματος πληροφοριών* ***(Information Filtering Agents)***, οι *πράκτορες αναζήτησης πληροφοριών* ***(Information Discovery Agents)*** και μηχανισμοί που υποστηρίζουν την κατανομή των πόρων και την εξέλιξη.

- ✓ Η μηχανή ανάκτησης εγγράφων από το WWW.
- ✓ Η μηχανή που επεξεργάζεται τα έγγραφα και αναζητά τις ρίζες των λέξεων ***(stemming)*** καθώς και την απόδοση βάρους σε κάθε όρο, ώστε να παραχθούν τα διανύσματα των λέξεων κλειδιών.
- ✓ Μια βάση δεδομένων με τις ανακτημένες URLs των εγγράφων.



Σχήμα 7.4.: Η Αρχιτεκτονική της Amalthaea

7.2. LawBot

7.2.1. Εισαγωγή

Το LawBot είναι ένα σύστημα *βασισμένο σε πράκτορες του Internet* ***(Internet-based Agents)*** που βοηθούν στη συλλογή και στην οργάνωση των καταστατικών και των ιστορικών στοιχείων, σχετικά με μια αναζήτηση πάνω σε θέματα νομικής. Το σύστημα περιλαμβάνει οντολογία που περιγράφει την αντίστοιχη νομική ορολογία, απλουστεύοντας κατά συνέπεια τη χρήση του συστήματος από ανθρώπους που δεν γνωρίζουν το νομικό επάγγελμα.

Οι ευφυείς πράκτορες επεκτείνονται σε διαφορετικές περιοχές εφαρμογών. Οι πράκτορες αναπτύχθηκαν για να βοηθήσουν τους χρήστες στην επεξεργασία των πληροφοριών. Τα *Shopbots* είναι ένα τέτοιο συνηθισμένο παράδειγμα. Το LawBot παρέχει βοήθεια στους νομικούς ερευνητές με σκοπό την εύρεση των ηλεκτρονικών εγγράφων που είναι αποθηκευμένα σε διάφορες βάσεις δεδομένων που διατηρούνται από τοπικές αρχές, κρατικές αρχές και ομοσπονδιακές. Το LawBot εφαρμόζεται ως συλλογή των πρακτόρων που είναι υιοθετημένα σύμφωνα με τις προτιμήσεις του χρήστη να συλλέγουν, να φιλτράρουν, να οργανώνουν, και να προτείνουν ιστορικά στοιχεία και νόμους σχετικά με μια συγκεκριμένη αναζήτηση. Ο στόχος του LawBot είναι η δημιουργία ενός συστήματος που να μπορεί να χρησιμοποιηθεί αποτελεσματικά όχι μόνο από τους δικηγόρους και τους άλλους νομικούς επαγγελματίες, αλλά και από άλλους ανθρώπους.

7.2.2. Νομική Έρευνα

Οι νομικές πληροφορίες αποτελούνται από δύο κατηγορίες εγγράφων:

- ***Οι νόμοι (Laws)*** είναι αφηρημένες δηλώσεις των δικαιωμάτων, των προνομίων, των υποχρεώσεων και των απαγορεύσεων που ισχύουν για τους πολίτες μέσα σε ένα έθνος ή ένα κράτος.
- ***Οι Γνώμες*** είναι οι εφαρμογές ενός ή περισσότερων νόμων σε μια συγκεκριμένη κατάσταση γεγονότων.

7.2.3. Ηλεκτρονικού Τύπου Αλλαγές

Η επεξεργασία του κειμένου κατέστησε δυνατή την ηλεκτρονική αποθήκευση των παραδοσιακών και άλλων νομικών εγγράφων. Διάφορες ιδιωτικές επιχειρήσεις,

όπως Westlaw (http://westlaw.com/), Lexis (http://www.lexis.com/), και Law Office Information Systems ή LOIS (http://www.loislaw.com/) έχουν συλλέξει ή αναδημιουργήσει μεγάλες βάσεις δεδομένων αυτών των εγγράφων. Χρεώνουν την πρόσβαση σε αυτές είτε με βάση ένα συγκεκριμένο ποσό ή με βάση το χρόνο σύνδεσης σε αυτές.

Εκτός από τη δυνατότητα της άμεσης πρόσβασης σε περισσότερα έγγραφα από αυτά που θα μπορούσαν να αποθηκευτούν σε μια ιδιωτική βιβλιοθήκη, οι μηχανές αναζήτησης παρέχουν την αναζήτηση με βάση όλο το κείμενο. Το ευρετήριο που παράγεται από μια μηχανή αναζήτησης μπορεί να έχει περισσότερες καταχωρήσεις από ένα χειρωνακτικό ευρετήριο. Στην πραγματικότητα περιέχει μια αναφορά σε κάθε λέξη, σε κάθε έγγραφο της βάσης δεδομένων. Για εκείνους τους ερευνητές που είναι σε θέση να τη χρησιμοποιήσουν αποτελεσματικά, η αναζήτηση με βάση το πλήρες κείμενο έχει βοηθήσει σημαντικά στη διαδικασία της νομικής έρευνας με την αφαίρεση της ανάγκης ενός ευρετηρίου εννοιών και φέρνοντας τη διαδικασία ένα βήμα πιο κοντά στο υλικό που αναζητάται.

Τρεις άλλοι παράγοντες αναδιαμορφώνουν το μέλλον της ηλεκτρονικής νομικής έρευνας:

- Το κόστος της αποθήκευσης των εγγράφων έχει μειωθεί εντυπωσιακά, καθιστώντας την ηλεκτρονική αποθήκευση φθηνότερη απ' ό,τι της αποθήκευσης με χαρτί.
- Οι ισχυροί αλλά και ανέξοδοι *κεντρικοί υπολογιστές* ***(servers)*** έχουν γίνει ευρέως διαθέσιμοι, μειώνοντας το κόστος πρόσβασης των εγγράφων μέσω μιας μηχανής αναζήτησης.
- Το διαδίκτυο έχει απλοποιήσει και διευκολύνει την πρόσβαση στα μακρινά έγγραφα.

Αυτοί οι παράγοντες γίνονται όλο και σημαντικότεροι λόγω της ανάπτυξης της αναζήτησης στην τεχνολογία του ***browsing***. Λαμβάνοντας στο σύνολο, αυτές οι τεχνολογίες υποστηρίζουν μια γρήγορα επεκτάσιμη σφαιρική βάση δεδομένων των νομικών εγγράφων.

Δεδομένου ότι οι κυβερνήσεις δημοσιεύουν όλο και περισσότερο τα νομικά έγγραφά τους ηλεκτρονικά και είναι διαθέσιμα στο διαδίκτυο, είναι δυνατό για καθέναν με πρόσβαση στο διαδίκτυο να αναζητήσει και να εξετάσει νομικά έγγραφα χωρίς την απαίτηση της πρόσβασης σε ειδικές βιβλιοθήκες. Ενώ οι νέες ηλεκτρονικές μέθοδοι αναζήτησης δεν θα μπορέσουν ποτέ εντελώς να αντικαταστήσουν την παραδοσιακή

νομική αναζήτηση, είναι ήδη σε χρήση από τους επαγγελματίες και τους μη επαγγελματίες και αναμφισβήτητα θα συνεχίσουν να αυξάνονται.

Οι νομικές υπηρεσίες διαδικτύου όπως FindLaw (http://findlaw.com) και MoreLaw (http://www.morelaw.com) έχουν αναπτυχθεί για να παρέχουν ελεύθερη πρόσβαση στους νομικούς πόρους. Επειδή αυτές οι υπηρεσίες αναπτύχθηκαν ανεξάρτητα, η σύνταξη της ερώτησης και οι βάσεις δεδομένων ποικίλλουν από πεδίο σε πεδίο. Ένας νομικός ερευνητής που έχει πρόσβαση σε διάφορα πεδία πρέπει να μάθει πώς να χρησιμοποιεί τις διάφορες μηχανές αναζήτησης. Η βασισμένη στους πράκτορες τεχνολογία μπορεί να βοηθήσει στην επίλυση αυτού του προβλήματος με την παροχή ενός interface του χρήστη στα συστήματα. Οι πράκτορες μπορούν να επεκταθούν για να μεταφράσουν τα ερωτήματα σε διανεμημένες ερωτήσεις, μπορούν να επικοινωνούν μέσω του διαδικτύου με τις διάφορες μηχανές αναζήτησης χρησιμοποιώντας την κατάλληλη σύνταξη κάθε μηχανής και συγκεντρώνουν, ταξινομούν και τοποθετούν κατάλληλα τα έγγραφα που ανακτήθηκαν από πολλές διανεμημένες ερωτήσεις. Κατά συνέπεια, η τεχνική πολυπλοκότητα μιας εκτενούς νομικής αναζήτησης μπορεί να κρυφτεί από τον ερευνητή, που του επιτρέπει να επικεντρωθεί στη λογική ανάλυση της ίδιας της έρευνας.

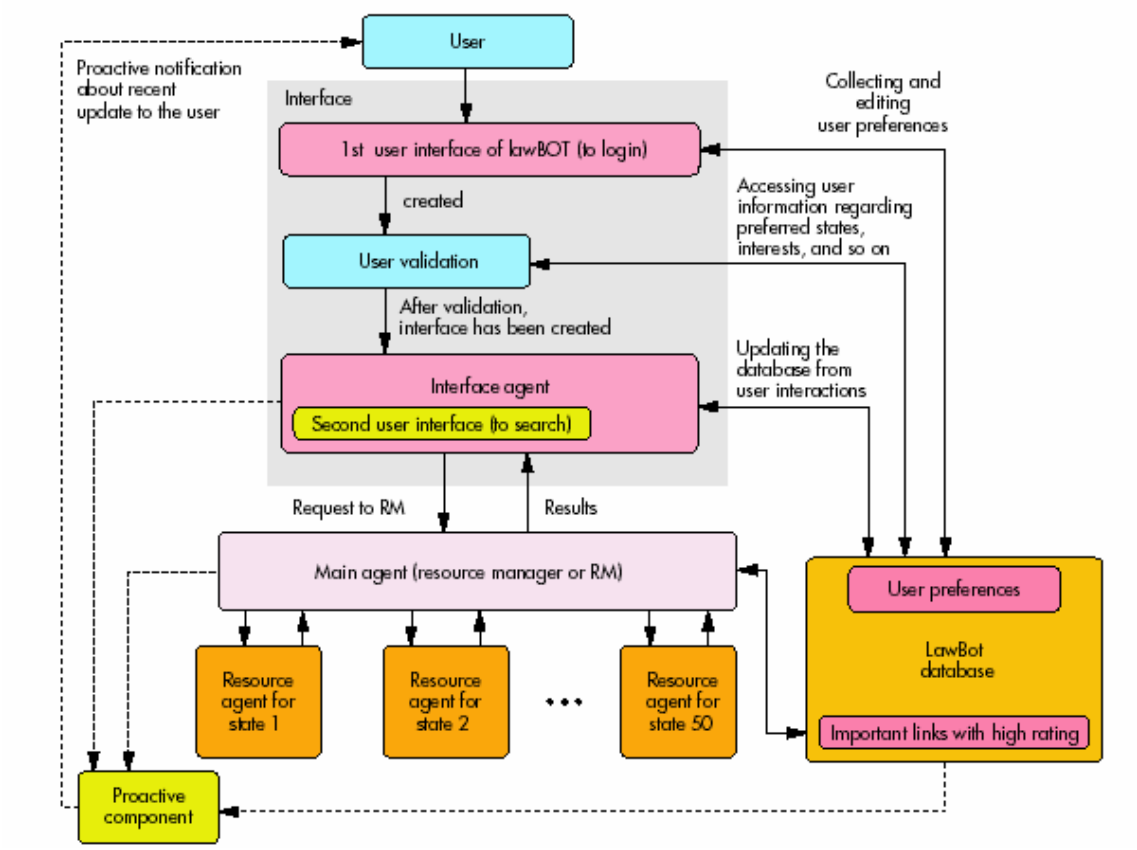
7.2.4. Αρχιτεκτονική της LawBot

Το LawBot σχεδιάστηκε με σκοπό να παρέχει αυτή τη λειτουργία και να διευκολύνει τη νομική έρευνα μέσω του διαδικτύου. Η αρχιτεκτονική του συστήματος παρουσιάζεται στο σχήμα 7.5. Περιλαμβάνει πέντε μέρη με την ακόλουθη λειτουργία:

Interface: Η πρώτη οθόνη χρησιμοποιείται για να έχει πρόσβαση κάποιος στο σύστημα μέσω των *user profiles* που αποθηκεύονται μέσα στη βάση δεδομένων της LawBot. Τα προφίλ περιλαμβάνουν τις προτιμήσεις του χρήστη και πρόσφατα αποτελέσματα αναζήτησης. Μετά από την επικύρωση, ο ***interface agent*** χρησιμοποιεί τα προφίλ για να παρουσιάσει μια προσαρμοσμένη δεύτερη οθόνη, μέσω της οποίας οι χρήστες μπορούν να υποβάλουν νέες τροποποιημένες πρόσφατες αναζητήσεις, αναζητήσεις προς συγκεκριμένα κράτη, κ.λ.π. Ο *interface agent* μεταβιβάζει τις ερωτήσεις των χρηστών, και εμφανίζει τα αποτελέσματα από τον ***resource manager***.

Resource Manager: Ο *resource manager* παίζει τον σημαντικό ρόλο της οργάνωσης, της ανάκτησης και της επεξεργασίας των πληροφοριών. Αυξάνει την ερώτηση του χρήστη με τη συμβολή των οντολογικών πεδίων και έπειτα την αναθέτει

στους σχετικούς *resource agents* όπως απαιτείται από την ερώτηση. Ο *resource manager* συντάσσει πληροφορίες που ανακτώνται από τους *resource agents* μαζί με αυτές που είναι αποθηκευμένες στην υπάρχουσα βάση της LawBot, και τις επιστρέφει στον *interface agent*. Ο *resource manager* μπορεί επίσης να χρησιμοποιήσει ένα *δυναμικό στοιχείο* ***(proactive component)*** από τον *interface agent* για να ειδοποιήσει τους χρήστες για τυχόν πρόσφατες αλλαγές σε προηγούμενα αποτελέσματα αναζήτησης εάν ζητηθούν. Αυτό το χαρακτηριστικό διατηρεί το χρήστη ενήμερο καθώς νέοι νόμοι και περιπτώσεις διατίθενται.



Σχήμα 7.4.: Η Αρχιτεκτονική του LawBot

Resource agents: Ο χρήστης επιλέγει τις αρμοδιότητες που θα αναζητηθούν, οι οποίες καθορίζουν τον *Resource agent ή agents* που θα λάβουν τα αιτήματα αναζήτησης. Οι *Resource agents* είναι προγραμματισμένοι να έχουν πρόσβαση σε συγκεκριμένα *sites* στον ιστό και να διαμορφώνουν τα αιτήματα σύμφωνα με την προβλεπόμενη σύνταξη των μηχανών αναζήτησης που βρήκαν το συγκεκριμένο site. Ένας *αλγόριθμος φιλτραρίσματος λέξεων* ***(word-filtering algorithm)*** επιλέγει και ταξινομεί τα έγγραφα με βάση την πυκνότητα των λέξεων. Τα ταξινομημένα έγγραφα

επιστρέφονται στον *resource manager* χρησιμοποιώντας μια γλώσσα επικοινωνίας μεταξύ των πρακτόρων.

Η βάση δεδομένων της LawBot: Ο *resource manager* μπορεί να ξεκαθαρίσει τις ερωτήσεις και να ταξινομήσει τα αποτελέσματα μέσω της πρόσβασής του στις προτιμήσεις των χρηστών, στις οντολογίες πεδίων, και στις βοηθητικές πληροφορίες που συχνά χρησιμοποιούν συνδέσεις που αποθηκεύονται στη βάση δεδομένων της LawBot.

Δυναμικό στοιχείο(proactive component): Η LawBot περιλαμβάνει ένα interface που μπορεί να υιοθετήσει τις προτιμήσεις των χρηστών και να απαντήσει σε ερωτήσεις του παρελθόντος εκτελώντας αναζήτηση χωρίς να είναι απαραίτητη η σύνδεση του χρήστη. Εάν η *offline* αναζήτηση προσδιορίσει νέα έγγραφα με υψηλή σχετικότητα, το *proactive component* ενημερώνει τον χρήστη για τη διαθεσιμότητά τους.

ΚΕΦΑΛΑΙΟ 8

ΠΡΟΤΥΠΟ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

Το World Wide Web παρουσιάζει έναν τεράστιο αριθμό πηγών πληροφορίας, ο οποίος αυξάνεται με πρωτοφανή ρυθμό. Για να γίνει αποτελεσματική η χρήση αυτού του πλούτου των πληροφοριών, οι χρήστες χρειάζονται διάφορους τρόπους για να εντοπίσουν τις πληροφορίες που τους ενδιαφέρουν. Οι τρόποι αυτοί είναι οι διάφορες *μηχανές αναζήτησης* ***(General Purpose Search Engines)***, *οι μηχανές πολλαπλής αναζήτησης* ***(Multiple Search Engines)***, *τα θεματικά ευρετήρια ή θεματικοί κατάλογοι* ***(Subject Directories- Subject Catalogues)*** και οι *μηχανές αναζήτησης με βάση το προφίλ του χρήστη* ***(User Profiling Search Engines)***, οι οποίοι έχουν περιγραφεί αναλυτικά στα προηγούμενα κεφάλαια της εργασίας.

Στο παρελθόν λίγα χρόνια πριν, η πλοήγηση της γνώσης, τα εργαλεία αναζήτησης των πόρων και οι μηχανές αναζήτησης που υπήρχαν είχαν αποκτήσει, κερδίσει την ευρεία αποδοχή στο διαδίκτυο. Σήμερα όμως λόγω του τεράστιου όγκου του διαδικτύου και της πάρα πολύ μεγάλης πολυπλοκότητάς του δεν είναι εύκολο για τα εργαλεία αναζήτησης να μπορούν να ανταποκρίνονται σε κάθε ερώτηση του χρήστη. Κάποια εργαλεία έχουν ευρετηριάσει όλες σχεδόν τις πληροφορίες του ιστού για μια συγκεκριμένη κατηγορία θεμάτων και συμπεριφέρονται πολύ καλά για ερωτήσεις που αφορούν αυτές τις κατηγορίες. Βέβαια θα πρέπει να τονιστεί ότι η συμπεριφορά τους για θέματα γενικής φύσεως δεν είναι η καλύτερη, κάτι που είναι αναμενόμενο. Από την άλλη πλευρά υπάρχουν εργαλεία αναζήτησης που στοχεύουν στον πλήρη ευρετηριασμό

του ιστού και να μπορούν να απαντούν ικανοποιητικά στο κάθε ερώτημα του χρήστη. Αυτά τα εργαλεία δεν καλύπτουν μόνο συγκεκριμένες κατηγορίες πεδίων.

Γενικά η δυνατότητα αναζήτησης και ανάκτησης πληροφοριών στον ιστό αποδοτικά και αποτελεσματικά είναι μια τεχνολογία που επιτρέπει την εκμετάλλευση όλων των δυνατοτήτων των εργαλείων αναζήτησης. Οι μηχανές αναζήτησης είναι ισχυρά εργαλεία για τον ανεξέλεγκτο στόχο της πλοήγησης του γρήγορα επεκτάσιμου World Wide Web.

8.1. Ανάπτυξη Ενός Μοντέλου-Προτύπου Ενιαίας Αναζήτησης

Έχοντας αναλύσει στα προηγούμενα κεφάλαια της παρούσας εργασίας τα διάφορα χαρακτηριστικά και τις τεχνικές που χρησιμοποιούν, στην ενότητα αυτή θα γίνει μια ολική σύγκριση και αξιολόγηση όλων των εργαλείων αναζήτησης του ιστού, με σκοπό την ανάπτυξη ενός μοντέλου ενιαίας αναζήτησης.

Η σύγκριση των παραπάνω εργαλείων αναζήτησης πραγματοποιείται για κάθε σημαντικό τμήμα αυτών. Οι μηχανές αναζήτησης μπορεί να θεωρηθούν ότι αποτελούνται γενικά από πέντε κύρια μέρη, όπως αναφέρθηκε στην ενότητα 1.3.3. (Randolph Hock, 2001):

- *Το Ειδικό Λογισμικό* ***(robot, spider, crawler κ.λ.π.)***
- *Η Βάση Δεδομένων ή αλλιώς ο Κατάλογος* ***(database of information)***
- *Το Πρόγραμμα Ευρετηρίασης και το Ευρετήριο* ***(the indexing program and the index)***
- *Η Μηχανή Ανάκτησης, Μηχανή Αναζήτησης, το Ειδικό Πρόγραμμα* ***(retrieval engine)*** και
- *Η Γραφική Διεπαφή* ***(the graphical HTML (Hyper-text Markup Language) interface)***

Κάθε ένα από τα παραπάνω τμήματα περιλαμβάνεται στη σύγκριση καθώς το πρότυπο πρέπει να αποτελείται από αυτά τα μέρη. Αρχικά θα αναπτυχθεί το πρότυπο της γραφικής διεπαφής ***(the graphical HTML (Hyper-text Markup Language) interface)*** ή αλλιώς του ***User Interface***. Έπειτα θα αναπτυχθεί το πρότυπο του ειδικού λογισμικού ***(robot, spider, crawler κ.λ.π.)*** δηλαδή το ***crawling***. Περαιτέρω θα μελετηθεί η *βάση δεδομένων ή αλλιώς ο κατάλογος* ***(database of information)*** για το σύνολο των εργαλείων αναζήτησης. Στη συνέχεια μελετάται το πρότυπο του *προγράμματος ευρετηρίασης και το ευρετήριο* ***(the indexing program and the index)***. Τέλος θα αναπτυχθεί η *μηχανή ανάκτησης, μηχανή αναζήτησης, το ειδικό πρόγραμμα* ***(retrieval engine)*** που είναι υπεύθυνη για την ευρετηρίαση της βάσης δεδομένων της εκάστοτε μηχανής και την εμφάνιση των αποτελεσμάτων ταξινομημένα στο χρήστη. Αυτές είναι δύο διαδικασίες που δεν είναι ευδιάκριτες στον χρήστη και δεν πρέπει να είναι. Άρα θα αναπτύξουμε και το πρότυπο της διάταξης και ταξινόμησης των σελίδων αποτελεσμάτων ***(Ranking)***.

Το πρότυπο αναπτύχθηκε με βάση τις πληροφορίες που συλλέχθησαν στα προηγούμενα κεφάλαια της παρούσας εργασίας για κάθε εργαλείο ξεχωριστά. Οι πληροφορίες αυτές αλλά και άλλα χαρακτηριστικά των εργαλείων αναζήτησης βρίσκονται συγκεντρωμένα στα παραρτήματα Ε και ΣΤ στο τέλος της εργασίας. Για τις μηχανές πολλαπλής αναζήτησης υπάρχουν συγκεντρωμένα τα χαρακτηριστικά τους στον πίνακα 5.4. Το πρότυπο αναπτύχθηκε με βάση τα 26 εργαλεία αναζήτησης που υπάρχουν στους παραπάνω πίνακες, τα οποία είναι:

- ✓ ***Οι Μηχανές Αναζήτησης:*** AltaVista, Google, Hotbot, Infoseek, Lycos, AllTheWeb, Teoma, Aol Search, Northernlight, Direct Hit, Inktomi,
- ✓ ***Τα Θεματικά Ευρετήρια:*** Yahoo!, LookSmart, MSN, Excite και
- ✓ ***Οι Μηχανές Πολλαπλής Αναζήτησης:*** Metacrawler, Ixquick, ProFusion, Dogpile, SavvySearch, Search.com, Cyber411, Highway61, Ask Jeeves, Mamma, SurfWax.

8.1.1. Το Πρότυπο της Γραφικής Διεπαφής (The Graphical HTML (Hyper-text Markup Language) Interface)

Εξετάζοντας προσεκτικά τα *interfaces* των διάφορων μηχανών αναζήτησης παρατηρούμε ότι υπάρχουν μεγάλες διαφορές αλλά και πολλές ομοιότητες. Είναι γεγονός ότι οι αρχικές σελίδες ***(Homepages)*** των θεματικών ευρετηρίων είναι γεμάτες με πολλές συνδέσεις. Υπάρχουν συνδέσεις για τα αντιπροσωπευτικά πεδία (κατηγορίες), καθώς και συνδέσεις με *ειδήσεις* ***(news)*** και με εικόνες ***(pictures)***. Από την άλλη πλευρά υπάρχουν *homepages* που ακολουθούν μια μινιμαλιστική μέθοδο. Αυτές οι αρχικές σελίδες εκτός από το *κουτί της αναζήτησης* ***(search box)*** δεν περιέχουν αγγελίες, κινούμενα αρχεία GIF, ή ***applets της Java*** που να αποσπούν τους χρήστες. Τέτοια *interfaces* έχουν οι Google, Ixquick, κ.τ.λ. Ο Larry Page, ένας από τους δημιουργούς της Google, ίσως της δημοφιλέστερης μηχανής αναζήτησης, είπε ότι η αρχική σελίδα της Google δεν είναι πολύ φορτωμένη με συνδέσεις, επειδή η έρευνα δείχνει ότι ο χρόνος που χρειάζεται για να ολοκληρωθεί μια εργασία είναι ανάλογος με τον αριθμό των διαθέσιμων επιλογών (η υπόθεση είναι ότι είναι περισσότερο χρονοβόρα μια αναζήτηση σε μια αρχική σελίδα με πολλές συνδέσεις από ό,τι σε μια αρχική σελίδα με λίγες). Φαίνεται ότι ο David Bodnick, ο σχεδιαστής της Ixquick, συμφωνεί με τον Page.

Βέβαια πέρα από το κουτί της αναζήτησης οι περισσότερες μηχανές πολλαπλής αναζήτησης καθώς και οι μηχανές αναζήτησης παρέχουν ένα *link* στην αρχική σελίδα

που οδηγεί στην προχωρημένη αναζήτηση (για ανθρώπους που τους αρέσουν οι πολλές επιλογές), επιτρέποντας στους χρήστες να καθορίσουν τις περιοχές αναζήτησης, τον αριθμό των εγγράφων που ανακτώνται, ή τον αριθμό των σελίδων που απαριθμούν από κάθε περιοχή. Αυτό είναι ένα πάρα πολύ σημαντικό χαρακτηριστικό για κάθε μηχανή αναζήτησης.

Επίσης στην αρχική σελίδα των περισσότερων μηχανών αναζήτησης υπάρχουν σύνδεσμοι υπερκειμένου σε σελίδες με πληροφορίες για την κάθε μηχανή αναζήτησης, και σελίδες της μηχανής αναζήτησης για άλλες γλώσσες, αλλά πολύ λίγο αποσπούν την προσοχή του χρήστη. Τέλος παίζει πολύ σημαντικό ρόλο για τους χρήστες και η ταχύτητα εμφάνισης της αρχικής σελίδας. Αυτή εξαρτάται από το πόσο φορτωμένη είναι η σελίδα αυτή.

Από τα παραπάνω γίνεται φανερό ότι το ***πρότυπο Interface*** είναι αυτό που έχουν καθιερώσει οι σημαντικότερες μηχανές αναζήτησης, όπως η Google. Το interface αυτό περιέχει το κουτί της αναζήτησης και ένα link που συνδέεται με την προχωρημένη αναζήτηση, ώστε να μπορεί ο χρήστης να καθορίσει με περισσότερη λεπτομέρεια την αναζήτηση που θα πραγματοποιήσει.

8.1.2. Το Πρότυπο του Ειδικού Λογισμικού (Crawling)

Το να αναπτυχθεί ένα πρότυπο ειδικού λογισμικού είναι πάρα πολύ δύσκολο με τα σημερινά μεγέθη του ιστού και με τις σημερινές ανάγκες των χρηστών. Το ειδικό λογισμικό κάθε μηχανής αναζήτησης είναι διαφορετικό και μπορεί να διαφέρει από το πόσο συχνά επισκέπτεται τον ιστό, στον αριθμό των crawlers που χρησιμοποιεί μέχρι και την αρχιτεκτονική του. Είναι ιδιαίτερα απίθανο μια μηχανή να χρησιμοποιεί το ίδιο ειδικό λογισμικό και να πραγματοποιεί με τον ίδιο τρόπο το crawling, κάτι το οποίο έγινε φανερό από τα προηγούμενα κεφάλαια της εργασίας όπου περιγράφηκε αναλυτικά ο τρόπος του crawling για κάθε εργαλείο αναζήτησης.

Στον παραπάνω Πίνακα 8.1. φαίνεται το χρονικό διάστημα που ο crawler της κάθε μηχανής αναζήτησης επισκέπτεται τη σελίδα μετά την υποβολή της για ευρετηρίαση και το χρονικό διάστημα που θα την επισκεφτεί για τυχόν αλλαγές μετά την πρώτη ευρετηρίασή της. Επίσης για ορισμένα εργαλεία αναζήτησης έχουν τοποθετηθεί και διάφορα χαρακτηριστικά του ειδικού λογισμικού που δείχνει πόσο διαφορετικό μπορεί να είναι μεταξύ των εργαλείων.

Εργαλείο Αναζήτησης	Επίσκεψη της σελίδας μετά την αίτηση υποβολής της	Επίσκεψη της σελίδας για τυχόν αλλαγές της	Αριθμός των συνδέσεων που χειρίζεται ο spider	Πόσο χρόνο ξοδεύει σε κάθε σελίδα
AltaVista	1 μήνα	4-6 εβδομάδες		
Infoseek	1 εβδομάδα	2 μήνες		
Excite	21 ημέρες			30 sec
Google			300 συνδέσεις	
Yahoo!	1-2 μήνες			
Lycos	2 εβδομάδες			
Infoseek	Σε πραγματικό χρόνο ευρετήριο			
Open Text	1 μήνα			
WebCrawler	Κάτι περισσότερο από μήνα			

Πίνακας 8.1.: Η σύγκριση των εργαλείων αναζήτησης με βάση τα χαρακτηριστικά του ειδικού λογισμικού τους

Ο παραπάνω Πίνακας 8.1. θα μπορούσε να γίνει πάρα πολύ μεγάλος περιέχοντας πολλά διαφορετικά χαρακτηριστικά για το crawling των εργαλείων αναζήτησης. Αυτά τα ιδιαίτερα χαρακτηριστικά περιέχονται στα κεφάλαια 4, 5 και 6 της παρούσας εργασίας καθώς και στην ενότητα 3.3.3.5. του κεφαλαίου 3.

Μια άλλη σκοπιά με την οποία μπορούμε να ερευνήσουμε το crawling των εργαλείων αναζήτησης και να προσπαθήσουμε να αναπτύξουμε το πρότυπο είναι οι διάφοροι τρόποι crawling που εφαρμόζονται από τα παραπάνω εργαλεία. Παρακάτω υπάρχουν οι διάφοροι τρόποι του crawling, οι οποίοι έχουν αναπτυχθεί αναλυτικά στην ενότητα 4.7. της παρούσας εργασίας:

- *Εξαντλητικό Crawl (Deep Crawl)*
- *Άμεση Ευρετηρίαση (Instant Indexing)*
- *Υποστήριξη Πλαισίων (Frames Support)*
- *Καταγραφή Εικόνων (Image Maps)*
- *Robots.txt*
- *Meta Robots Tag*
- *Το Πλήθος των Συνδέσεων Βοηθά στο Εξαντλητικό Crawl (Link Popularity Helps Deep Crawl)*

- *Συχνότητα Αλλαγών της Σελίδας (Learns Frequency)*

Στον παρακάτω Πίνακα 8.2. πραγματοποιείται για τα περισσότερα εργαλεία αναζήτησης αυτή η σύγκριση του τρόπου του crawling. Όλες οι πληροφορίες για την κάθε μηχανή ξεχωριστά βρίσκονται στα ***Παραρτήματα Ε και ΣΤ*** της παρούσας εργασίας. Παρατηρούμε ότι και η AltaVista και η Google εφαρμόζουν το *εξαντλητικό Crawl* ***(Deep Crawl)***. Σύμφωνα με το *εξαντλητικό Crawl* θα τοποθετηθούν σε μια λίστα πολλές σελίδες από μια ιστοσελίδα, ακόμα και αν οι σελίδες αυτές δεν έχουν υποβληθεί ρητά σε αυτές. Οι υπόλοιπες μηχανές αναζήτησης θα απαριθμήσουν συνήθως πολύ λιγότερες σελίδες από ένα *site*. Γενικά, όσο μεγαλύτερο είναι το ευρετήριο μιας μηχανής αναζήτησης, το πιθανότερο να έχει πολλές σελίδες ανά *site*.

Crawling	Ναι	Όχι	Παρατηρήσεις
Εξαντλητικό Crawl (Deep Crawl)	AltaVista, Google	Excite, Lycos	
Άμεση Ευρετηρίαση (Instant Indexing)	AltaVista	Excite, Lycos, Google	Οι σελίδες εμφανίζονται σε μια ή δυο μέρες μετά την καταχώρησή τους
Υποστήριξη Πλαισίων (Frames Support)	AltaVista, Google	Excite, Lycos	Lycos παρέχει την ελάχιστη υποστήριξη
Καταγραφή Εικόνων (Image Maps)	AltaVista	Excite, Google, Lycos	
Robots.txt	Όλες	-	
Meta Robots Tag	Όλες	-	
Πλήθος Συνδέσεων που Βοηθά στο Εξαντλητικό Crawl	Lycos	Excite, AltaVista	
Συχνότητα Αλλαγών της Σελίδας	AltaVista	Excite, Google, Lycos	

Πίνακας 8.2.: Η σύγκριση των εργαλείων αναζήτησης με βάση το Crawling (Bikkannavar, 1999)

Επίσης, από τον παραπάνω πίνακα παρατηρούμε ότι όλα τα εργαλεία αναζήτησης

υποστηρίζουν το αρχείο Robots.txt (είναι ένας τρόπος για τους webmasters να κρατούν τις μηχανές αναζήτησης μακριά από τις ιστοσελίδες τους) και το Meta Robots Tag (επιτρέπει στους ιδιοκτήτες ιστοσελίδων να καθορίσουν ότι μια σελίδα δεν πρέπει να ευρετηριασθεί).

Παρατηρούμε ότι η Lycos χρησιμοποιεί *το πλήθος των συνδέσεων που βοηθά στο εξαντλητικό crawl* ***(Link Popularity Helps Deep Crawl)***. Σύμφωνα με αυτή τη μέθοδο καθορίζεται η δημοτικότητα μιας σελίδας με την ανάλυση του αριθμού των *συνδέσεων* ***(links)*** που καταλήγουν σε αυτή από άλλες σελίδες. Μερικές μηχανές χρησιμοποιούν αυτό ως μέσο να καθορίσουν ποιες σελίδες θα συμπεριληφθούν στο ευρετήριο.

Τέλος μια άλλη παράμετρος που πρέπει να ληφθεί υπόψη είναι η *συχνότητα αλλαγών της σελίδας* ***(Learns Frequency)***, δηλαδή πόσο συχνά οι σελίδες αλλάζουν (ανανεώνουν) τα περιεχόμενά τους. Σελίδες που αλλάζουν τα περιεχόμενά τους συχνά, επισκέπτονται περισσότερο. Παραδείγματος χάριν η Excite επίσης τρέχει ένα ***Fresh-Spider,*** το οποίο κάθε εβδομάδα διερευνά δύο εκατομμύρια σημαντικές ιστοσελίδες, όπως ορίζεται από την Excite.

Δεδομένου ότι ο ιστός είναι τόσο δυναμικός από την άποψη των νέων σελίδων που προστίθενται σε αυτόν (και διαγράφονται, επίσης), οι spiders ξαναεπισκέπτονται τις ήδη ευρετηριασμένες σελίδες με σκοπό να έχουν το ευρετήριό τους όσο το δυνατό περισσότερο ενημερωμένο. Στην πραγματικότητα, ένα εργαλείο αναζήτησης μπορεί να θεωρηθεί ότι έχει

- Ένα *spider ενημέρωσης* ***(freshness spider):*** επισκέπτεται τα δύο πρώτα εκατομμύρια σελίδων της μηχανής αναζήτησης μία φορά την εβδομάδα και
- Ένα *spider πληρότητας* ***(completeness spider):*** επισκέπτεται το υπόλοιπο των ευρετηριασμένων σελίδων περίπου μία φορά το μήνα.

Είναι ίσως το πιο σημαντικό χαρακτηριστικό που θα πρέπει να διαθέτει ο spider ή οι spiders, να μπορούν να κρατούν ενημερωμένες τις βάσεις δεδομένων των εργαλείων αναζήτησής τους. Άρα το ***πρότυπο του ειδικού λογισμικού*** ξεκινά με αυτή την βασική αρχή και φυσικά αυτό για κάθε εργαλείο αναζήτησης ποικίλλει και εξαρτάται από το ειδικό λογισμικό του. Το ιδανικό για ένα εργαλείο αναζήτησης (συγκρίνοντας με τις τιμές του πίνακα 8.1.) θα ήταν, το ειδικό λογισμικό να επισκέπτεται μέσα σε 15 μέρες τη σελίδα που έκανε αίτηση υποβολής. Και βέβαια σε 1 μήνα να την ξαναεπισκεφθεί για τυχόν αλλαγές ή πρόσθετα στοιχεία. Το πρότυπο θα πρέπει να περιλαμβάνει και έναν τρόπο καθορισμού της σημαντικότητας της κάθε σελίδας που επισκέπτεται το

ειδικό λογισμικό, έτσι ώστε οι πιο σημαντικές ιστοσελίδες να μπορούν να επισκέπτονται συχνότερα. Επίσης θα πρέπει να υπάρχει και τρόπος καθορισμού των σημαντικών σελίδων που αλλάζουν συχνά, έτσι ώστε να τις επισκέπτεται το ειδικό λογισμικό συχνότερα. Για την πρώτη περίπτωση θα μπορούσε να χρησιμοποιηθεί η τεχνική της Lycos, *το πλήθος των συνδέσεων που βοηθά στο εξαντλητικό crawl* ***(Link Popularity Helps Deep Crawl)***. Για να επιτευχθεί και το δεύτερο θα πρέπει να χρησιμοποιηθούν από το ίδιο το εργαλείο αναζήτησης περισσότεροι από ένας spiders και με δυνατότητα χειρισμού πολλών συνδέσεων ταυτοχρόνως (τεχνική της Google).

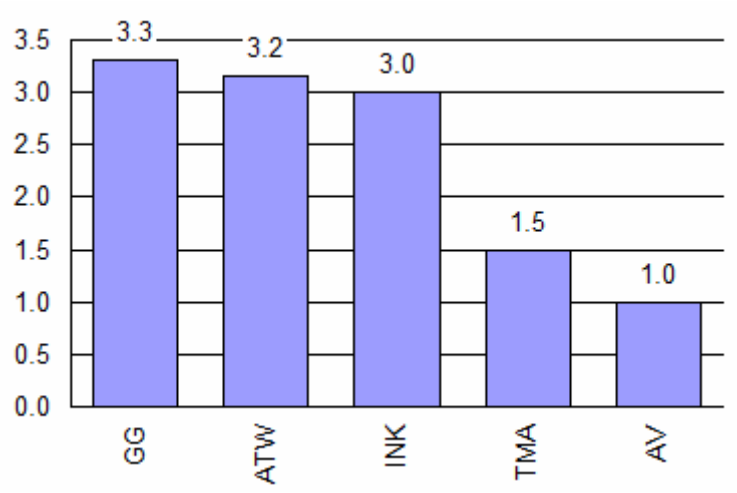
Με βάση τον πίνακα 8.2. και τους πίνακες των παραρτημάτων Ε και ΣΤ παρατηρούμε ότι το ***πρότυπο του ειδικού λογισμικού*** πρέπει να υποστηρίζει το ***Εξαντλητικό Crawl (Deep Crawl)***, καθώς οι τέσσερις από τις εφτά (57%) σημαντικότερες μηχανές αναζήτησης το υποστηρίζουν. Επίσης το αρχείο ***Robots.txt*** και το ***Meta Robots Tag*** πρέπει να υποστηρίζεται από το ***πρότυπο του ειδικού λογισμικού***, αφού όλα τα εργαλεία αναζήτησης (100%) τα υποστηρίζουν.

Επίσης από τον πίνακα του Παραρτήματος Ε που αναφέρεται στο ειδικό λογισμικό προκύπτει ότι η ***αναζήτηση τίτλων*** (5 από τα 15 εργαλεία αναζήτησης) και η ***αναζήτηση URL*** (8 από τα 15 εργαλεία αναζήτησης) πρέπει να υποστηρίζεται από το ***πρότυπο του ειδικού λογισμικού***. Από τον ίδιο πίνακα προκύπτει ότι η αναζήτηση anchor δεν υποστηρίζεται από κανένα εργαλείο αναζήτησης.

8.1.3. Το Πρότυπο της Βάσης Δεδομένων (Database of Information)

Όπως γίνεται εύκολα αντιληπτό, η ακρίβεια των αποτελεσμάτων σε μια ερώτηση του χρήστη εξαρτάται σε πάρα πολύ μεγάλο βαθμό και από το μέγεθος της βάσης δεδομένων της κάθε μηχανής. Είναι γνωστό ότι όταν ο χρήστης θέτει μια ερώτηση στη μηχανή αναζήτησης, η μηχανή δεν αναζητά σε πραγματικό χρόνο τον ιστό, αλλά αναζητά πληροφορίες σχετικές με το ερώτημα που βρίσκονται στη βάση δεδομένων της. Άρα γίνεται κατανοητό ότι όσο μεγαλύτερη βάση δεδομένων έχει η κάθε μηχανή αναζήτησης τόσο περισσότερες πληροφορίες θα έχει αποθηκεύσει και τόσο περισσότερες σελίδες σχετικές με την ερώτηση του χρήστη θα εμφανίσει σε αυτόν.

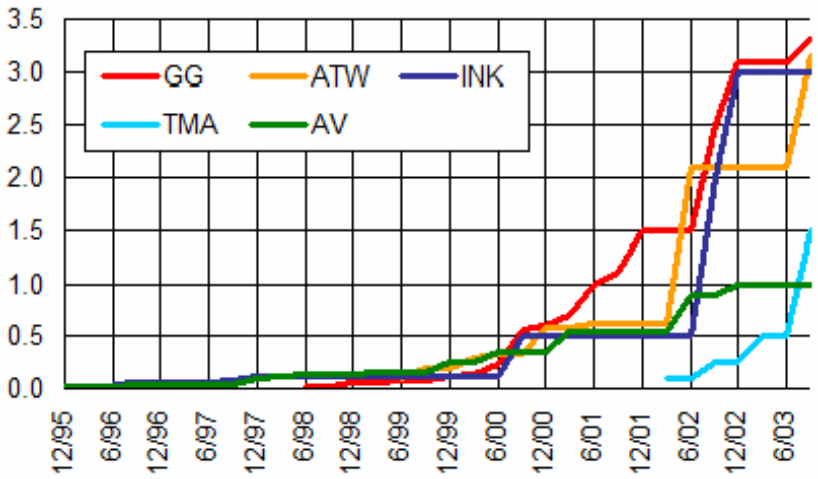
Στις παρακάτω γραφικές παραστάσεις φαίνεται η ραγδαία αύξηση του ευρετηρίου και συνάμα της βάσης δεδομένων της κάθε μηχανής αναζήτησης από τον Δεκέμβριο του 1995 έως και τον Σεπτέμβριο του 2003.



Σχήμα 8.1.: Τα Δισεκατομμύρια των Αρχείων Κειμένου που Ευρετηριάστηκαν τον Σεπτέμβριο του 2003 από τις Κυριότερες Μηχανές Αναζήτησης (SearchEngineWatch.com)

Όπου: ***GG= Google*** ***ATW= AllTheWeb*** ***INK= Inktomi***
TMA= Teoma ***AV= AltaVista***

Το Σχήμα 8.1. παρουσιάζει το μέγεθος των αρχείων (δισεκατομμύρια αρχείων) κειμένων που έχουν ευρετηριαστεί τον Σεπτέμβριο του 2003 από τις σημαντικότερες μηχανές αναζήτησης, τα οποία περιέχουν HTML αρχεία, αρχεία κειμένων, αρχεία PDF, αρχεία του Microsoft Office και άλλα παρόμοια αρχεία. Εικόνες και αρχεία πολυμέσων δεν περιέχονται.



Σχήμα 8.2.: Το Μέγεθος των Δισεκατομμυρίων Αρχείων Κάθε Μηχανής Αναζήτησης από το 1995 έως και το 2003 (SearchEngineWatch.com)

Το παραπάνω Σχήμα 8.2. δείχνει πως το μέγεθος των μηχανών αναζήτησης (δισεκατομμύρια αρχείων) μεταβλήθηκε στη διάρκεια των χρόνων 1995 έως και 2003. Οι παραπάνω μηχανές αναζήτησης είναι αυτές που αναζητούν ακόμη *(crawling)* πληροφορίες στον ιστό.

Στον παρακάτω Πίνακα 8.3. φαίνονται τα κύρια χαρακτηριστικά της βάσης κάθε μηχανής αναζήτησης, δηλαδή τους τύπους των αρχείων που περιέχει καθώς και το μέγεθός τους.

Μηχανές Αναζήτησης	***Database***
Google (google.com)	Πλήρες κείμενο των σελίδων, .pdf, .doc, .xls, .ps, .wpd και άλλα (4,3B από τις οποίες το 1B των URLs μερικώς ευρετηριασμένο) **Συν:** Ειδήσεις που ενημερώνονται συνεχώς (4500 πηγές), Εικόνες και Groups: Usenet
Yahoo! Search (search.yahoo.com)	Πλήρες κείμενο των σελίδων, .pdf, .ps, flash και άλλα (Περισσότερα από 3Billions) **Συν:** Ειδήσεις (7000 πηγές)- Εικόνες-Χάρτες-Άνθρωποι- Yellow Pages- Ταξίδια- Προϊόντα
Teoma (teoma.com)	Πλήρες κείμενο των σελίδων (περίπου 1B)
Ask Jeeves (www.ask.com)	Τα αποτελέσματα τα παίρνει από τη βάση της Teoma. **Συν:** Για αναζήτηση προϊόντων χρησιμοποιεί priceGrabber.com. Για αναζήτηση εικόνων χρησιμοποιεί Picsearch.com. Για αναζήτηση ειδήσεων χρησιμοποιεί Moreover.com.
Πολλαπλές Μηχανές Αναζήτησης	***Database***
Vivisimo (vivisimo.com)	Χρησιμοποιεί τα παρακάτω εργαλεία αναζήτησης: Netscape, MSN, Lycos, LookSmart και άλλες.
Dogpile (dogpile.com)	Χρησιμοποιεί τα παρακάτω εργαλεία αναζήτησης: Yahoo!, LookSmart, Go.com, Lycos, Direct Hit, RealNames, Sprinks, Dogpile Directory, GoTo.com, Open Directory, AltaVista, Google, AskJevesAbout.com.

Θεματικοί Κατάλογοι	***Database***
Librarians' Index to the Internet (lii.org)	Πηγές χρήσιμες στους χρήστες των δημοτικών βιβλιοθηκών, αξιολογημένες και δημιουργημένες από βιβλιοοικονόμους.(περίπου 13K).
Yahoo (dir.yahoo.com)	Υποβαλλόμενες ιστοσελίδες (περίπου 2M).
Infomine (infomine.ucr.edu)	Πολύ προσεκτικά επιλεγμένοι πόροι του διαδικτύου από τις βιβλιοθήκες των πανεπιστημίων (περίπου 100K).
Academic Info (academicinfo.net)	Επίπεδο κολεγίων και αναζήτησης (περίπου 25K).

Πίνακας 8.3: Η σύγκριση των εργαλείων αναζήτησης με βάση τις βάσεις δεδομένων τους (http://infopeople.org/search/chart.html)

Τα στοιχεία του παραπάνω πίνακα ισχύουν για το 2004 και συγκεκριμένα μέχρι τον Απρίλιο του 2004.Από τον παραπάνω πίνακα βλέπουμε ότι απουσιάζει μια πολύ σημαντική μηχανή αναζήτησης, η AltaVista. Τον Μάρτιο του 2004 η AllTheWeb και η AltaVista έπαψαν να είναι διαθέσιμες. Από τη στιγμή που το Yahoo! δημιούργησε το 2004 τη δική του βάση δεδομένων για τις αναζητήσεις του (Yahoo!Search database), οι AllTheWeb και AltaVista ξανασχεδιάστηκαν για να παρέχουν αποτελέσματα παρόμοια με το Yahoo!.

Είναι φανερό από τον παραπάνω πίνακα ότι τα θεματικά ευρετήρια διατηρούν τις μικρότερες βάσεις δεδομένων και αυτό είναι απόλυτα κατανοητό, αφού απευθύνονται σε συγκεκριμένα γνωστικά πεδία. Βέβαια το θεματικό ευρετήριο του Yahoo! έχει ξεπεράσει σε μέγεθος (περίπου 2M) κατά πολύ όλους τους ανταγωνιστές του και ίσως αυτός είναι ο σημαντικότερος λόγος που έχει γίνει τόσο δημοφιλές.

Η Google, που θεωρείται μια από τις καλύτερες μηχανές αναζήτησης, αν όχι η καλύτερη, διατηρεί μια πολύ μεγάλη βάση (4,3 δισεκατομμύρια εγγράφων) και αυτό είναι το μεγάλο της πλεονέκτημα. Από τη στιγμή που το διαδίκτυο μεγαλώνει με ραγδαίους ρυθμούς θα πρέπει και οι μηχανές αναζήτησης να ακολουθήσουν αυτή την αύξηση.

Από τα παραπάνω γίνεται απόλυτα κατανοητό ότι το ***πρότυπο της βάσης δεδομένων μιας μηχανής αναζήτησης*** θα πρέπει να είναι μια μεγάλη βάση δεδομένων, με μέγεθος κοντά στα 4 δισεκατομμύρια έγγραφα και να περιέχει ταυτοχρόνως όλους

τους δυνατούς τύπους των αρχείων που υπάρχουν στον ιστό, όπως .pdf, .doc, .xls, .ps, .wpd και άλλα.

8.1.4. Το Πρότυπο του Προγράμματος Ευρετηρίασης και το Ευρετήριο (The Indexing Program and the Index)

Σε αυτή την ενότητα περιγράφονται εν συντομία οι διάφορες τεχνικές σύνταξης ευρετηρίου που χρησιμοποιούνται από τα διάφορα εργαλεία αναζήτησης και από τα οποία θα προκύψει και το πρότυπο ευρετήριο. Όπως έχει αναφερθεί και στην ενότητα 3.3.3.2. για τη δημιουργία ευρετηρίων ακολουθούνται δύο διαφορετικές τεχνικές:

1. ***Χειρωνακτική (manual)*** δημιουργία ευρετηρίου, που πραγματοποιείται από ανθρώπους. Αν και δεν είναι τέλεια θεωρούνται περισσότερο ακριβή, καθώς οργανώνονται από ειδικούς με βάση διάφορα δημοφιλή θεματικά αντικείμενα και συντάσσονται με τέτοιο τρόπο, ώστε να διευκολύνουν τη διαδικασία αναζήτησης.
2. ***Αυτόματη (automatic)*** δημιουργία ευρετηρίου, που πραγματοποιείται με τη βοήθεια ειδικών προγραμμάτων ή *ευφυών πρακτόρων* ***(intelligent agents)***. Γνωστά με την ονομασία ***robots, spiders, wanderers, Web walkers ή και Web agents***, τα προγράμματα αυτά κινούνται συνεχώς στο διαδίκτυο, επισκέπτονται τη μια ιστοσελίδα μετά την άλλη, συλλέγουν πληροφορίες και δημιουργούν τα ευρετήρια όπου καταχωρούν το περιεχόμενο του Web.

Υπάρχουν γενικά δύο είδη αυτόματα δημιουργημένων ευρετηρίων:

- ✓ Το *σταθμικό* ***(weighted)*** στο οποίο αποδίδεται στους όρους ένα βάρος ανάλογα με τη συχνότητά τους εντός του εγγράφου και
- ✓ Το *μη-σταθμικό* ***(unweighted)*** στο οποίο κάθε όρος αποθηκεύεται με μία τιμή, η οποία περιγράφει την τοποθεσία του και λίγες ή καθόλου περαιτέρω πληροφορίες.

Βέβαια το **σταθμικό ευρετήριο** αποτελείται από διάφορα μοντέλα όπως:

- ➢ Το ***Vector Space Model*** είναι μία συνηθισμένη προσέγγιση *IR* του σταθμικού ευρετηρίου και της επακόλουθης ανάκτησης. Μία εκδοχή αυτής της προσέγγισης χρησιμοποιείται από τη μηχανή αναζήτησης Excite, ως μέρος της διαδικασίας ***Intelligent Concept Extraction (ICE)***.
- ➢ Το *πιθανοθεωρητικό μοντέλο* ***(probabilistic model),*** το πιο συνηθισμένο από το

οποίο είναι το ***Bayesian Model***

Βέβαια οι πιο προχωρημένες τεχνικές *indexing* προσπαθούν να ορίσουν τις *έννοιες* ***(concepts),*** οι οποίες χρησιμοποιούνται σε ένα έγγραφο, χρησιμοποιώντας στατιστικές μεθόδους, οι οποίες συσχετίζουν την εμφάνιση λέξης και έννοιας. Συχνά διατηρείται τόσο η συχνότητα εμφάνισης των λέξεων (φράσεων ή εννοιών) όσο και η τοποθεσία τους. Μερικές μηχανές αναζήτησης αξιώνουν τη χρήση επεξεργασίας της φυσικής γλώσσας. Η συχνή συνύπαρξη όρων σε ένα εύρος εγγράφων χρησιμοποιείται επίσης για την αναγνώριση φράσεων και εννοιών.

Οι παραπάνω τεχνικές φυσικά εστιάζονται και εφαρμόζονται σε συγκεκριμένα τμήματα των σελίδων, όπως στους τίτλους τους, στους πρώτους διακόσιους χαρακτήρες, στις πρώτες είκοσι γραμμές του κειμένου, στις εικόνες, στα metatags, κ.τ.λ.

Από τον πίνακα του Παραρτήματος Ε που αναφέρεται στο *Indexing* παρατηρούμε αρχικά ότι τα περισσότερα εργαλεία αναζήτησης ευρετηριάζουν ***ολόκληρη τη σελίδα*** και όχι ένα μέρος της. Επίσης θα πρέπει το ***πρότυπο του προγράμματος ευρετηρίασης*** να ευρετηριάζει τις παρακάτω πηγές: Web pages, Usenet, ειδήσεις, e-mail διευθύνσεις, ταχυδρομικοί κώδικες, αριθμοί τηλεφώνων, Yellow Pages. Ακόμη ένα μεγάλο μέρος των εργαλείων αναζήτησης (έντεκα από τα είκοσι έξι) υποστηρίζει την ***Αναζήτηση Φράσεων (Phrase Searching)*** είτε τοποθετώντας τις φράσεις μέσα σε διπλά εισαγωγικά, είτε κάνοντας click στην επιλογή που διαθέτει το εργαλείο αναζήτησης. Άρα το ***πρότυπο του προγράμματος ευρετηρίασης*** πρέπει να υποστηρίζει την αναζήτηση φράσεων. Η αναζήτηση Μικρών και Κεφαλαίων γραμμάτων δεν υποστηρίζεται από το μεγαλύτερο σύνολο των εργαλείων.

Μια άλλη σκοπιά με την οποία μπορούμε να ερευνήσουμε το indexing των εργαλείων αναζήτησης και να προσπαθήσουμε να αναπτύξουμε το πρότυπο είναι οι διάφοροι τρόποι indexing που εφαρμόζονται από τα παραπάνω εργαλεία. Παρακάτω υπάρχουν οι διάφοροι τρόποι του indexing, οι οποίοι έχουν αναπτυχθεί αναλυτικά στην ενότητα 4.7. της παρούσας εργασίας:

- *Πλήρες Κείμενο (Full Body Text)*
- *Stop Words*
- *Meta Description*
- *Meta Keywords*
- *ALT Text*

- *Σχόλια*
- *Stemming*

Από τον παρακάτω Πίνακα 8.4. και από τους πίνακες των Παραρτημάτων Ε και ΣΤ που αναφέρονται στο Indexing, συμπεραίνουμε ότι όλα τα εργαλεία αναζήτησης υποστηρίζουν την *ευρετηρίαση πλήρους κειμένου* ***(Full Body Text)***. Η δεύτερη σημαντική παρατήρηση είναι ότι ο μεγαλύτερος αριθμός των εργαλείων αναζήτησης υποστηρίζει τα *meta description* (23 από τα 26 εργαλεία), τα *meta keywords* (22 από τα 26 εργαλεία). Αντιθέτως τα σχόλια και το stemming δεν ευρετηριάζονται από το σύνολο των εργαλείων αναζήτησης.

Με βάση όλα τα παραπάνω μπορούμε να συμπεράνουμε ότι το ***πρότυπο του προγράμματος ευρετηρίασης*** θα αποτελείται καταρχήν από ένα αυτόματα ευρετήριο καθώς τα περισσότερα εργαλεία αναζήτησης το έχουν υιοθετήσει καθώς επίσης και λόγω της πάρα πολύ μεγάλης αύξησης του www μόνο αυτό το είδος του ευρετηρίου μπορεί να ανταποκριθεί καλύτερα σε κάθε ερώτηση, που μπορεί να τεθεί από τον χρήστη.

Indexing	**Ναι**	**Όχι**	**Παρατηρήσεις**
Πλήρες Κείμενο (Full Body Text)	Όλες	-	Μερικές stop words μπορεί να μην ευρετηριάζονται
Stop Words	AltaVista, Lycos, Google, Excite	Καμιά	
Meta Description	Όλες	Google, Lycos, Yahoo!	
Meta Keywords	Όλες	Excite, Google, Lycos, Yahoo!	
ALT Text	AltaVista, Lycos	Excite, Google	
Σχόλια	Καμιά	Άλλες	
Stemming			

Πίνακας 8.4.: Η σύγκριση των εργαλείων αναζήτησης με βάση το Indexing (Bikkannavar, 1999)

Το αυτόματο ευρετήριο θα είναι σταθμικό ευρετήριο, το οποίο και θα υποστηρίζει

την *ευρετηρίαση πλήρους κειμένου* καθώς και την *αναζήτηση-ευρετηρίαση του ευρύτερου νοήματος* ***(concept-based indexing or searching)***, η οποία δεν αναζητά μόνο τους όρους που έχει εισαγάγει ο χρήστης αλλά, προχωρώντας ένα βήμα παραπέρα, προσπαθεί να προσδιορίσει το ευρύτερο νόημα των όρων αυτών, προκειμένου να ανακτήσουν όσο το δυνατόν πιο συναφείς πληροφορίες.

Τέλος συγκρίνοντας το μέγεθος των κειμένων κάθε σελίδας που ευρετηριάζουν τα σημαντικότερα εργαλεία αναζήτησης (η AltaVista ευρετηριάζει τα 100Kb κειμένου κάθε σελίδας, η Infoseek τους πρώτους 200 χαρακτήρες, η Lycos τις 100 σημαντικότερες λέξεις και τις 20 πρώτες γραμμές κειμένου και η Excite τους 395 πρώτους χαρακτήρες και τους 70 χαρακτήρες τίτλου), το ***πρότυπο του προγράμματος ευρετηρίασης*** θα πρέπει να ευρετηριάζει τουλάχιστον τους 200 πρώτους χαρακτήρες του κειμένου της κάθε σελίδας. Επίσης θα πρέπει να ευρετηριάζει πρώτα από όλα τους τίτλους και τις κεφαλίδες της κάθε σελίδας και να μην λαμβάνει καθόλου υπόψη τα metatags, αφού τις περισσότερες φορές ο σκοπός τους είναι η αύξηση της σχετικότητας και κατά συνέπεια της σειράς ταξινόμησης των σελίδων. Βέβαια το ***πρότυπο του προγράμματος ευρετηρίασης*** πρέπει να υποστηρίζει την *ευρετηρίαση πλήρους κειμένου* ***(Full Body Text)***, τα *meta description* και τα *meta keywords*.

8.1.5. Το Πρότυπο της Μηχανής Ανάκτησης- Μηχανής αναζήτησης- Το Ειδικό Πρόγραμμα- Τρόπος Ταξινόμησης (Retrieval Engine- Ranking)

Όπως έχει αναφερθεί και παραπάνω η *μηχανή ανάκτησης, μηχανή αναζήτησης, το ειδικό πρόγραμμα* ***(retrieval engine)*** είναι υπεύθυνη για την ευρετηρίαση της βάσης δεδομένων της εκάστοτε μηχανής και την εμφάνιση των αποτελεσμάτων *ταξινομημένα* ***(ranking)*** στο χρήστη. Επίσης οι αναλυτικές πληροφορίες σχετικά με τους αλγορίθμους ταξινόμησης που χρησιμοποιούνται από τις κυριότερες μηχανές αναζήτησης δεν είναι δημόσια διαθέσιμες, ωστόσο φαίνεται ότι οι περισσότεροι χρησιμοποιούν:

- Την τεχνική *απόδοσης βάρους σε κάθε όρο* ***(Indexing Term Weighting)*** ή παραλλαγές αυτής και
- Το ***Vector Space Model***

Οι τεχνικές αυτές έχουν αναπτυχθεί αναλυτικά στην ενότητα 3.3.3.4 του κεφαλαίου 3 της παρούσας εργασίας. Στην ίδια ενότητα περιγράφονται και οι πιο

συνηθισμένοι παράγοντες που χρησιμοποιούν τα εργαλεία αναζήτησης για να πραγματοποιήσουν την ανάκτηση και ταξινόμηση των σχετικών με την ερώτηση του χρήστη εγγράφων από τις βάσεις δεδομένων τους.

Γενικά τα περισσότερα εργαλεία αναζήτησης κάνουν χρήση της θέσης και της συχνότητας εμφάνισης των λέξεων-κλειδιών σε μια ιστοσελίδα με στόχο την ταξινόμησή της. Ο ακριβής μηχανισμός λειτουργεί λίγο διαφορετικά για κάθε εργαλείο αναζήτησης.

Σε αντίθεση της θέσης και της συχνότητας, μερικές μηχανές αναζήτησης δίνουν μεγαλύτερη σημασία σε άλλους παράγοντες. Αυτοί οι παράγοντες βοηθούν λίγο και δεν εγγυώνται την καλύτερη δυνατή ταξινόμηση. Μερικοί σημαντικοί παράγοντες φαίνονται στον Πίνακα 8.5 και στους πίνακες των Παραρτημάτων Ε και ΣΤ που αναφέρονται στο Ranking. Στον πίνακα 8.5. αυτό ένα σημαντικό χαρακτηριστικό είναι το *πλήθος των συνδέσεων* ***(Link Popularity)*****,** το οποίο μπορεί να καθορίσει τη δημοτικότητα μιας σελίδας αναλύοντας πόσες συνδέσεις (αναφορές) καταλήγουν σε αυτή από άλλες σελίδες. Μερικές μηχανές ταξινομούν υψηλότερα σελίδες με πολλές συνδέσεις, ή συνδέσεις από σημαντικές ιστοσελίδες.

Ranking	**Ναι**	**Όχι**	**Παρατηρήσεις**
Meta Tags βοηθούν στην Ταξινόμηση	Καμιά	AltaVista, Lycos, Google, Excite	
Reviewed Status βοηθά στην Ταξινόμηση	Καμιά	AltaVista, Lycos, Google, Excite	
Το πλήθος συνδέσεων βοηθά στην Ταξινόμηση	AltaVista, Google, Excite	Lycos	Πολύ σημαντικό για το Google
Direct Hit βοηθά στην Ταξινόμηση	Καμιά	Όλες	

Πίνακας 8.5.: Σύγκριση των εργαλείων αναζήτησης με βάση τα χαρακτηριστικά ταξινόμησης (Bikkannavar, 1999)

Το ***πρότυπο της μηχανής ανάκτησης- τρόπου ταξινόμησης*** θα πρέπει να λαμβάνει υπόψη τις τεχνικές της ενότητας 3.3.3.4 του κεφαλαίου 3 και πιο συγκεκριμένα:

- ✓ Την *τοποθεσία* ***(location)*** και τη *συχνότητα* ***(frequency)*** των όρων της αναζήτησης.

- ✓ Οι μηχανές αναζήτησης ελέγχουν επίσης αν οι λέξεις-κλειδιά της αναζήτησης εμφανίζονται κοντά στην κορυφή μίας ιστοσελίδας, όπως για παράδειγμα στην επικεφαλίδα ή στις πρώτες παραγράφους του κειμένου.
- ✓ Η συχνότητα είναι ο άλλος σημαντικός παράγοντας σχετικά με τον τρόπο που οι μηχανές αναζήτησης καθορίζουν τη σχετικότητα.
- ✓ Η τεχνική *τοποθεσίας/ συχνότητας* ***(location/frequency method)***.
- ✓ Ο αριθμός των όρων σε μια ιστοσελίδα που ταιριάζουν με την ερώτηση του χρήστη.
- ✓ Το πόσο σπάνιος είναι ο όρος μιας ερώτησης.
- ✓ Η *εγγύτητα* ***(proximity)*** των όρων.
- ✓ Η ημερομηνία των εγγράφων- τα πιο πρόσφατα αρχεία ταξινομούνται υψηλότερα από τα παλαιότερα αρχεία.
- ✓ Κάποιες μηχανές αναζήτησης καταχωρούν σε ευρετήριο πιο πολλές ιστοσελίδες από άλλες.
- ✓ Ακόμα, μερικές μηχανές αναζήτησης αυξάνουν τη *βαθμολογία* ***(score)*** σχετικότητας των σελίδων, στις οποίες καταλήγουν πολλά *links*, με τη λογική ότι αυτές είναι δημοφιλείς σελίδες και άρα ο κόσμος θα θέλει να τις ανακτήσει.
- ✓ Μερικές μηχανές αναζήτησης μπορεί ακόμα, να αποκλείσουν κάποιες σελίδες από το ευρετήριο, αν εντοπίσουν '***spamming'***.
- ✓ Μερικές μηχανές αναζήτησης που υποστηρίζουν την περιγραφή των meta και των λέξεων-κλειδιών θα ταξινομήσουν καλύτερα μια σελίδα εάν οι αναζητήσιμοι όροι εμφανίζονται σε αυτές τις περιοχές. Όλες οι μηχανές αναζήτησης που υποστηρίζουν αυτά τα *tags* δεν ταξινομούν υψηλότερα τις σελίδες.

Στην ενότητα αυτή έγινε μια προσπάθεια να αναπτυχθεί το πρότυπο- μοντέλο μιας ενιαίας αναζήτησης με βάση τα κύρια μέρη από τα οποία αποτελείται ένα εργαλείο αναζήτησης. Είναι κατανοητό ότι λόγω του πλήθους των εργαλείων αναζήτησης, αλλά και του συγκεκριμένου έργου που επιτελεί κάθε ένα από αυτά, υπάρχουν πάρα πολλά χαρακτηριστικά που είναι διαφορετικά σε κάθε εργαλείο. Οπότε δεν είναι δυνατό να καλυφθεί η σύγκρισή τους και ταυτόχρονα να δημιουργηθεί και το πρότυπο της ενιαίας αναζήτησης.

ΚΕΦΑΛΑΙΟ 9
ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη σημερινή εποχή, της Παγκοσμιοποίησης, των Επικοινωνιών και της ραγδαίας εξέλιξης της Τεχνολογίας της Πληροφορίας (IT), καθημερινά γινόμαστε αποδέκτες ενός τεράστιου όγκου πληροφορίας, Η εξάπλωση του Διαδικτύου σε παγκόσμια κλίμακα, έφερε επανάσταση στις δυνατότητες εντοπισμού και πρόσβασης της πληροφορίας. Ωστόσο, αν και ο όγκος της πληροφορίας που υπάρχει στο διαδίκτυο είναι ασύλληπτος, δεν υπάρχει η κατάλληλη οργάνωση της, ούτε εγγυάται κανείς για την ποιότητα της και την αξιοπιστία της πηγής της. Αυτό έχει ως αποτέλεσμα την ύπαρξη τεράστιων δυσκολιών, προκειμένου να μπορέσει κάποιος να εντοπίσει και να προσπελάσει την πληροφορία που επιθυμεί.

Η ανάγκη ενός εργαλείου, το οποίο να βοηθάει το χρήστη του διαδικτύου στην αναζήτηση πληροφοριών, οδήγησε στη δημιουργία και την ανάπτυξη των μηχανών αναζήτησης, των θεματικών ευρετηρίων, των «πυλών» και των μεταμηχανών αναζήτησης.

Αν και το κάθε εργαλείο έχει τα πλεονεκτήματα του, σε γενικές γραμμές υστερούν, διότι είναι:

- ***Ελλιπή*** - για παράδειγμα, δεν καλύπτουν όλη τη διαθέσιμη πληροφορία
- ***Ανακριβή*** - Περιέχουν πληροφορία η οποία δεν είναι ενημερωμένη, είναι λανθασμένη, ή δεν έχει κατηγοριοποιηθεί σωστά και επιστρέφουν στους χρήστες πληροφορία μη σχετική και χαμηλής ποιότητας.

Είναι προφανές ότι, στις περισσότερες περιπτώσεις, περισσότερη προσπάθεια

δαπανάται στη συλλογή νέων εγγράφων, από ότι στην επαλήθευση των υπαρχόντων, εξαιτίας της συχνής ύπαρξης, «νεκρών», ανενημέρωτων links στα αποτελέσματα.

Τα χειρωνακτικά διατηρημένα θεματικά ευρετήρια, αν και είναι εύκολα στη χρήση και αρκετά ακριβή, καλύπτουν μόνο ένα μικρό μέρος της διαθέσιμης πληροφορίας. Εξάλλου, η έλλειψη αυτοματισμού τα κάνει ιδιαίτερα ανενημέρωτα, ενώ η ανθρώπινη παρέμβαση μπορεί να δημιουργήσει προκατειλημμένα και περιορισμένα αποτελέσματα.

Οι αυτοματοποιημένες μηχανές αναζήτησης, ενώ είναι τα πιο περιεκτικά εργαλεία όσον αφορά την κάλυψη του δικτύου, είναι ιδιαίτερα επιρρεπείς στην ανακρίβεια. Η αυτόματη ανάλυση και η κατηγοριοποίηση είναι μία πολύπλοκη διαδικασία, και μολονότι οι προσπάθειες που γίνονται για την ομαδοποίηση των σχετιζόμενων εγγράφων βάσει έννοιας είναι ελπιδοφόρες, υπάρχει ακόμα μία τάση επιστροφής πολύ περισσότερων ανακριβών αποτελεσμάτων από ότι στην περίπτωση των χειρωνακτικά διατηρημένων, κατηγοριοποιημένων πόρων.

Οι μετά-μηχανές αναζήτησης είναι ένα καλό σημείο εκκίνησης για τους αρχάριους, αλλά δεν επιτρέπει την πρόσβαση στις επιλογές αναζήτησης (Boolean, κ.λ.π.) της κάθε μηχανής αναζήτησης χωριστά.

Η ιδέα των subject – specific gateways είναι υποσχόμενη, αλλά μόνο αν υλοποιηθεί παγκόσμια μία δεδομένη μέθοδος, ούτως ώστε να μπορούν όλες να συνδεθούν ενιαία. Εάν δεν έχουν κάποιου είδους κεντρικό πόρο για να τους συνδέει, θα είναι τόσο δύσκολο να εντοπιστούν όσο και η κάθε μία πληροφορία, που αποθηκεύουν.

Υπάρχουν γενικά διαφορές στον τρόπο λειτουργίας των διαφόρων μηχανών αναζήτησης. Η κάθε μία έχει τα δικά της χαρακτηριστικά. Τα βασικά σημεία διαφοροποίησης των διαφόρων μηχανών αναζήτησης είναι:

- ***Μέγεθος βάσης δεδομένων***
- ***Τρόπος δημιουργίας του ευρετηρίου***
- ***Ενημέρωση τον ευρετηρίου***
- ***Κατάταξη των αποτελεσμάτων***
- ***Τεχνικές αναζήτησης***

Συνεπώς, μία μηχανή αναζήτησης φαίνεται να μην είναι αρκετή, προκειμένου να προσφέρει αφενός τον όγκο και αφετέρου την ποιότητα της πληροφορίας, που απαιτεί κάποιος χρήστης.

Εξάλλου, υπάρχει το λεγόμενο «αόρατο» Δίκτυο, το οποίο περιέχει τεράστιο όγκο πληροφορίας, η οποία όμως δεν καταχωρείται στο ευρετήριο καμιάς μηχανής αναζήτησης και συνεπώς δεν είναι διαθέσιμη. Επίσης, υπάρχει ακόμα πρόβλημα στην ανίχνευση πληροφοριών, οι οποίες δεν είναι σε HTML μορφή (π.χ. PDF αρχεία, οπτικοακουστικό υλικό, κ.λ.π.).

Γενικότερα, αν και υπάρχει πλούσια βιβλιογραφία σχετικά με το θέμα της αξιολόγησης των μηχανών αναζήτησης, παρουσιάζονται έντονες δυσκολίες κατά την πραγματοποίηση μίας τέτοιας αξιολόγησης, ενώ λίγες είναι οι έρευνες που λαμβάνουν υπόψη τον ίδιο τον χρήστη και τον τρόπο που αλληλεπιδρά με τη μηχανή αναζήτησης.

Οι περισσότερες μελέτες αξιολόγησης των εργαλείων αναζήτησης γίνεται με βάση τα παρακάτω κριτήρια:

- ***Σχεδίαση- περιβάλλον αρχικής σελίδας***
- ***Ταχύτητα εμφάνισης αρχικής σελίδας***
- ***Ταχύτητα εμφάνισης αποτελεσμάτων***
- ***Επάρκεια πληροφόρησης που συνοδεύει τα αποτελέσματα- μορφή τους***
- ***Ποσότητα των αποτελεσμάτων***
- ***Ακρίβεια των αποτελεσμάτων***
- ***Ευελιξία δυνατότητας***
- ***Ενημερότητα των αποτελεσμάτων***

Τα παραπάνω κριτήρια είναι περισσότερο ορατά και κατανοητά στον χρήστη. Βέβαια είναι και αυτά που τον ενδιαφέρουν περισσότερο, όπως η ακρίβεια των αποτελεσμάτων. Η παρούσα εργασία όμως δεν είχε στόχο να μελετήσει επιφανειακά όλα τα παραπάνω εργαλεία αναζήτησης και να τις συγκρίνει με γνώμονα τα βασικά τους χαρακτηριστικά που είναι προφανή στους περισσότερους χρήστες κάθε εργαλείου αναζήτησης.

Στη συγκεκριμένη εργασία έγινε μια προσπάθεια να εξεταστούν τα τεχνικά χαρακτηριστικά κάθε εργαλείου και ο τρόπος λειτουργίας του. Τι κρύβεται πίσω από μια προσεγμένη σχεδιαστικά τις περισσότερες φορές ***homepage*** ενός εργαλείου αναζήτησης. Έτσι για κάθε εργαλείου εξετάστηκε ο *τρόπος ευρετηρίασης* ***(indexing)*** των πληροφοριών, οι *τεχνικές ανάκτησης* ***(Retrieval Engine)*** των πληροφοριών που εφαρμόζουν καθώς και η *εμφάνιση* ***(Display Features)*** σε συνάρτηση με τη *διάταξη* ***(Ranking)*** των αποτελεσμάτων που προσφέρει το κάθε εργαλείο. Επίσης μελετήθηκε και η *συλλογή των πληροφοριών* ***(crawling)*** από το διαδίκτυο. Όλα τα παραπάνω

χαρακτηριστικά αποτελούν την καρδιά του εργαλείου αναζήτησης.

Από τα παραπάνω γίνεται αντιληπτό ότι και το πρότυπο της ενιαίας αναζήτησης θα βασίζεται σε αυτά τα τμήματα του εργαλείου αναζήτησης. Αναλυτικότερα αναπτύχθηκε το πρότυπο της γραφικής διεπαφής ***(the graphical HTML (Hyper-text Markup Language) interface)*** ή αλλιώς του ***User Interface***. Έπειτα αναπτύχθηκε το πρότυπο του ειδικού λογισμικού ***(robot, spider, crawler κ.λ.π.)*** δηλαδή το ***crawling***. Περαιτέρω, η *βάση δεδομένων ή αλλιώς ο κατάλογος* ***(database of information)*** για το σύνολο των εργαλείων αναζήτησης. Στη συνέχεια αναφέρθηκε το πρότυπο του *προγράμματος ευρετηρίασης και το ευρετήριο* ***(the indexing program and the index)***. Τέλος αναπτύχθηκε η *μηχανή ανάκτησης, μηχανή αναζήτησης, το ειδικό πρόγραμμα* ***(retrieval engine)*** που είναι υπεύθυνη για την ευρετηρίαση της βάσης δεδομένων της εκάστοτε μηχανής και την εμφάνιση των αποτελεσμάτων ταξινομημένα στο χρήστη. Αυτές είναι δύο διαδικασίες που δεν είναι ευδιάκριτες στον χρήστη και δεν πρέπει να είναι. Έτσι μελετήσαμε και το πρότυπο της διάταξης και ταξινόμησης των σελίδων αποτελεσμάτων ***(Ranking)***.

Προσπαθήσαμε να εντοπίσουμε τις περισσότερες ομοιότητες μεταξύ αυτών των τεχνικών τμημάτων του κάθε εργαλείου αναζήτησης με σκοπό τη δημιουργία ενός τέτοιου εργαλείου που θα περιείχε τα περισσότερα πλεονεκτήματα από κάθε εργαλείο αναζήτησης ξεχωριστά. Αυτή η διαδικασία δεν είναι καθόλου εύκολη γιατί η κάθε μηχανή αναζήτησης έχει τη δική της φιλοσοφία, έχει τους δικούς της αλγορίθμους, εφαρμόζει τις δικές της τεχνικές και κανόνες. Εκτός από τα καθαρά παραπάνω μέρη, επίσης πολύ σημαντικό είναι ο σκοπός του κάθε εργαλείου αναζήτησης και το είδος των πληροφοριών που συλλέγει και εμφανίζει στους χρήστες του. Για παράδειγμα τα θεματικά ευρετήρια πληροφορούν τους χρήστες καλύτερα για κάποιες κατηγορίες θεμάτων από ότι οι μηχανές γενικής αναζήτησης. Άρα και η κατασκευή τους θα διαφέρει.

Έχοντας υπόψη όλα τα προηγούμενα, θα μπορούσαμε να διαπιστώσουμε ότι οι εξελίξεις στα εργαλεία αναζήτησης πληροφοριών στον παγκόσμιο ιστό αναμένονται ραγδαίες. Παρά το γεγονός ότι οι μηχανισμοί που χρησιμοποιούν οι μηχανές και τα άλλα εργαλεία αναζήτησης γίνονται ολοένα πιο αποτελεσματικοί, ο ρυθμός αύξησης των σελίδων που δημοσιεύονται στο *internet* είναι τέτοιος, ώστε δεν αφήνει περιθώρια για μια συνολική κάλυψη μεγαλύτερη του 40-50%, στην καλύτερη των περιπτώσεων. Εκτός αυτού, ο μέσος χρήστης, όντας αντιμέτωπος με τις εκατοντάδες των αποτελεσμάτων που θα έχει ανακτήσει και βελτιώσει, δύσκολα θα κατορθώσει να

διερευνήσει περισσότερες από 50-60 από αυτές τις σελίδες, αφού οι παραδοσιακοί μηχανισμοί (crawlers, spiders) που χρησιμοποιούν σήμερα τα περισσότερα εργαλεία αναζήτησης δε διαθέτουν την «κρίση» που απαιτείται ώστε να επιλέξουν το πιο κατάλληλο και συναφές αποτέλεσμα σύμφωνα με την αναζήτηση.

Οι μηχανές αναζήτησης και οι θεματικοί κατάλογοι που διαθέτουν ανθρώπινο εξειδικευμένο προσωπικό, το οποίο αναλαμβάνει να ευρετηριάσει τα περιεχόμενα των πληροφοριακών πηγών και να κατατάξει τη συσσωρευμένη γνώση στις κατάλληλες κατηγορίες, είναι τα εργαλεία εκείνα που επιτυγχάνουν τα καλύτερα αποτελέσματα. Έτσι, οι σημερινές τάσεις, δείχνοντας να έχουν καταλάβει τη σημασία του ανθρώπινου παράγοντα, αναπτύσσουν τεχνολογίες που ουσιαστικά μιμούνται την ανθρώπινη συμπεριφορά. Τα πρώτα δείγματα τέτοιων εργαλείων (Google, Direct Hit) δείχνουν το μέλλον στις μηχανές αναζήτησης. Οι μηχανισμοί που διαθέτουν οι συγκεκριμένες μηχανές δεν κάνουν τίποτα άλλο παρά να ανακτούν τα αποτελέσματα που άλλοι χρήστες έχουν επιλέξει ως απάντηση στα ερωτήματα τους ή να ανακτούν αποτελέσματα που οδηγούν ή αναφέρονται ως ευρύτερα χρησιμοποιούμενες πληροφοριακές πηγές σε άλλα sites. Έτσι, βασιζόμενες ουσιαστικά στον ανθρώπινο παράγοντα και την ανθρώπινη κρίση, οι μηχανές αυτές πετυχαίνουν με πολύ μικρότερη κάλυψη να ανακτούν σε μεγάλο ποσοστό αποτελέσματα άμεσα σχετιζόμενα με το αρχικό ερώτημα και σαφώς λιγότερα στον αριθμό, άρα πιο εύκολα επεξεργάσιμα από το χρήστη.

Σε ό,τι αφορά τη μορφή που θα έχουν τα εργαλεία αναζήτησης πληροφοριών τα επόμενα χρόνια, οι χρήστες δείχνουν να έχουν κάνει τις επιλογές τους. Ο υψηλός βαθμός παραμετροποίησης και η αυτοματοποιημένη λειτουργία είναι τα βασικά χαρακτηριστικά των εργαλείων αυτών και αποτελούν το κύριο στοιχείο διαφοροποίησης σε σχέση με τις μηχανές αναζήτησης και τα θεματικά ευρετήρια. Τα εν λόγω προγράμματα δεν είναι τίποτα περισσότερο από μηχανισμοί που παραμετροποιούνται σε εξαντλητικό βαθμό από το χρήστη, προκειμένου η αναζήτηση τους να είναι όσο το δυνατόν πιο συγκεκριμένη και αποτελεσματική. Επιπλέον, εφόσον γίνει ο αρχικός προγραμματισμός τους, τα προγράμματα αυτά αναλαμβάνουν να αναζητούν, να ανακτούν και να παρουσιάζουν στο χρήστη τις πληροφορίες που έχει προδιαγράψει, λειτουργώντας αυτόνομα και χωρίς να απαιτούν από το χρήστη να τα ενεργοποιεί κάθε φορά. Όπως είναι φανερό, μηχανισμοί τέτοιου είδους προσδίδουν μοναδικά πλεονεκτήματα στους χρήστες που επιθυμούν να ενημερώνονται σε μόνιμη βάση για συγκεκριμένες εξελίξεις (τιμές μετοχών, παρακολούθηση ειδήσεων) που

σημειώνονται είτε καθημερινά είτε σε τακτά χρονικά διαστήματα, εξασφαλίζοντας παράλληλα σε αυτούς πολύτιμο χρόνο.

Η ραγδαία αύξηση των περιεχομένων του διαδικτύου, ο γιγαντισμός των ευρετηρίων και των θεματικών καταλόγων, ο μετασχηματισμός του διαδικτύου σε ένα εμπορικό μέσο με παγκόσμια απήχηση καθώς και η σταδιακή μείωση του διαθέσιμου χρόνου των επαγγελματιών για αναζήτηση πληροφοριών καθιστούν τη χρήση των «έξυπνων πρακτόρων» επιτακτική για εκείνους που θεωρούν ότι το διαδίκτυο μπορεί πραγματικά να τους προσφέρει το ανταγωνιστικό πλεονέκτημα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

I. Ξένη Βιβλιογραφία

AltaVista Software (2002), *AltaVista Enterprise Search Technical Overview*.

Angelaccio M., Buttarazzi B. (2002), *Local Searching the Internet*, IEEE Internet Computing, 1089-7801, (http://computer.org/internet/).

Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley Pyb Co, (http://www.dcc.ufmg.br/irbook/)

Bikkannavar N. (1999), *Search Engines – A Survey Report*.

Brin, S. & Page, L. (1998), *The anatomy of a large-scale hypertextual web search engine*, In Proceedings of WWW (pp. 107-117), Amsterdam, Elsevier.

Chen H. and Houston A. L., Sewell R. R., Schatz B. R. (1998), *Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques*, Journal of the American Society for Information Science, 49 (7): 582–603.

Chen H., Fan H., Chau M., and Zeng D. (2001), *MetaSpider: Meta-Searching and Categorization on the Web*, Journal of the American Society for Information Science and Technology.

Chignell, M., Gwizdka, J., Bodner, R. (1999), *Discriminating meta-search: a framework for evaluation*, Information Processing and Management, Elsevier Science, 35, p. 337-362.

Chu, H. & Rosenthal, M. (1996), *Search engines for the world wide web: A comparative study and evaluation methodology*. ASIS 1996 Annual Conference Proceedings, Baltimore, MD, (http://www.asis.org/annual-96/ElectronicProceedin g s/chu .html).

Cooke, A. (1999), *A Guide to Finding Quality Information on the Internet*, Library Association Publications Ltd.

Chakrabarti S., Martin van den Berg, Byron D. (1999), *Focused crawling: a new approach to topic-specific Web resource discovery*, Computer Networks, Elsevier Science, 31, pp 1623–1640.

Chau M., Chen H. (2003), *Comparison of Three Vertical Search Spiders*, IEEE Computer Society, 0018-9162.

Chau M., Zeng D., Chen H., Huang M., Hendriawan D. (2003), *Design and evaluation of a multi-agent collaborative Web mining system*, Decision Support Systems, Elsevier Science, 35, pp 167– 183.

Chen C. C., Chen C. M., Sun Y. (2001), *PVA: A Self-Adaptive Personal View Agent System*.

Cheung D, W., Kao B., Lee J. (1998), *Discovering user access patterns on the World Wide Web*, Knowledge Based Systems, 10, pp 463-470.

Christophi C., Dikaiakos M. (2003), *Automatic profile generation in eRACE*, University of Cyprus.

Dreilinger, D., & Howe, A. E. (1997), Experiences with selecting search engines using metasearch, ACM Transactions on Information Systems, 15(3), 195-222.

Filman R., Feniosky P.-M. (1998), *Seek, and YE Shall Find*, IEEE Internet Computing, (http://computer.org/internet/).

Filman R., Pant S. (1998), *Searching the Internet*, IEEE Internet Computing, 1089-7801, (http://computer.org/internet/).

Frakes, W.B, Baeza- Yates (1992), *Information Retrieval Data Structures and Algorithms*, Prentice Hall.

Fructl M., Kreuziger J., Beigl M. (1977), *Assistant for an Information Database*.

Gehmeyr A., Muller J., Schappert A., *Mobile Information Agents on the Web*.

Gillia M., Winker P. (2003), *A global optimization heuristic for estimating agent based models*, Computational Statistics & Data Analysis, Elsevier Science, 42, pp 299 – 312.

Glover E. J., Lawrence S., Birmingham W. P., Giles C. L. (1999), *Architecture of a Metasearch Engine that Supports User Information Needs*, ACM, CIMK, pp 210-216.

Godoy D. and Amandi A., *PersonalSearcher: An Intelligent Agent for Searching Web Pages*.

Gordon, M., Pathak, P. (1999), *Finding information on the world wide web: The retrieval effectiveness of search engines, Information Processing and Management*, Elsevier Science, 35, p.141-180

Hawkins J. (2001), *ixquick: Evaluating a New Metasearch Engine*.

Hearst M. A. (2000), *Next Generation Web Search: Setting Our Sites***,** University of California, Berkeley.

Henziger, M., *Web Information Retrieval- an Algorithmic Perspective*.

Henziger, M, Motwani R., Silverstein C. (2002), *Challenges in Web Search Engines*.

Hernández J. C., Sierra J. M., Ribagorda A., Ramos B. (2001), *Search Engines as a Security Threat*, IEEE 0018-9162, University Madrid

Hock R. (2001), *The Extreme Searcher's Guide to Web Search Engines: A Handbook for the Serious Searcher*, CyberAge Books.

Hou, M. (1998), *Comparison for three internet search tools (Yahoo, Alta Vista, Lycos)*, Unpublished manuscript, Department of Mechanical and Industrial Engineering, University of Toronto.

Hsinchun C., Yi-Ming C., Ramsey M., Yang C. C. (1998), *An intelligent personal spider agent for dynamic Internet/Intranet searching*, Decision Support Systems, Elsevier Science, 23, pp 41–58.

Huang L., *A Survey On Web Information Retrieval Technologies*, Computer Science Department, State University of New York.

Jenkins, C., Jackson, M., Burden, P., Wallis, J. (2000), *Searching the World Wide Web: an Evaluation of Available Tools and Methodologies*, Elsevier Science.

Junghoo Cho, Hector Garcia-Molina, Lawrence Page, *Efficient Crawling Through URL Ordering*, Department of Computer Science, Stanford University.

Kobayashi M. and Takeda K. (2000), *Information Retrieval on the Web***,** ACM Computing Surveys, Vol. 32, No. 2.

Lai H., Yang T. - C. (2000), *A system architecture for intelligent browsing on the Web*, Decision Support Systems, Elsevier Science, 28, pp 219–239.

Lawrence S. and Giles C. (1997), *Context and page analysis for improved Web Search*, IEEE Internet Computing.

Lawrence, S., & Giles, C. L. (1998), *Inquirus, the NECI meta search engine*, Elsevier Science, pp. 95-105.

Lawrence, S., & Giles, C. L. (1998), *Searching the World Wide Web*, Science, 5360(280), 98-100.

Leighton, & Srivastava. (1997), *Precision among World Wide Web search services (search engines): AltaVista, Excite, HotBot, Infoseek, Lycos,* (http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm).

Li Y. (1998), *Toward a Qualitative Search Engine*, IEEE Internet Computing, 1089-7801, (http://computer.org/internet/)

Lieberman H., *Letizia: An Agent That Assists Web Browsing.*

Menczer F. (2003), *Complementing search engines with online web mining agents*, Decision Support Systems, Elsevier Science, 35, pp 195– 212.

Moldovan D. I., Mihalcea R. (2000), *Using WordNet and Lexical Operators to Improve Internet Searches*, IEEE Internet Computing, 1089-7801, Southern Methodist University, (http://computer.org/internet/).

Moukas A. G. (1997), *Amalthaea: Information Filtering and Discovery Using A Multiagent Evolving System*, Massachusetts Institute of Technology.

Muller M. E., *An Intelligent Multi-Agent Architecture for Information Retrieval from the Internet*.

Pant G. and Menczer F., *MySpiders: Evolve your own intelligent Web crawlers*, Department of Management Sciences, University of Iowa.

Rus D. and Subramanian D. (1995), *Information Retrieval, Information Structure and Information Agents*.

Sandip D. AND Sandip S., Blackstock B. (2000), LawBot: *A Multiagent Assistant for Legal Research*, IEEE Internet Computing, 1089-7801, (http://computer.org/internet/)

Savoy, J., Picard, J. (2001), *Retrieval effectiveness on the web*, Information Processing and Management, Elsevier Science, 37, pp. 543-569

Seacord R. C., Hissam S. A., Wallnau K. C. (1998), *AGORA: A Search Engine for Software Components*, IEEE Internet Computing, 1089-7801, Carnegie Mellon University, (http://computer.org/internet/).

Shaw N. G., Mian A., Yadav S. B. (2002), *A comprehensive agent-based architecture for intelligent information retrieval in a distributed heterogeneous environment,* Decision Support Systems, Elsevier Science, 32, pp 401– 415.

Spink A, (2002), *A user-centered approach to evaluating human interaction with Web search engines: an exploratory study*, Information Processing and Management, Elsevier Science, 38, p, 401-426.

Srinivasan P., Menczer F., Pant G., *A General Evaluation Framework for Topical Crawlers*, University of Iowa

Tewari G., Youll J., Maes P. (2002), *Personalized location-based brokering using an agent-based intermediary architecture*, Decision Support Systems, Elsevier Science, 34, pp 127– 137.

Thomas B. (1998), *Rank and File*, SPIE — The International Society for Optical Engineering, IEEE Internet Computing, (http://computer.org/internet/).

Tillman H. N. (2000), *Evaluating Quality on the Net*, Babson College, Babson Park, Massachusetts, (www.hopetillman.com/findqual.html)

Vaughan, J. (1999), *Considerations in the choice of an Internet search tool*, MCB University Press, 17(1), pp. 89-106.

Yang Christopher C., Yen Jerome, Chen Hsinchun (2000), *Intelligent internet searching agent based on hybrid simulated annealing*, Decision Support Systems, 28, pp. 269–277

Yao J. T., Yao Y. Y., *Information Granulation for Web based Information Retrieval Support Systems*, Department of Computer Science, University Regina.

Zacharis Z. N., Panayiotopoulos T. (2001), *Web Search Using a Genetic Algorithm*, IEEE Internet Computing, 1089-7801, University of Piraeus, (http://computer.org/internet/).

Zantout H., Marir F. (1999), *Document management systems from current capabilities towards intelligent information retrieval: an overview*, International Journal of Information Management, Elsevier Science, 19, pp 471- 484.

Zhou X., Yates D. J., Chen G. (2001), *Using Visual Spatial Search Interface for WWW Applications*, Information Systems, Elsevier Science Ltd., Vol. 26, No. 2, pp. 61-74.

ΙΙ. Ελληνική Βιβλιογραφία

Κωνσταντινίδης, Σ. (2000), *Τεχνικές Αναζήτησης Επιχειρηματικών Πληροφοριών στο Διαδίκτυο*, Εκδόσεις Anubis.

Μακριδάκης, Γ. (2003), *Αξιολόγηση Μηχανών Αναζήτησης στο Διαδίκτυο και Ανάλυση Συμπεριφοράς των Χρηστών τους*, Πολυτεχνείο Κρήτης.

Φωστιέρη, Μ. (2001), *Ανάπτυξη Πολυκριτήριου Πληροφοριακού Συστήματος Πολλαπλών Πρακτόρων για την Αναζήτηση Πληροφοριών στο Διαδίκτυο*, Πολυτεχνείο Κρήτης.

III. Internet Sites

All-in-One: (http://www.albany.net/allinone)

AltaVista (www.altavista.com)

A Comparison of Seven Search Engines
(http://www.iwaynet.net/~lsci/Search/home.html)

A Standard for Robot Exclusion (http://www.robotstxt.org/wc/norobots.html)

Directories and Virtual Libraries
(http://www.webliminal.com/search/search-web04.html#Virtual Libraries: Directories with a Difference)

Dogpile (www.dogpile.com)

DirectHit: (http://www.directhit.com)

Excite (www.excite.com)

Glossary of Internet & Web Jargon
(http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Glossary.html)

Google (www.google.com)

GoTo: (http://www.goto.com)

Guide to Meta-Search Engines (http://www.indiana.edu/~librcsd/search/meta.html)

Highway61 : (http://www.highway61.com)

Hotbot (www.hotbot.com)

How Search Engines Rank Web Pages
(http://www.searchenginewatch.com/webmasters/article.php/2167961)

How Search Engines Work
(http://searchenginewatch.com/webmasters/article.php/2168031)

Infoseek (www.infoseek.com)

Inktomi (http://www.inktomi.com/)

Internet Tools for the Advanced Searcher (http://www.philb.com/adint.htm)

Ixquick (www.ixquick.com)

LookSmart: (http://www.looksmart.com)

Lycos (www.lycos.com)

Major Search Engines and Directories
(http://searchenginewatch.com/links/article.php/2156221)

Mamma (www.mamma.com)

Metasearch: (http://www.metasearch.com)

Meta-Search Engines
(http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html)

Metacrawler (www.metacrawler.com)

Metacrawlers – (http://searchenginewatch.com/links/Metacrawlers/)

Metacrawlers and Metasearch Engines
(http://searchenginewatch.com/links/article.php/2156241)

Multi-search Engines - a comparison (http://www.philb.com/msengine.htm)

NetSearch: (http://www.ais.net/netsearch/search_entry.html)

northernlight (http://www.northernlight.com/index.html)

northernlight (http://www.northernlight.com/library.html)

northernlight (http://www.northernlight.com/portal.html)

northernlight (http://www.northernlight.com/services.html)

northernlight (http://www.northernlight.com/technology.html)

One-Global: (http://one-global.com)

OneSeek: (http://www.oneseek.com)

ProFusion (www.profusion.com)

Recommended Subject Directories
(http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SubjDirectories.html)

Robots Exclusion (http://www.robotstxt.org/wc/exclusion.html)

Search.com (www.search.com)

Search Assistance Features (http://searchenginewatch.com/facts/article.php/2155971)

Search Engine Comparison: Features & Syntax
(http://www.iwaynet.net/~lsci/Search/tablevt.htm)

Search Engine Glossary (http://searchenginewatch.com/facts/article.php/2156001)

Search Engine Math (http://searchenginewatch.com/facts/article.php/2156021)

Search Engine Sizes http://searchenginewatch.com/reports/article.php/2156481

Search Engine Tutorial for Web Designers (http://northernwebs.com/set/index.html)

Searchengineguide http://searchengineguide.org/principa.htm

Search Features Chart (http://searchenginewatch.com/facts/article.php/2155981)

Searching and Researching on the Internet and the Web Glossary (http://www.webliminal.com/search/glossary.htm#directory)

Searching Beyond Text: Multimedia Search Tools (http://www.onlineinc.com/onlinemag/OL2000/net11.html)

Search Engines, Indexes Directories and Libraries (http://www.netstrider.com/search/)

SurfWax (www.surfwax.com/servlet/com.surfwax.FrontEnd.home

The Anatomy of a Large-Scale Hypertextual Web Search Engine (http://www-db.stanford.edu/~backrub/google.html#pr)

The BEST Search Engines (http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html)

The Web Robots FAQ... (http://www.robotstxt.org/wc/faq.html)

Virtual Yellow Pages: (http://www.vyp.com/yp/search.html)

Yahoo! (www.yahoo.com)

Web Searching Tips (http://searchenginewatch.com/facts/index.php)

webworkshop (http://webworkshop.net/pagerank.html#what_is_pagerank)

Worth A Look: Searching the Invisible Web (http://websearch.about.com/library/searchwiz/bl_invisibleweb_apra.htm?once=true&)

http://www.actionsearch.com/howtousesearch.htm

http://www.advmediaproductions.com/

http://www.ala.org/ala/lita/litaresources/toolkitforexpert/toolkitexpert.htm

http://www.ceenet.org/workshops/lectures2000/Miroslav_Milinovic/mmsi2000/tsld001.htm

http://www.cs.jhu.edu/~weiss/glossary.html

http://ecampus.bentley.edu/dept/li/INVISIBLE_WEB.HTM

http://ecampus.bentley.edu/dept/li/METASEARCH.HTM

http://ecampus.bentley.edu/dept/li/SEARCH_ENGINES.HTM

http://ecampus.bentley.edu/dept/li/SEARCH_TIPS.HTM

http://ecampus.bentley.edu/dept/li/SUBJECT_DIRECTORIES.HTM

http://www.evolt.org/article/The_Meta_Search_Engines/12/41694/

http://computer.howstuffworks.com/search-engine.htm

http://computer.howstuffworks.com/search-engine1.htm

http://computer.howstuffworks.com/search-engine2.htm

http://computer.howstuffworks.com/search-engine4.htm

http://ixquick.com/eng/metasearch_comparison.html

http://ixquick.com/eng/search_faq.html

http://ixquick.com/eng/sws_about.html

http://www.k-praxis.com/archives/cat_news_analysis.html

http://www.laisha.com/JanZine/webcrawler.html

http://www.lexisone.com/legalresearch/legalguide/internet_search_engines/metasearch_engines.htm

http://www.microsoft.com/mind/default.asp

http://www.netstrider.com/search/conclusions.html

http://www.netstrider.com/search/features.html

http://www.netstrider.com/search/ranking.html

http://www.pandia.com/metasearch/

http://www.positioning-search-engines.com/searchengines.htm

http://www.robotstxt.org/wc/exclusion-admin.html

http://www.robotstxt.org/wc/robots.html

http://www.searchengineguide.com

http://searchengineguide.org/yahoo.htm

http://searchenginewatch.com/

http://www.sc.edu/beaufort/library/pages/engines.shtml

http://www.submitcorner.com/Guide/SE/

http://www.tisl.ukans.edu/~sgauch/papers/webnet97

http://www.webmasterworld.com/glossary/

http://webreference.com/content/search/features.html

http://webreference.com/content/search/how.html

http://www2.widener.edu/Wolfgram-Memorial-Library/pyramid/metasearch.htm

http://www2.widener.edu/Wolfgram-Memorial-Library/pyramid/setbldes.htm

http://www2.widener.edu/Wolfgram-Memorial-Library/pyramid/setblsch.htm

http://www2.widener.edu/Wolfgram-Memorial-Library/pyramid/subdrtbl.htm

ΓΛΩΣΣΑΡΙ

Όρος	*Ορισμός*
Πράκτορας (Agent)	Μια οντότητα λογισμικού που τοποθετείται μέσα και ένα μέρος ενός περιβάλλοντος που αισθάνεται αυτό το περιβάλλον και ενεργεί σε αυτό, στο χρόνο, όπως προγραμματίζεται. Οι πράκτορες έχουν αυτονομία - η δυνατότητα να ενεργήσουν χωρίς ανθρώπινη επίδραση, και η συνεργασία με άλλους πράκτορες ή άλλες οντότητες λογισμικού. Ένας πράκτορας μπορεί να είναι στατικός ή κινητός. Μπορεί να είναι ικανός να ενεργήσει κατάλληλα ή να εκτελεί έναν συγκεκριμένο σκοπό. Οι πράκτορες μπορούν να εκτελέσουν μια ή περισσότερες λειτουργίες όπως συνεργάζονται, διασυνδέουν, συλλέγουν τις πληροφορίες, περιμένουν για γεγονότα, κ.τ.λ.. Όταν ένας πράκτορας μπορεί να μάθει από την εμπειρία του αναφέρεται συχνά ως έξυπνος πράκτορας. Οι κινητοί πράκτορες παρουσιάζουν ειδικό ενδιαφέρον για P2P. Ένας κινητός πράκτορας είναι ο πράκτορας που κινείται, ή μπορεί να κινηθεί, μέσω ενός δικτύου προκειμένου να εκτελέσει ένα δεδομένο στόχος εξ ονόματος ενός χρήστη. Παράδειγμα: ένας πράκτορας που εκτελεί μια διανεμημένη αναζήτηση, ανά παραμέτρους αναζήτησης που παρέχονται από το χρήστη. Οι όροι "ευφυής πράκτορας" και "BOT" χρησιμοποιούνται συχνά ταυτόσημα

Όρος	*Ορισμός*
	με "πράκτορα". **Εναλλακτικός ορισμός:** Μια ευφυής διαδικασία λογισμικού που μπορεί να διαμορφωθεί για να αποκριθεί αυτόματα όταν η περίσταση το απαιτήσει.
Boolean:	Οι τελεστές Boolean είναι οι βοηθητικοί όροι (AND, NOT, OR) που χρησιμοποιούνται σε ένα ερώτημα προκειμένου να περιορίσουν τον όγκο των αποτελεσμάτων και να εξειδικεύσουν τη ζητούμενη πληροφορία.
Browsing:	Ο όρος browsing στην πιο κοινή σημασία του χρησιμοποιείται για να περιγράψει την περιήγηση στο web, τον παγκόσμιο ιστό. Όταν μια αναζήτηση γίνεται μέσω κάποιου θεματικού καταλόγου, ο όρος χρησιμοποιείται για να περιγράψει τη θεματική αναζήτηση στα περιεχόμενα του καταλόγου αυτού.
Browser:	Είναι το πρόγραμμα που χρησιμοποιούμε προκειμένου να αποκτήσουμε πρόσβαση και να περιηγηθούμε στο Internet. Ο πιο γνωστός Internet Browser είναι ο ***Microsoft Explorer***.
Classification:	Είναι η ταξινόμηση, η διαδικασία κατά την οποία αποφασίζεται, είτε από άνθρωπο είτε από μηχανή (λογισμικό), η θεματική κατηγορία στην οποία ανήκει ένα έγγραφο.
Cluster:	Η ομαδοποίηση σχετιζόμενων εγγράφων.
Context-based searching ή Concept-based searching:	Η αναζήτηση όρων με Βάση τα συμφραζόμενα.
Crawler:	Το ειδικό λογισμικό που αναλαμβάνει να συγκεντρώνει και να ανανεώνει τις καταχωρίσεις σε μια βάση δεδομένων σαρώνοντας τις web pages και τα περιεχόμενά τους (συνδεόμενες σελίδες και έγγραφα, συνδεόμενα sites, κ.λ.π.)
Database:	Ως βάση δεδομένων θα μπορούσαμε να χαρακτηρίσουμε τη συγκέντρωση και την αποθήκευση πληροφοριών οργανωμένων με βάση κάποιο *κλειδί **(index key)***.
Directory Search:	Η αναζήτηση με τη χρήση των καταλόγων ή θεματικών καταλόγων, όπως

Όρος	*Ορισμός*
	ευρύτερα είναι γνωστοί, αποτελεί την εναλλακτική λύση έναντι των μηχανών αναζήτησης. Αυτοί αποτελούν καταλόγους οργανωμένους ιεραρχικά σε κατηγορίες και υποκατηγορίες από το γενικότερο όρο/ θέμα προς το ειδικότερο.
False Drops:	Τα αποτελέσματα της αναζήτησης που δεν έχουν σχέση με το ζητούμενο θέμα.
Full-text Indexing:	Η ευρετηρίαση του συνόλου των περιεχομένων μιας σελίδας. Οι μηχανές αναζήτησης που υποστηρίζουν την ευρετηρίαση αυτού του τύπου συγκεντρώνουν μεγαλύτερη βάση δεδομένων και διαθέτουν ευρετήριο.
Hierarchical:	Η ιεραρχική δομή ενός ή περισσότερων θεμάτων από το γενικότερο προς το ειδικότερο
Hits:	Τα αποτελέσματα, έγγραφα ή αναφορές σε άλλες web pages ή σε άλλα έγγραφα, τα οποία επιστρέφει η μηχανή αναζήτησης ως αποτέλεσμα του ερωτήματος που τέθηκε.
Hypertext Link:	Η υπογραμμισμένη ή εντονότερη λέξη ή εικόνα που παρουσιάζεται σε μια web page, προκειμένου να υποδηλώσει ότι η ενεργοποίηση της παραπέμπει σε κάποια συνδεόμενη σελίδα ή σε έγγραφο με σχετική πληροφορία. Τα links, όπως έχουν καθιερωθεί να ονομάζονται, αποτελούν τον κατεξοχήν τρόπο μετακίνησης από σελίδα σε σελίδα ή από site σε site στο web.
Indexing (Ευρετηρίαση):	Η διαδικασία κατά την οποία ένα έγγραφο μετατρέπεται σε μορφή υπό την οποία θα είναι εύκολα αναζητήσιμο και ανακτήσιμο.
Information Agent:	Το ειδικό λογισμικό που αναλαμβάνει να «φιλτράρει» τις πληροφορίες εκ μέρους του χρήστη, προκειμένου να μειώσει τον όγκο τους.
Information Filtering:	Πρόκειται για το «φιλτράρισμα» των πληροφοριών, τη διαδικασία κατά την οποία ένα σύνολο πληροφοριών υποβάλλεται σε μια περαιτέρω «βελτίωση», προκειμένου να περιέχει μόνο τις πληροφορίες εκείνες που πραγματικά έχουν ζητηθεί.
Keyword Search:	Η αναζήτηση με τη χρήση θεματικών όρων ή λέξεων-κλειδί που περιγράφουν το θέμα για το οποίο ο χρήστης επιθυμεί να βρει πληροφορίες.
Multi-Search Engines:	Η αναζήτηση κατά την οποία ο χρήστης, εισάγοντας μία φορά το ερώτημα

Όρος	***Ορισμός***
	του, έχει πρόσβαση σε δύο ή περισσότερες μηχανές αναζήτησης ταυτόχρονα.
Operator:	Οι τελεστές αποτελούν τους βοηθητικούς όρους που θα πρέπει να χρησιμοποιήσει ο χρήστης, όταν θελήσει να προσδιορίσει το είδος της πληροφορίας που περιμένει να αντλήσει από τη μηχανή αναζήτησης.
Phrase Search:	Η εισαγωγή μιας ολόκληρης φράσης στη μηχανή αναζήτησης, ακριβώς στη μορφή που ο χρήστης θέλει να αναζητηθεί. Σε αυτή την περίπτωση, οι λέξεις που απαρτίζουν τη φράση θα πρέπει να είναι η μία μετά την άλλη και να συνδέονται με ιούς κατάλληλους *τελεστές* ***(operators)*** που υποστηρίζει κάθε μηχανή.
Precision:	Η ακρίβεια των αποτελεσμάτων που θα επιστρέψει στο χρήστη η μηχανή αναζήτησης ή όποιο άλλο εργαλείο χρησιμοποιηθεί. Η ακρίβεια των αποτελεσμάτων προσδιορίζεται από τον αριθμό των επιτυχημένων αναφορών που επιστράφηκαν σε σχέση με το σύνολο αυτών που επιστράφηκαν από τη μηχανή αναζήτησης.
Proximity:	Η εγγύτητα με την οποία εμφανίζονται οι ζητούμενες λέξεις-όροι στα αποτελέσματα της αναζήτησης.
Query:	Το ερώτημα είναι συνήθως ένας συνδυασμός λέξεων-όρων που προσδιορίζουν το ζητούμενο θέμα.
Query by Example:	Το ερώτημα κατά το οποίο γίνεται χρήση ενός παραδείγματος προκειμένου να αναζητηθούν συναφείς με αυτό πληροφορίες.
Ranking:	Η κατάταξη των αποτελεσμάτων με βάση τη σχέση τους με το ζητούμενο θέμα.
Recall ή Ανάκληση:	Πάντα σε σχέση με την ακρίβεια των αποτελεσμάτων, πρόκειται για το ποσοστό ανάκλησης συναφών αποτελεσμάτων που επιτυγχάνει η μηχανή-εργαλείο αναζήτησης.
Relevance:	Η σχετικότητα των αποτελεσμάτων έτσι όπως αυτά επιστρέφονται από ένα ερώτημα.
Robot:	Το ειδικό λογισμικό που αναλαμβάνει να συγκεντρώνει και να ανανεώνει τις καταχωρίσεις σε μια βάση δεδομένων, σαρώνοντας τις web pages και τα περιεχόμενά τους (συνδεόμενες σελίδες και έγγραφα, συνδεόμενα sites,

Όρος	***Ορισμός***
	κ.λ.π.).
Search Engine (Μηχανή Αναζήτησης):	Η μηχανή αναζήτησης αποτελεί το κατεξοχήν εργαλείο έρευνας στις σελίδες του web. Το ειδικό λογισμικό της μηχανής αναλαμβάνει να εντοπίσει στις αποθηκευμένες σελίδες που διαθέτει τους όρους που εισήγαγε ο χρήστης και να επιστρέψει ως απάντηση τις σχετιζόμενες αναφορές, έγγραφα ή άλλες web pages.
Search Tool:	Το ειδικό λογισμικό που αναλαμβάνει να διεκπεραιώσει την αναζήτηση για ένα ερώτημα στο web.
Site:	Μια «τοποθεσία» στο web η οποία χαρακτηρίζεται από μια *μοναδική διεύθυνση* ***(Uniform Resource Locator, URL)***.
Spider:	Το ειδικό λογισμικό που αναλαμβάνει να συγκεντρώνει και να ανανεώνει τις καταχωρίσεις σε μια βάση δεδομένων, σαρώνοντας τις web pages και τα περιεχόμενά τους (συνδεόμενες σελίδες και έγγραφα, συνδεόμενα sites, κ.λ.π.).
Stemming:	Η χρήση της ρίζας μιας λέξεως στη διαδικασία της αναζήτησης, προκειμένου να ανακτηθούν παράγωγα της λέξεως αυτής ως αποτελέσματα της ερώτησης.
Thesaurus (θησαυρός):	Πρόκειται για λίστα συνώνυμων όρων τους οποίους χρησιμοποιούν τα εργαλεία αναζήτησης ως εναλλακτική λύση, στην περίπτωση που οι όροι που τέθηκαν προς αναζήτηση δεν μπορούν να βρεθούν.
Uniform Resource Locator (URL):	Η μοναδική διεύθυνση που χαρακτηρίζει ένα site στο web.
Σύνολο των Υπηρεσιών Αρχιτεκτονικής (Sum of Services Architecture)	Μια αρχιτεκτονική βασισμένη σε ένα σύνολο υπηρεσιών που διαχωρίζονται μεταξύ τους. Το σύνολο αυτών των υπηρεσιών παρέχει την πλήρη προοριζόμενη λειτουργία.

ΠΑΡΑΡΤΗΜΑ Α

Πίνακας Συνηθέστερων Stop-Words

A	EVEN	NAMELY	THERE
ABOUT	EVER	NEITHER	THEREAFTER
ABOVE	EVERY	NEVER	THEREBY
ACROSS	EVERYONE	NEVERTHELESS	THEREFORE
AFTER	EVERYTHING	NEXT	THEREIN
AFTERWARDS	EVERYWHERE	NO	THEREUPON
AGAIN	EXCEPT	NOBODY	THESE
AGAINST	FEW	NONE	THEY
ALL	FIRST	NOONE	THIS
ALMOST	FOR	NOR	THOSE
ALONE	FORMER	NOT	THOUGH
ALONG	FORMERLY	NOTHING	THROUGH
ALREADY	FROM	NOW	THROUGHOUT
ALSO	FURTHER	NOWHERE	THRU
ALTHOUGH	HAD	OF	THUS
ALWAYS	HAS	OFF	TO
AMONG	HAVE	OFTEN	TOGETHER
AMONGST	HE	ON	TOO
AN	HENCE	ONCE	TOWARD
AND	HER	ONE	TOWARDS
ANOTHER	HERE	ONLY	UNDER
ANY	HEREAFTER	ONTO	UNTIL
ANYHOW	HEREBY	OR	UP
ANYONE	HEREIN	OTHER	UPON
ANYTHING	HEREUPON	OTHERS	US
ANYWHERE	HERS	OTHERWISE	VERY
ARE	HERSELF	OUR	VIA
AROUND	HIM	OURS	WAS
AS	HIMSELF	OURSELVES	WE
AT	HIS	OUT	WELL

BE	HOW	OVER	WERE
BECAME	HOWEVER	OWN	WHAT
BECAUSE	I	PER	WHATEVER
BECOME	IE	PERHAPS	WHEN
BECOMES	IF	RATHER	WHENCE
BECOMING	IN	SAME	WHENEVER
BEEN	INC	SEEM	WHERE
BEFORE	INDEED	SEEMED	WHEREAFTER
BEFOREHAND	INTO	SEEMING	WHEREAS
BEHIND	IS	SEEMS	WHEREBY
BEING	IT	SEVERAL	WHEREIN
BELOW	ITS	SHE	WHEREUPON
BESIDE	ITSELF	SHOULD	WHEREVER
BESIDES	LAST	SINCE	WHETHER
BETWEEN	LATTER	SO	WHITHER
BEYOND	LATTERLY	SOME	WHICH
BOTH	LEAST	SOMEHOW	WHILE
BUT	LESS	SOMEONE	WHO
BY	LTD	SOMETHING	WHOEVER
CAN	MANY	SOMETIME	WHOLE
CANNOT	MAY	SOMETIMES	WHOM
CO	ME	SOMEWHERE	WHOSE
COULD	MEANWHILE	STILL	WHY
DOWN	MIGHT	SUCH	WILL
DURING	MORE	THAN	WITH
EACH	MOREOVER	THAT	WITHIN
EG	MOST	THE	WITHOUT
EITHER	MOSTLY	THEIR	WOULD
ELSE	MUCH	THEM	YET
ELSEWHERE	MUST	THEMSELVES	YOU
ENOUGH	MY	THEN	YOUR
ETC	MYSELF	THENCE	YOURS

ΠΑΡΑΡΤΗΜΑ Β

Ενδεικτικός Πίνακας Καταλήξεων

ABILITIES
ABILITY
ABLE
ABLED
ABLEDLY
ABLENESS
ABLENESSES
ABLER
ABLES
ABLING
ABLINGFUL
ABLINGLY
ABLY
ACEOUS
ACEOUSLY
ACEOUSNESS
ACEOUSNESSES
ACIES
ACIDOUS
ACIDOUSLY
ACIOUSNESS
ACIOUSNESSES
ACITIES
ACITY
ACY
AE
AGE
AGED
AGEDLY
ACER
ALISTS
ALITIES
ALITY
ALIZATION
ALIZATIONAL
ALIZATIONALLY
AL1ZATIONS
ALIZE
ALIZED
ALIZEDLY
ALIZER
ALIZES
ALIZING
ALIZINGFUL
ALIZINGLY
ALIZINGS
ALLED
ALLEDLY
ALLIC
ALLICALLY
ALLICISM
ALLICISMS
ALLICS
ALLING
ALLINGFUL
ALLINGLY
ALLMENT
ALLY
ALMENT
ALNESS
ANTINGLY
ANTINGS
ANTLY
ANTMENT
ANTMENTS
ANTRESS
ANTRESSES
ANTRY
ANTS
AR
ARIAL
ARIALS
ARIAN
ARIANS
ARIC
ARJCISM
ARICISMS
ARICS
ARIES
ARILINESS
ARILY
ARINESS
ARINESSES
ARISABILITIES
ARISABILITY
ARISABLE
ARISATION
ARISATIONS
ARISE
ARISED
AROIDS
ARS
ARY
ASIS
ASISE
ASISEABLE
ASISED
ASISEDLY
ASISER
ASISES
ASISING
ASISINGFUL
ASISINGLY
ASISINGS
ASIZABLE
ASIZE
ASIZED
ASIZEDLY
ASIZER
AS1ZES
ASIZING
ASIZINGFUL
ASIZINGLY
ASIZINGS
ASM
ASMS
AST
ASTIC
ASTICAL
ASTICALLY

AGES
AGING
AGTNGFUL
AGINGLY
AIC
AICAL
AICALLY
AICALS
AICISM
AICISMS
AICS
AL
ALISATION
ALISATIONAL
ALISATIONALLY
ALISATIONS
ALISE
ALISED
ALISEDLY
ALISER
ALISES
ALISING
ALISINGFUL
ALISINGLY
ALISINGS
ALISM
ALISMS
ALIST
ALISTIC
ALISTICALLY
ALISTICISM
ALISTICISMS
ALISTICS
ALNESSES
ALS
ANCE
ANCED
ANCEDLY
ANGER
ANCES
ANCIAL
ANCIALS
ANCIES
ANCING
ANCINGFUL
ANCINGLY
ANCINGS
ANCY
ANEOUS
ANEOUSLY
ANEOUSNESS
ANT
ANTANEOUS
ANTANEOUSLY
ANTED ANTEDLY
ANTIALNESS
ANTIALNESSES
ANTIC
ANTICISM
ANTICISMS
ANTICS
ANTING
ANTINGFUL
ANTED ANTEDLY
ANTING
ANTINGFUL
ARISEDLY
ARISER
ARISES
ARISING
ARISINGFUL
ARISINGLY
ARISINuS
ARISM
ARISMS
ARIST
ARISTIC
ARISTICISM
ARISTICISMS
ARISTICS
ARISTS
ARITIES
ARITY
ARIZABILITIES
ARIZABILITY
ARIZABLE
ARIZATION
ARIZATIONS
ARIZE
ARISED
ARISEDLY
ARIZER
ARIZES
ARIZING
ARIZINGFUL
ARJZINGLY
ARIZINGS
ARLY
AROID
ASTICISM
ASTICISMS
ASTICS
ASTMENT
ASTMENTS
ASTRIES
ASTRY
ASTS
ASY
ATA
ATABILITIES
ATABILITY
ATABLE
ATABLES
ATABLY
ATAL
ATE
ATED
ATEDLY
ATELY
ATENESS
ATENESSES
ATER
ATES
ATIC
ATICAL
ATICALLY
ATICISM
ATICISMS
ATICS
ATING

ΠΑΡΑΡΤΗΜΑ Γ

Τα Κυριότερα Εργαλεία Αναζήτησης

Μηχανές Αναζήτησης:

AltaVista: http://www.altavista.com

DirectHit: http://www.directhit.com

Excite: http://www.excite.com

Google: http://www.google.com

GoTo: http://www.goto.com

HotBot: http://www.hotbot.com

Infoseek: http://www.infoseek.go.com

Lycos: http://www.lycos.com

NorthernLight: http://www.northernlight.com

Web Crawler: http://www.webcrawler.com

Μετα-Μηχανές Αναζήτησης:

All-in-One: http://www.albany.net/allinone

Highway61 : http://www.highway61.com

Mamma: http://www.mamma.com

Metacrawler: http://www.metacrawler.com

Metasearch: http://www.metasearch.com

OneSeek: http://www.oneseek.com

Θεματικά Ευρετήρια:

LookSmart: http://www.looksmart.com

NetSearch: http://www.ais.net/netsearch/search_entry.html

One-Global: http://one-global.com

Virtual Yellow Pages: http://www.vyp.com/yp/search.html

Yahoo!: http://www.yahoo.com

ΠΑΡΑΡΤΗΜΑ Δ

Αλγόριθμοι Για Το Ειδικό Λογισμικό (Crawlers, Robots, Spiders)

Στο παράρτημα αυτό παρατίθενται μερικοί αλγόριθμοι που χρησιμοποιούνται από το ειδικό λογισμικό για να επιτευχθεί το crawling.

❑ BackLink-Based Crawlers

Crawling algorithm (backward link based)

```
enqueue(url_queue, starting_url);
while (not empty(url_queue)) {
    url = dequeue(url_queue);
    page = crawl_page(url);
    enqueue(crawled_pages, (url, page));
    url_list = extract_urls(page);
    for each u in url_list
        enqueue(links, (url, u));
        if [u not in url_queue] and
           [(u,-) not in crawled_pages]
            enqueue(url_queue, u);
    reorder_queue(url_queue);
}
```

Function description

`enqueue(queue, element)`	:	append element at the end of queue.
`dequeue(queue)`	:	remove the element at the beginning of queue and return it.
`reorder_queue(queue)`	:	reorder queue using information in links. Refer to Fig 2.

❑ Similarity-Based Crawlers

Crawling algorithm (similarity-based)

```
enqueue(url_queue, starting_url);
while (not empty(hot_queue) and not empty(url_queue)) {
     url = dequeue2(hot_queue, url_queue);
     page = crawl_page(url);
     enqueue(crawled_pages, (url, page));
     url_list = extract_urls(page);
     for each u in url_list
          enqueue(links, (url, u));
          if [u not in url_queue] and
             [u not in hot_queue] and
             [(u,-) not in crawled_pages]
              if [u contains computer in anchor or url]
                   enqueue(hot_queue, u);
              else
                   enqueue(url_queue, u);
     reorder_queue(url_queue);
     reorder_queue(hot_queue);
}
```

Function description

```
dequeue2(queue1, queue2) :
     if (not empty(queue1)) dequeue(queue1);
     else                   dequeue(queue2);
```

❑ Τροποποιημένος Similarity-Based Crawling Αλγόριθμος

Crawling algorithm (modified similarity-based)

```
enqueue(url_queue, starting_url);
while (not empty(hot_queue) and not empty(url_queue)) {
   url = dequeue2(hot_queue, url_queue);
   page = crawl_page(url);
   if [page contains 10 or more computer in body
       or one computer in title]
      hot[url] = TRUE;
   enqueue(crawled_pages, (url, page));
   url_list = extract_urls(page);
   for each u in url_list
       enqueue(links, (url, u));
       if [u not in url_queue] and
          [u not in hot_queue] and
          [(u,-) not in crawled_pages]
           if [u contains computer in anchor or url]
               enqueue(hot_queue, u);
           else if [distance_from_hotpage(u) < 3]
               enqueue(hot_queue, u);
           else
               enqueue(url_queue, u);
   reorder_queue(url_queue);
   reorder_queue(hot_queue);
}
```

Function description

```
distance_from_hotpage(u) :
   return 0 if [hot[u] = TRUE];
   return 1 if [hot[v] = TRUE] and [(v, u) in links]
             for some v;
   return 2 if [hot[v] = TRUE] and
               [(v, w) in links] and [(w, u) in links]          for
   some v, w;
```

ΠΑΡΑΡΤΗΜΑ Ε

Τα Κυριότερα Χαρακτηριστικά των Εργαλείων Αναζήτησης (1)

Στο συγκεκριμένο παράρτημα υπάρχουν συγκεντρωμένα με τη μορφή πίνακα όλα τα χαρακτηριστικά των εργαλείων αναζήτησης. Σε αυτό το παράρτημα βρίσκονται τέσσερις πίνακες οι οποίοι περιγράφουν τα χαρακτηριστικά για κάθε εργαλείο αναζήτησης. Υπάρχουν σε αυτούς τους πίνακες περισσότερα εργαλεία από αυτά που αναφέρθηκαν στα προηγούμενα κεφάλαια. Με βάση αυτούς τους πίνακες, καθώς και τους πίνακες του επόμενου παραρτήματος (ΠΑΡΑΡΤΗΜΑ ΣΤ) προέκυψαν οι πίνακες του κεφαλαίου 8, οι οποίοι αποτέλεσαν τη βάση της ανάπτυξης του ***Ενιαίου Προτύπου Αναζήτησης***.

Οι πίνακες που υπάρχουν είναι οι παρακάτω:

- ***ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ CRAWLING-RANKING ΚΑΙ BOOLEAN***
- ***ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΕΥΡΕΤΗΡΙΑΣΗΣ***
- ***ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΕΜΦΑΝΙΣΗΣ ΚΑΙ ΕΙΔΙΚΟΥ ΛΟΓΙΣΜΙΚΟΥ***
- ***ΠΙΝΑΚΑΣ ΔΙΑΦΟΡΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ***

	Εργαλεία Αναζήτησης	Crawling							Ranking				Boolean και Τελεστές Εγγύτητας
		Εξαντλητικό Crawl	Άμεση Ευρετηρίαση	Υποστήριξη Πλαισίων	Εικόνες	Robots.txt	Meta Robots Tag	Αλλαγή Σελίδων	Meta Tags	Reviewed Status	Πλήθος Συνδέσεων	Direct Hit	
Μηχανές Αναζήτησης	**AltaVista**	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ	AND, OR, AND NOT, NEAR στην προχωρημένη αναζήτηση
	Google	ΝΑΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ	AND, το OR δεν τον υποστηρίζει
	HotBot					ΝΑΙ	ΝΑΙ					ΟΧΙ	AND, OR, NOT Αναζήτηση με βάση την εγγύτητα δεν παρέχεται
	Infoseek					ΝΑΙ	ΝΑΙ					ΟΧΙ	ΟΧΙ
	Lycos	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	AND, OR, NOT
	AllTheWeb	ΝΑΙ				ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Teoma	ΟΧΙ				ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Aol Search					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Northernlight					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Direct Hit					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Inktomi	ΝΑΙ				ΝΑΙ	ΝΑΙ					ΟΧΙ	
Θεματικά Ευρετήρια	**Yahoo!**					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	LookSmart					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	MSN					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Excite	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ	AND, OR, AND NOT
Μηχανές Πολλαπλής Αναζήτησης	**Metacrawler**					ΝΑΙ	ΝΑΙ					ΟΧΙ	Το (+) προσθέτει όρους, το (-) αφαιρεί όρους
	Ixquick					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	ProFusion					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Dogpile					ΝΑΙ	ΝΑΙ					ΟΧΙ	AND, OR, NOT, NEAR, Παρενθέσεις ()
	SavvySearch					ΝΑΙ	ΝΑΙ					ΟΧΙ	AND, OR, NOT, NEAR, Παρενθέσεις ()
	Search.com					ΝΑΙ	ΝΑΙ					ΟΧΙ	
	Cyber411					ΝΑΙ	ΝΑΙ					ΟΧΙ	AND, OR, NOT, NEAR, Παρενθέσεις ()
	Highway61					ΝΑΙ	ΝΑΙ					ΟΧΙ	AND, OR
	Ask Jeeves					ΝΑΙ	ΝΑΙ					ΟΧΙ	ΟΧΙ
	Mamma					ΝΑΙ	ΝΑΙ					ΟΧΙ	Το (+) προσθέτει όρους, το (-) αφαιρεί όρους
	SurfWax					ΝΑΙ	ΝΑΙ					ΟΧΙ	

ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ CRAWLING-RANKING ΚΑΙ BOOLEAN

	Εργαλεία Αναζήτησης	Indexing											
		Πηγές που Ευρετηριάζονται	Μέρη της Σελίδας που Ευρετηριάζονται	Αποκοπή (Truncation)	Αναζήτηση Φράσεων	Αναζήτηση Μικρών-Κεφαλαίων	Πλήρες Κείμενο	Stop Words	Meta Description	Meta Keywords	ALT Text	Σχόλια	Stemming
Μηχανές Αναζήτησης	AltaVista	Web pages, Usenet, Αναζήτηση Ατόμων (ταχυδρομικοί κώδικες, αριθμοί τηλεφώνων), Αναζήτηση Επιχειρήσεων	Ολόκληρη η σελίδα	Η αναζήτηση των όρων γίνεται με βάση την αρχική τους μορφή	Μέσα σε διπλά εισαγωγικά (")	Χρησιμοποιώντας κεφαλαία θα γίνει ακριβής αναζήτηση της λέξης	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ		
	Google	Web pages	Ολόκληρη η σελίδα	Η αναζήτηση των όρων γίνεται με βάση την αρχική τους μορφή	Μέσα σε διπλά εισαγωγικά (")	ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ		
	HotBot	Web pages, Usenet, κυριότερες ειδήσεις, e-mail διευθύνσεις, ταχυδρομικοί κώδικες	Ολόκληρη η σελίδα	Η αναζήτηση των όρων γίνεται με βάση την αρχική τους μορφή	Μέσα σε διπλά εισαγωγικά (")	Μόνο σε συγκεκριμένες λέξεις	ΝΑΙ		ΝΑΙ	ΝΑΙ			ΝΑΙ
	Infoseek	Web pages, Usenet, ειδήσεις, e-mail διευθύνσεις, ταχυδρομικοί κώδικες, αριθμοί τηλεφώνων, Yellow Pages	Ολόκληρη η σελίδα	Οι αναζητήσιμοι όροι μετατρέπονται στις ρίζες τους	Μέσα σε διπλά εισαγωγικά (")	Μόνο σε συγκεκριμένες λέξεις και ιδιαίτερες περιπτώσεις	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Lycos	Web pages, ειδήσεις του Reuter, Αναζήτηση Ατόμων (e-mail διευθύνσεις, ταχυδρομικοί κώδικες, αριθμοί τηλεφώνων), Yellow Pages	Δημιουργεί περιλήψεις των σελίδων βασιζόμενη στις κεφαλίδες, στους τίτλους, στις συνδέσεις και στις πρώτες λιγοστές λέξεις των παραγράφων κλειδιών (key paragraphs)	Η αναζήτηση των όρων γίνεται με βάση την αρχική τους μορφή	Μέσα σε διπλά εισαγωγικά (")	ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ		
	AllTheWeb						ΝΑΙ		ΝΑΙ	ΝΑΙ	ΟΧΙ		
	Teoma						ΝΑΙ		ΝΑΙ	ΝΑΙ	ΝΑΙ		
	Aol Search						ΝΑΙ		ΝΑΙ	ΝΑΙ			ΝΑΙ
	Northernlight						ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Direct Hit						ΝΑΙ		ΝΑΙ	ΝΑΙ			ΝΑΙ
	Inktomi						ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΝΑΙ	ΝΑΙ
Θεματικά Ευρετήρια	Yahoo!						ΝΑΙ		ΟΧΙ	ΟΧΙ			
	LookSmart						ΝΑΙ		ΝΑΙ	ΝΑΙ			
	MSN						ΝΑΙ		ΝΑΙ	ΝΑΙ			ΝΑΙ
	Excite	Web pages, Πληροφορίες Εταιρειών, Αναζήτηση Ατόμων (ταχυδρομικοί κώδικες, αριθμοί τηλεφώνων), Yellow Pages	Ολόκληρη η σελίδα	Οι αναζητήσιμοι όροι μετατρέπονται στις ρίζες τους	Μέσα σε διπλά εισαγωγικά (")	Βάζοντας κεφαλαία στο πρώτο γράμμα κάθε λέξης	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ		
Μηχανές Πολλαπλής Αναζήτησης	Metacrawler			ΟΧΙ	Μέσα σε διπλά εισαγωγικά (") ή Επιλέγει ο χρήστης το "Phrase Search"	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Ixquick						ΝΑΙ		ΝΑΙ	ΝΑΙ			
	ProFusion						ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Dogpile			ΟΧΙ	Μέσα σε διπλά εισαγωγικά (")	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	SavvySearch			ΟΧΙ	Μέσα σε διπλά εισαγωγικά (") ή Επιλέγει ο χρήστης το "Phrase Search"	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Search.com						ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Cyber411			ΟΧΙ	Μέσα σε διπλά εισαγωγικά (")	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Highway61			ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Ask Jeeves			ΟΧΙ	όπως την έχει διατυπώσει ο χρήστης	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	Mamma			ΟΧΙ	Επιλέγει ο χρήστης το "Phrase Search"	ΟΧΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			
	SurfWax						ΝΑΙ		ΝΑΙ	ΝΑΙ			

	Εργαλεία Αναζήτησης	Χαρακτηριστικά Εμφάνισης							Ειδικό Λογισμικό					
		Αριθμός των Συνδέσεων	20 Αποτελέσματα	50 Αποτελέσματα	100 Αποτελέσματα	Ταξινόμηση με Βάση την Ημερομηνία	Εμφάνιση Τίτλων	Εμφάνιση Ημερομηνίας	Αναζήτηση Τίτλων	Αναζήτηση Site	Αναζήτηση URL	Αναζήτηση Συνδέσεων	Wildcard	Αναζήτηση Anchor
Μηχανές Αναζήτησης	**AltaVista**	ΝΑΙ	ΝΑΙ	ΝΑΙ			ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ
	Google	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ				ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ
	HotBot	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ		ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ
	Infoseek													
	Lycos	ΝΑΙ								ΟΧΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ
	AllTheWeb	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ				ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ
	Teoma								ΝΑΙ					
	Aol Search	ΝΑΙ								ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ
	Northernlight	ΝΑΙ				ΝΑΙ		ΝΑΙ		ΟΧΙ	ΝΑΙ	ΟΧΙ	ΝΑΙ	
	Direct Hit	ΝΑΙ								ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ
	Inktomi								ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΝΑΙ	ΟΧΙ
Θεματικά Ευρετήρια	**Yahoo!**									ΟΧΙ	ΝΑΙ			
	LookSmart	ΝΑΙ								ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	
	MSN	ΝΑΙ	ΝΑΙ	ΝΑΙ		ΝΑΙ	ΝΑΙ			ΟΧΙ	ΟΧΙ	ΝΑΙ	ΟΧΙ	
	Excite	ΝΑΙ	ΝΑΙ	ΝΑΙ			ΝΑΙ			ΝΑΙ	ΝΑΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ
Μηχανές Πολλαπλής Αναζήτησης	**Metacrawler**													
	Ixquick													
	ProFusion													
	Dogpile													
	SavvySearch													
	Search.com													
	Cyber411													
	Highway61													
	Ask Jeeves													
	Mamma													
	SurfWax													

ΠΙΝΑΚΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΕΜΦΑΝΙΣΗΣ ΚΑΙ ΕΙΔΙΚΟΥ ΛΟΓΙΣΜΙΚΟΥ

	Εργαλεία Αναζήτησης	Διάφορα Χαρακτηριστικά				
		Clustering	Εύρεση Ομοίων	Αναζήτηση ανάμεσα στα Αποτελέσματα	Αναζήτηση με βάση τη Γλώσσα	Μετάφραση Σελίδων
Μηχανές Αναζήτησης	AltaVista	NAI	NAI	NAI	NAI	NAI
	Google	NAI	NAI	NAI	NAI	NAI
	HotBot	NAI		NAI	NAI	
	Infoseek					
	Lycos			NAI	NAI	NAI
	AllTheWeb	NAI			NAI	
	Teoma					
	Aol Search		NAI			
	Northernlight	NAI			NAI	
	Direct Hit					
	Inktomi					
Θεματικά Ευρετήρια	Yahoo!					
	LookSmart					
	MSN	NAI			NAI	
	Excite	NAI			NAI	
Μηχανές Πολλαπλής Αναζήτησης	Metacrawler					
	Ixquick					
	ProFusion					
	Dogpile					
	SavvySearch					
	Search.com					
	Cyber411					
	Highway61					
	Ask Jeeves					
	Mamma					
	SurfWax					

ΠΙΝΑΚΑΣ ΔΙΑΦΟΡΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

ΠΑΡΑΡΤΗΜΑ ΣΤ

Τα Κυριότερα Χαρακτηριστικά των Εργαλείων Αναζήτησης (2)

Στο συγκεκριμένο παράρτημα υπάρχουν συγκεντρωμένα με τη μορφή πίνακα όλα τα χαρακτηριστικά των εργαλείων αναζήτησης. Σε αυτό το παράρτημα βρίσκονται πέντε πίνακες οι οποίοι περιγράφουν τα χαρακτηριστικά για κάθε εργαλείο αναζήτησης. Η διαφορά με τους πίνακες του προηγούμενου παραρτήματος είναι ότι οι πίνακες αυτοί αναφέρουν ποια εργαλεία έχουν κάποιο συγκεκριμένο χαρακτηριστικό και όχι για κάθε εργαλείο αναζήτησης ξεχωριστά. Υπάρχουν σε αυτούς τους πίνακες περισσότερα εργαλεία από αυτά που αναφέρθηκαν στα προηγούμενα κεφάλαια. Με βάση αυτούς τους πίνακες, καθώς και τους πίνακες του προηγούμενου παραρτήματος (ΠΑΡΑΡΤΗΜΑ Ε) προέκυψαν οι πίνακες του κεφαλαίου 8, οι οποίοι αποτέλεσαν τη βάση της ανάπτυξης του ***Ενιαίου Προτύπου Αναζήτησης***.

Crawling				Διάφορα Χαρακτηριστικά	
Crawling	**Ναι**	**Όχι**	**Παρατηρήσεις**	**Χαρακτηριστικό**	**Εργαλεία Αναζήτησης**
Εξαντλητικό Crawl (Deep Crawl)	AllTheWeb, AltaVista, Google, Inktomi	Teoma, Excite, Lycos		Related Searches	AltaVista, AllTheWeb, Excite, HotBot, Lycos, MSN, Yahoo Not yet updated, but may be still correct: iWon
Άμεση Ευρετηρίαση (Instant Indexing)	AltaVista	Excite, Lycos, Google	Οι σελίδες εμφανίζονται σε μια ή δυο μέρες μετά την καταχώρησή τους	Clustering	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN, Northern Light
Υποστήριξη Πλαισίων (Frames Support)	AltaVista, Google	Excite, Lycos	Lycos παρέχει την ελάχιστη υποστήριξη	Εύρεση Ομοίων (Find Similar)	AltaVista, AOL Search, Google
Καταγραφή Εικόνων (Image Maps)	AltaVista	Excite, Google, Lycos		Search Within	AltaVista, Google, HotBot, Lycos
Robots.txt	Όλες	-		Spidered Version	Google
Meta Robots Tag	Όλες	-		Search By Language	AltaVista, AllTheWeb, Excite, Google, HotBot, Lycos, MSN, Northern Light
Πλήθος Συνδέσεων που Βοηθά στο Εξαντλητικό Crawl	Lycos	Excite, AltaVista		Μετάφραση Σελίδων (Page Translation)	AltaVista, Google, Lycos
Συχνότητα Αλλαγών της Σελίδας	AltaVista	Excite, Google, Lycos		Porn Filter	AltaVista, AllTheWeb, Google
Indexing				Porn Warning	HotBot, MSN, Northern Light
Indexing	**Ναι**	**Όχι**	**Παρατηρήσεις**	**Χαρακτηριστικό Εμφάνισης**	**Εργαλεία Αναζήτησης**
Πλήρες Κείμενο (Full Body Text)	Όλες	-	Μερικές stop words μπορεί να μην ευρετηριάζονται	Number Of Listings Shown (10 unless noted)	AltaVista, AllTheWeb, AOL Search (5), Direct Hit, Excite, Google, HotBot, LookSmart (15), Lycos, MSN (15), Northern Light Not yet updated, but may be still correct: iWon, Netscape, Yahoo (20)
Stop Words	AltaVista, Inktomi, Lycos, Google, Excite	FAST		Ability To Increase Number Of Listings	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN Not yet updated, but may be still correct: Yahoo
Meta Description	Όλες	Google, Lycos, Yahoo!	AllTheWeb, AltaVista, Teoma χρησιμοποιούν περισσότερο αυτό το tag	20 Αποτελέσματα	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN Not yet updated, but may be still correct: Yahoo
Meta Keywords	Όλες	Excite, Google, Lycos, Yahoo!		50 Αποτελέσματα	AltaVista, AllTheWeb, Excite, Google, HotBot, MSN Not yet updated, but may be still correct: Yahoo
ALT Text	AltaVista, Lycos, Google, Teoma	Excite, AllTheWeb, Inktomi		100 Αποτελέσματα	AllTheWeb, Google, HotBot, Not yet updated, but may be still correct: Yahoo
Σχόλια	Inktomi	Άλλες		Ταξινόμηση με βάση την ημερομηνία (Sort By Date)	MSN Search, Northern Light
Stemming	AOL Search, Direct Hit, HotBot, Inktomi (HotBot, MSN)			Date Range	AltaVista, Google, HotBot, MSN, Northern Light Not yet updated, but may be still correct: iWon, Yahoo
Ranking				Εμφάνιση Ημερομηνίας (Date Displayed)	AltaVista, HotBot (for Inktomi results), Northern Light
Ranking	**Ναι**	**Όχι**	**Παρατηρήσεις**	Εμφάνιση Τίτλων (Display Titles)	AltaVista, Excite, HotBot (URLs only option), MSN
Meta Tags βοηθούν στην Ταξινόμηση	Καμιά	AltaVista, Lycos, Google, Excite		Other Major Customize Options	AltaVista, AllTheWeb, Google
Reviewed Status βοηθά στην Ταξινόμηση	Καμιά	AltaVista, Lycos, Google, Excite			
Το πλήθος συνδέσεων βοηθά στην Ταξινόμηση	AltaVista, Google, Excite	Lycos	Πολύ σημαντικό για το Google		
Direct Hit βοηθά στην Ταξινόμηση	Καμιά	Όλες			

Χαρακτηριστικά	None	AllTheWeb, AOL Search, Direct Hit, Excite, Google, Inktomi, HotBot, Lycos
Αναζήτηση Τίτλων (Title Search)	title:	AltaVista, AllTheWeb, Inktomi
	intitle:	Google, Teoma
	allintitle:	Google
Αναζήτηση Site (Site Search)	host:	AltaVista
	site:	Excite, Google (Netscape, Yahoo)
	url.host:	AllTheWeb, Lycos (for AllTheWeb results only)
	domain:	Inktomi (HotBot, iWon, LookSmart)
	none	AOL, Direct Hit, HotBot, LookSmart, Lycos, MSN, Netscape, Northern Light, Open Directory, Yahoo
Αναζήτηση URL (URL Search)	url:	AltaVista, Excite, Northern Light
	url.all:	AllTheWeb, Lycos (for AllTheWeb results only)
	allinurl: inurl:	Google
	originurl:	Inktomi (AOL, GoTo, HotBot)
	u:	Yahoo
	none	AOL, Direct Hit, HotBot, LookSmart, MSN **Not yet updated, but may be still correct:** Open Directory
Αναζήτηση Συνδέσεων (Link Search)	link:	AltaVista, Google, Northern Light
	linkdomain:	Inktomi (AOL, HotBot, iWon, MSN) (NOTE: measures links to entire domains)
	link.all:	AllTheWeb, Lycos (for AllTheWeb results only)
	none	AOL, Direct Hit, Excite, HotBot, LookSmart, Northern Light **Not yet updated, but may be still correct:** Netscape, Yahoo (n/a)
Wildcard	*	AltaVista, Inktomi (iWon), Northern Light **Not yet updated, but may be still correct:** Yahoo
	?	AOL Search, Inktomi (iWon)
	%	Northern Light
	none	AllTheWeb, Direct Hit, Excite, Google, HotBot, LookSmart, Lycos, MSN (MSN's help says it offers wildcard, but it failed to during testing)
Αναζήτηση Anchor	anchor:	AltaVista

Boolean		Εργαλεία Αναζήτησης
Or	OR	AltaVista, AOL Search, Excite, Google, Inktomi (HotBot, MSN), Lycos, Northern Light
	None	AllTheWeb, Direct Hit, LookSmart, **Not yet updated, but may be still correct:** Yahoo
And	AND	AltaVista, AOL Search, Excite, Inktomi (HotBot, MSN) Lycos, Northern Light
	None	AllTheWeb, Direct Hit, Google, LookSmart **Not yet updated, but may be still correct:** Yahoo
Not	NOT	AOL Search, Excite, Inktomi (HotBot), Lycos, Northern Light
	AND NOT	AltaVista, Inktomi (MSN) **Not yet updated, but may be still correct:** Netscape
	None	AllTheWeb, Direct Hit, Google, LookSmart, **Not yet updated, but may be still correct:** Yahoo
Nesting	( )	AltaVista, AOL Search, Excite, Inktomi (MSN), Northern Light
	None	AllTheWeb, Direct Hit, Google, Inktomi (HotBot), LookSmart, Lycos **Not yet updated, but may be still correct:** Yahoo
Near	NEAR	AltaVista (10 words), AOL Search (specify number), Lycos (25 words)
	None	AllTheWeb, Direct Hit, Google, Inktomi (HotBot, MSN), LookSmart
Notes At AltaVista, Boolean only works on advanced search page. At Excite, Google & MSN, Boolean commands must be in UPPERCASE At Inktomi-powered services, set menu to "Boolean"		

Μηχανές Αναζήτησης	Πήγές που Ευρετηριάζονται στο Internet (Types of Internet Resources Indexed)	Μέρη της Σελίδας που Ευρετηριάζονται (Parts of Web Pages Indexed)
Alta Vista	Web pages, Usenet newsgroups, *People Search* (i.e., postal addresses, phone numbers), *Business Search* (by category or name)	Ολόκληρη η σελίδα
Excite	Web pages, Company information, Stock quotes, Email addresses, *People Finder* (i.e., postal addresses, phone numbers), Yellow pages ***Power Search* Options:** *Selected Web sites* (i.e., selected by Excite), *Current news*, Excite UK, Excite France, Excite Germany, Excite Sweden	Ολόκληρη η σελίδα
Google	Web pages	Ολόκληρη η σελίδα
HotBot	Web pages, Usenet newsgroups, *Top news sites*, Email adresses, Postal addresses, etc.	Ολόκληρη η σελίδα
Infoseek	Web pages, Usenet newsgroups, News, Company profiles (US public and private companies), Stock quotes, Email addresses, Postal addresses, Phone numbers, Yellow pages	Ολόκληρη η σελίδα
Lycos	Web pages, Reuters news, *Top 5%* (i.e., Web site reviews), Cities, Stock quotes, *PeopleFind* (e.g., Email addresses, Postal addresses, Phone numbers, Business Web sites), etc.	Δημιουργεί περιλήψεις των σελίδων βασιζόμενη στις κεφαλίδες, στους τίτλους, στις συνδέσεις και στις πρώτες λιγοστές λέξεις των παραγράφων κλειδιών (key paragraphs)

Μηχανές Αναζήτησης	Αυτόματη Στρατηγική Αναζήτησης (Default Search Strategy)	Αποκοπή (Truncation)	Boolean και τελεστές Εγγύτητας (Boolean & Proximity Operators)	Αναζήτηση Φράσεων (Phrase Searching)	Αναζήτηση Μικρών και κεφαλαίων (Case-Sensitive Searching)
Alta Vista	***OR operator implied*** Search results ranked by location and frequency of search terms on Web pages --- + require term in results - exclude term in results ***Simple Searches Only***	***Search terms are searched as typed in*** **Note:** Placing an * after a search term will find variant endings **Example:** imagin* will find imagine, imagines, etc.	**AND, OR, AND NOT, NEAR (i.e., within 10 words)** ***Advanced Searches Only*** **Note:** Operators may be typed in upper or lower case letters	***Enclose phrase within double quotation marks (")*** **Example:** "better business bureau"	***Using capital letters will force an exact case match on the entire word*** **Example:** ***"NeXt"*** will only find the term spelled with an upper case ***"N" and "X"***
Excite Search	***"Concept search"*** Searches Web pages for search terms as typed in **and also** for Web pages with terms ***related*** to the user specified search terms --- + require term in results - exclude term in results	***Search terms default to their apparent stems*** **Example:** imagination is stemmed to imagine	***AND, OR, AND NOT*** **Note:** Operators **must be** typed in **ALL UPPER CASE LETTERS "with a space on each side"**	***Enclose phrase within double quotation marks (")*** **Example:** "better business bureau"	***Capitalize the first letter of each word in a proper name*** **Example:** Bill Gates
HotBot	***AND operator implied*** Search results ranked by frequency of search terms on Web pages --- + require term in results - exclude term in results	***Search terms are searched as typed in***	***AND, OR, NOT*** **Note:** Proximity searching not available	***Enclose phrase within double quotation marks (") or select "exact phrase" from the "Look for" drop down menu*** **Example:** "better business	***Case sensitive searches will be performed for words that have "interesting case" (i.e., "words with mixed upper and lower case characters")***

				bureau"	
Google	***AND operator implied.*** Web pages returned are ranked by the number of "high importance" pages linked to them + requires "stop word" in results	***Search terms are searched as typed in*** **Note:** Only pages that contain **ALL** the words will be retrieved	***AND*** **Note:** Does not support the operator **OR**	***Enclose phrase within double quotation marks (")*** **Example:** "better business bureau"	***Not Case Sensitive*** **Example:** "Bill Gates" and "bill gates" will retrieve the same web sites
Infoseek	***Or operator implied*** Search results ranked by location and frequency of search terms on Web pages + require term in results **-** exclude term in results	***Search terms default to their apparent stems*** **Example:** "imagination" is stemmed to "imagine"	***No***	***Enclose phrase within double quotation marks (")*** Adjacent capitalized words are automatically treated as a single phrase **Examples:** "better business bureau" "Bill Gates"	***Case is ignored when the search term is typed in all lowercase letters. Use of mixed upper and lowercase letters will force an exact case match on the entire word.*** **Example:** ***"NeXt"*** will only find the term spelled with an upper case ***"N" and "X"***
Lycos	***AND operator implied*** Web pages containing all user specified search terms are listed first in the results + require term in results **-** exclude term in results	***Search terms are searched as typed in***	***AND, OR, NOT*** **Note:** Operators may be typed in upper or lower case letters	***Enclose phrase within double quotation marks (")*** **Example:** "better business bureau"	***No***