ΤΕΙ ΚΡΗΤΗΣ-ΠΑΡΑΡΤΗΜΑ ΧΑΝΙΩΝ
ΤΜΗΜΑ:ΤΗΛΕΠΙΚΟΙΝΩΝΙΕΣ & ΔΙΚΤΥΑ Η/Υ
(ΠΣΕ)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ
ΠΩΣ ΝΑ ΒΡΩ ΑΥΤΟ ΠΟΥ ΘΕΛΩ ΣΤΟ INTERNET

ΣΠΟΥΔΑΣΤΗΣ: ΔΕΣΥΛΛΑΣ ΑΛΕΞΑΝΔΡΟΣ

ΕΠΙΒΛΕΠΩΝ: ΜΑΝΩΛΗΣ ΑΝΤΩΝΙΔΑΚΗΣ

Σεπτέμβριος 2005

ΕΝΟΤΗΤΕΣ:

1.ΕΙΣΑΓΩΓΗ

2.ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

3.ΟΙ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ ΓΕΝΙΚΑ

4.ΤΡΟΠΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

5. ΤΡΟΠΟΙ ΑΝΑΖΗΤΗΣΗΣ ΚΑΙ ΕΡΓΑΛΕΙΑ

6.ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

7.ΜΗΧΑΝΕΣ ΠΟΛΛΑΠΛΗΣ ΑΝΑΖΗΤΗΣΗΣ

8.ΟΙ ΠΙΟ ΓΝΩΣΤΕΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

9. ΓΛΩΣΣΑΡΙ

10. ΒΙΒΛΙΟΓΡΑΦΙΑ

1. ΕΙΣΑΓΩΓΗ

Πολλοί θεωρούν το διαδίκτυο και το WWW ως το σημαντικότερο μέσο ταχείας διάδοσης πληροφοριών και ως το παράθυρο στον κυβερνοχώρο. Όμως ο μεγάλος όγκος των πληροφοριών έχει δημιουργήσει ένα σημαντικό ερευνητικό πρόβλημα.

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 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.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. 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.

2. ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

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, και 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. Μερικοί αναλυτές υποστηρίζουν ότι, ο ιστός διπλασιάζεται σε μέγεθος κάθε 100 έως 125 ημέρες (Morgan, 1996).

Η 10η Σεπτεμβρίου του **1990** έμελλε να μείνει γνωστή ως η ημερομηνία που θα εισήγαγε την έννοια των μηχανών αναζήτησης στο INTERNET. Ο Peter Deutsch, μαζί με τους Alan Emtage και Bill Heelan,όλοι τους φοιτητές του πανεπιστημίου Mcgiil στον Καναδά ,ανήγγειλαν στο Usenet τη λειτουργία του Archie,καλώντας τους χρήστες του δικτύου να το χρησιμοποιήσουν .Το Archie (συντομογραφία του Archiver,αρχειοθέτης)ήταν ένα σύστημα καταγραφής ,σε καθημερινή βάση ,των περισσοτέρων διακομιστών FTP που λειτουργούσαν ,καθώς και των αρχείων που αυτοί περιλάμβαναν .Διατηρώντας μια λίστα με διακομιστές FTP (αρχικά 210),τους επισκεπτόταν τις νυχτερινές ώρες ,ώστε να μην προκαλεί κίνηση και αργοπορία στη δικτυακή κυκλοφορία ,και κατέγραφε τα αρχεία που αυτοί διέθεταν. Οποιοσδήποτε χρήστης , από κάθε γωνιά της γης , μπορούσε να συνδεθεί στο Archie και να ρωτήσει για τη διαθεσιμότητα ενός αρχείου.

Το **1991** έκανε την εμφάνιση του ένα νέο πρωτόκολλο, το Gopher το οποίο χρησίμευε για την κατηγοριοποίηση και την παρουσίαση των εγγράφων ενός διακομιστή .Πολλοί δικτυακοί τόποι ξεκίνησαν να κατασκευάζουν τις σελίδες τους ώστε να υποστηρίζουν το Gopher,οι πρώτοι browser έκαναν την εμφάνισή τους τότε ,ενώ το INTERNET άρχισε σιγά σιγά να παίρνει τη μορφή που έχει σήμερα. Το

1992στο πανεπιστήμιο της Νεβάδα, αναπτύχθηκε από τους Steven Foster και Fred Barrie η Veronica, μια μηχανή αναζήτησης που χρησιμοποιούσε το πρωτόκολλο Gopher.

Η ΠΡΩΤΗ «ΑΡΑΧΝΗ». Το **1993** ήταν η χρονιά που άνθησε και καθιερώθηκε ο Παγκόσμιος Ιστός (World Wide Web), ιδιαίτερα μετά την εμφάνιση του πρώτου φυλλομετρητή με το όνομα NCSA Mosaic. Ο Mosaic αναπτύχθηκε στο Εθνικό Κέντρο Εφαρμογών Υπερυπολογιστών, στο Πανεπιστήμιο του Ιλινόϊς. Τον Ιούνιο της ίδιας χρονιάς ο Matthew Gray, φοιτητής στο MIT, ορμώμενος από την επιθυμία του να ανακαλύπτει τους νέους δικτυακούς τόπους που ξεπηδούσαν με γοργούς ρυθμούς στο web, έφτιαξε ένα αυτόματο σύστημα που τους έψαχνε και τους εντόπιζε. Το δημιούργημά του το ονόμασε **wanderer** (περιπλανώμενος). Αν και το αρχικό κίνητρο του Matthew ήταν απλώς η ανεύρεση καινούριων σελίδων, λίγο αργότερα, καθώς το web είχε αρχίσει να μεγαλώνει με γεωμετρική πρόοδο, τον συνεπήρε η ιδέα της καταμέτρησης αυτής της ίδιας της ανάπτυξης. Προκειμένου να πετύχει το σκοπό του ο Matthew χρησιμοποίησε το wanderer και έγινε ο άνθρωπος που μας χάρισε τα πρώτα στατιστικά για την ανάπτυξη του διαδικτύου. Ενώ αρχικά το wanderer δεν αξιοποιούσε τα ευρήματά του με τρόπο ώστε να είναι δυνατές μετέπειτα αναζητήσεις, σιγά σιγά άρχισε να αποθηκεύει τις διευθύνσεις που έβρισκε στο δρόμο του και να χτίζει την πρώτη σχετική βάση δεδομένων, τη wandex. Η πρώτη αράχνη ήταν γεγονός και ανέπνευσε τον τρόπο λειτουργίας των σημερινών μηχανών αναζήτησης.

Καθώς το web μεγάλωνε με ταχυτάτους ρυθμούς, γινόταν όλο και πιο δύσκολη υπόθεση η καταχώρηση και η ταξινόμηση των νέων τόπων και σελίδων που καθημερινά εμφανίζονταν. Το wanderer αποτέλεσε την έμπνευση για πολλούς προγραμματιστές να ασχοληθούν σοβαρά με την αυτόματη χαρτογράφηση του διαδικτύου.Το σκεφτικό ήταν: εφόσον οι περισσότερες σελίδες διαθέτουν συνδέσμους που οδηγούν σε άλλες σελίδες η τόπους, ένα πρόγραμμα θα μπορούσε να τους ακολουθεί και από σύνδεσμο σε σύνδεσμο να ανακαλύπτει όλο και περισσότερα από το Web. Έτσι, τον Δεκέμβριο του 1993, τρεις μηχανές αναζήτησης, εξοπλισμένες με αράχνες που σάρωναν τον παγκόσμιο ιστό έκαναν την εμφάνισή τους. Επρόκειτο για τις jumbstation WWWWorm και RBSE [Repository-Based-Software Engineering]. Οι αράχνες τους ήταν ικανές να συλλέγουν πληροφορίες από τους τίτλους και τις κεφαλίδες των σελίδων που επισκέπτονταν.

Το 1995 έκανε την εμφάνισή του το Excite, δημιούργημα έξι φοιτητών από το πανεπιστήμιο του Στάνφορντ. Ήταν η πρώτη μηχανή που βασίστηκε σε στατιστικές αναλύσεις σχετικές με τη συγγένεια των λέξεων ,ώστε να κάνει πιο αποδοτικές τις αναζητήσεις των χρηστών του. Το 1994 ήταν η εποχή του Web Crawler, πνευματικού παιδιού του Brian Pinkerton, φοιτητή στο πανεπιστήμιο Ουάσινγκτον της ομώνυμης πολιτείας. Το Web Crawler ήταν η πρώτη μηχανή που κατέγραφε ολόκληρο το περιεχόμενο των σελίδων που επισκεπτόταν , τη στιγμή που οι ανταγωνιστές του κρατούσαν συνήθως μόνο τις πρώτες εκατό λέξεις. Επόμενους σταθμούς εξέλιξης αποτέλεσαν οι μηχανές Lykos και Altavista ,με την πρώτη να καλύπτει τον εντυπωσιακά μεγάλο αριθμό σελίδων της εποχής (εξήντα εκατομμύρια , 1996).Η Altavista έμεινε γνωστή για της αμείωτες επιδόσεις της ,παρά τα εκατομμύρια των επισκέψεών που δεχόταν καθημερινά . Επίσης το 1996,το Meta Crawler ήρθε να εισάγει την έννοια των μετα-μηχανών:μηχανές σε ρόλο διαμεσολαβητή ,οι οποίες μεταβιβάζουν τα ερωτήματα μας σε πλήθος <<πραγματικών>> μηχανών αναζήτησης και μας επιστρέφουν τα συγκεντρωτικά αποτελέσματα.

Το 1998 δύο φοιτητές από το Στάνφορντ, οι Larry Page και Sergey Brin,έφεραν τα πάνω κάτω στο χώρο εφαρμόζοντας ένα προηγμένο σύστημα αξιολόγησης των δικτυακών τόπων. Η μηχανή αναζήτησης που ανέπτυξαν είχε το παράξενο όνομα Google και έμελλε να αλλάξει για τα καλά τη δικτυακή μας ζωή.

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, μμορφωτικό 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.

3. ΟΙ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ ΓΕΝΙΚΑ

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 μία προσπάθεια να καθιερωθούν σε μία 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 μία 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 καμιά 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 κινουμένης 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 παγκοσμίου ιστού.

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 μμεγαλύτερη 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 διαθεσίμων 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 ότι 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 ανταπεξέλθει, έστω και μερικώς, στα προβλήματα που προκαλεί η αλόγιστη συσσώρευση πληροφορίας.

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, της τηλεόρασης, του ραδιόφωνου 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.

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.

Οι δυνατότητες που παρέχει το διαδίκτυο είναι απεριόριστες και οι πληροφορίες για οτιδήποτε χρειαζόμαστε σίγουρα υπάρχουν μέσα σ' αυτό. 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 πως εξελίσσετε ένας ποδοσφαιρικός αγώνας 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 της Αλίκης Βουγιουκλάκη 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.

Στο διαδίκτυο σίγουρα βρίσκονται αρκετοί άνθρωποι με τους οποίους έχουμε κοινά ενδιαφέροντα, προβληματισμούς, χόμπι κ.λπ., ασχολούνται με το αντικείμενο που μας ενδιαφέρει και κατέχουν πληροφορίες που θα θέλαμε να γνωρίζουμε. Το διαδίκτυο δεν το ελέγχει κανείς και 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.

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 συνδέοντας 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 μμεγαλώσει, 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.

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

• Το ειδικό λογισμικό (robot, spider, crawler κ.λπ.)

• 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)

Το Ειδικό Λογισμικό

Το ειδικό λογισμικό (robot, spider, crawler κ.λπ.) 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)) εξερευνώνται λεπτομερώς και συχνότερα από τους crawlers από ότι οι λιγότερο δημοφιλείς σελίδες.

Το ειδικό αυτό λογισμικό μπορεί να προγραμματιστεί να εξερευνάει τις ιστοσελίδες σε βάθος (depth) ή σε εύρος (breadth), ή και τα δύο. Εκείνο που προγραμματίζεται για σε βάθος εξερεύνηση όχι μόνο προσδιορίζει τις κυρίες 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 καθεμία 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 στον 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 ποιες 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 πολυτίμων πληροφοριών ευρετηρίασης.

Οι γνώστες της HTML γνωρίζουν ότι τα πλαίσια (frames) χρησιμοποιούνται σε εκατομμύρια ιστοσελίδες. (Τα πλαίσια είναι μια συσκευή της HTML που μμεταχειρίζεται 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 ανακτωμένης 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.

4. ΤΡΟΠΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

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.λπ.

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.

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 μμεγάλος 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.λπ., 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, **έβδομων** , 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.

5. ΤΡΟΠΟΙ ΑΝΑΖΗΤΗΣΗΣ ΚΑΙ ΕΡΓΑΛΕΙΑ

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 προηγουμένη 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 ότι 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).
• 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.

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 μμετατραπεί στην αναζήτηση με τη χρήση θεματικών ευρετηρίων.

Αναζήτηση με Χρήση Θεματικών Ευρετηρίων (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.

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).

Αναζήτηση με βάση το Προφίλ του Χρήστη(User's Profiille Searchiing)

User profiling μπορεί να οριστεί ως η προσπάθεια να δημιουργηθεί ένα προφίλ (profile) του χρήστη με βάση τα ενδιαφέροντα 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.

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 καταλόγου 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.

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 (επόμενη ενότητα)

*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.**

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 :

(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).

• 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 δυικής 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.

(Informatiion Retriievall 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.

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) συμφωνά 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.

6. ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Σε εκατοντάδες εκατομμύρια ανέρχεται σήμερα ο αριθμός των σελίδων στο Διαδίκτυο, ενώ χιλιάδες είναι οι νέες σελίδες που προστίθενται καθημερινά. Για έναν απλό χρήστη που επιθυμεί μόνος του, χωρίς τη χρήση κάποιου κατάλληλου βοηθητικού, λογισμικού εργαλείου, να ανακτήσει από το Διαδίκτυο πληροφορία σχετικά με ένα θέμα που τον ενδιαφέρει, οι προοπτικές να στεφθεί με επιτυχία η προσπάθειά του αυτή είναι σχεδόν ελάχιστες : ο χρήστης θα περιπλανιέται επί ώρες στο Ίντερνετ άσκοπα, ενώ πολλοί σχετικοί δικτυακοί τόποι θα παραμένουν για αυτόν ανεξερεύνητοι.

Από την άλλη, η αναζήτησή του θα διευκολυνόταν αφάνταστα με τη χρήση ενός λογισμικού εργαλείου που θα πραγματοποιούσε για λογαριασμό του την αναζήτηση σε όλο το Διαδίκτυο και θα του επέστρεφε όλα τα αποτελέσματα της αναζήτησης, με μόνο απαιτούμενο από την πλευρά του χρήστη να πληκτρολογήσει το θέμα που τον ενδιαφέρει με όσο το δυνατό μεγαλύτερη σαφήνεια. Τέτοια εργαλεία είναι 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.

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 μμεγαλύτερου 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.

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 συναντιούνται 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 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.

Η μηχανή αναζήτησης μπορεί να θεωρηθεί ότι αποτελείται γενικά από πέντε κύρια εσωτερικά στοιχεία (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)

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 μμεγαλύτερες 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 περιεχόμενων 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 περιεχόμενα 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 περιεχόμενα 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. Συμφωνά 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 καμιά 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 -συμφωνά 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 αναζητήσημους 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 μμορφοποιήσει 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 προσμετράτε 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.

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.

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.

Ο Crawler των πληροφοριών ανιχνεύει τον ιστό και αποθηκεύει τις πληροφορίες σε μια βάση δεδομένων, η οποία είναι πρόσβαση 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 και τα αποτελέσματα εμφανίζονται σε μορφή ενιαίας λίστας. Συλλέγοντας αποτελέσματα από διαφορετικές μηχανές αναζήτησης, η ανάκληση των εγγράφων καθώς και η ακρίβειά τους αυξάνεται.

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 τακτοποιούν κάθε αντικείμενο (βιβλίο, περιοδικό, κ.λπ.). 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 διαθεσίμων 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 καμιάς μηχανής αναζήτησης. Οι Lawrence S. και Giles C. 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 καμιά 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.λπ.). 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. .ς 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).

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.

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 .

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. Καμιά 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 συναντάτε 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, συμφωνά 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 συμφωνά 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 αναπαρίστανται 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 αντιστραμμένου ευρετηρίου αρχείων (inverted file index). Δηλαδή, κάθε πιθανός όρος έχει μία καταχώρηση σε ένα αρχείο ευρετηρίου μαζί με μία λίστα συνδεόμενων αυξόντων αριθμών των εγγράφων. Αυτοί οι αύξοντες αριθμοί μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτίωση περαιτέρω μεταδεδομένων (metadata) και συχνά για ένα τοπικό αντίγραφο του πλήρους κειμένου (full text) του εγγράφου.

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.

Set Theoretic Models

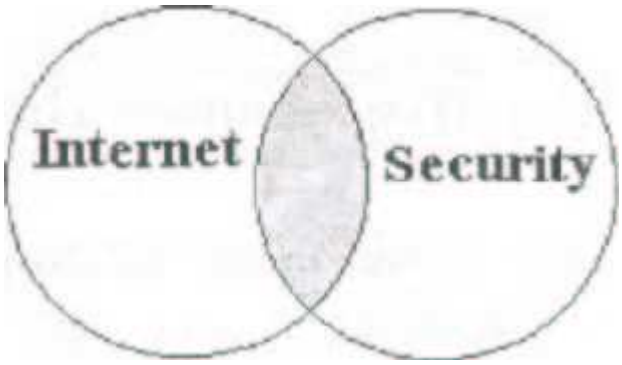
BOOLEAN LOGIC

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:

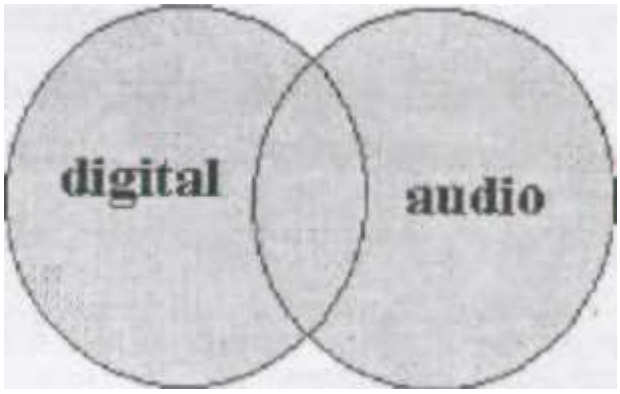


Το παραπάνω διάγραμμα απεικονίζει τη χρήση του τελεστή AND. Στον αριστερό κύκλο περιλαμβάνονται όλα τα έγγραφα που περιέχουν τον όρο Internet. Ο δεξιός κύκλος περιλαμβάνει όλα τα έγγραφα που περιέχουν τον όρο security. Το αποτέλεσμα θα είναι τελικά η μηχανή αναζήτησης να επιστρέψει τα έγγραφα εκείνα 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.

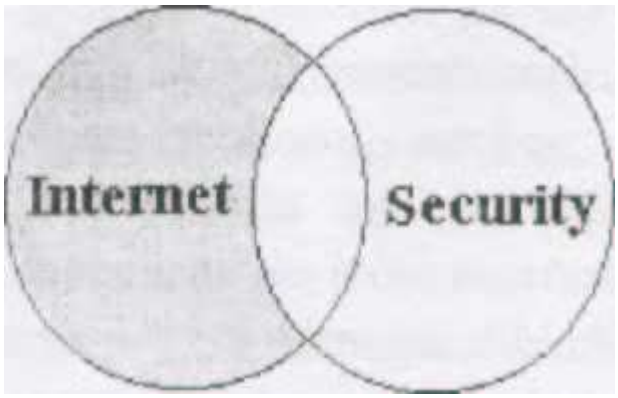


Όπως φαίνεται στο παραπάνω διάγραμμα γίνεται αναζήτηση για όλα τα έγγραφα που περιέχουν τον όρο digital (αριστερός κύκλος) και για εκείνα που περιέχουν τον όρο audio (δεξιός κύκλος). Τα έγγραφα που θα ανακτηθούν θα είναι όλα εκείνα που περιέχονται στους δύο κύκλους.

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 τον τελεστή :



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 αναβρεθεί στις δύο υπό-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 λεγομένη 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. Αναζητώντας 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.λπ. θα αποκοπεί.

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

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={d1,d2, ...dn} όπου di με (1<i<η) 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={q1,q2, ...qm}, όπου qi 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 μμεγαλύτερη αξιοπιστία να ανακτήσουν και να κατατάξουν τα καταλληλότερα κάθε φορά έγγραφα.

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.

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.λπ.

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 ερώτημα 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.

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.

Ειδικό Λογισμικό (Robot, Spiders ,Crawler, Agent)

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 P0. Ανακτά αυτή τη σελίδα P0, 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 αλγόριθμους 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 περιεχόμενο 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.

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 ομοιότητα της με το κείμενο της ερώτησης του χρήστη.

Λόγω του πολλαπλασιασμού των σελίδων του www, πολλά νεότερα 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 περιεχόμενο 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 δημοτικότητα 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 πλατφορμών 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.λπ.). 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.

Αλγόριθμοι Spidering

Ο ιστός έχει πληθώρα χρησίμων 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 ερώτημα 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 περιεχόμενο 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.

Μερικές τεχνικές θεωρούν URLs με λιγότερες καθέτους πιο χρήσιμες από εκείνες με περισσότερες καθέτους.

10λογος των Μηχανών Αναζήτησης

Οι μηχανές αναζήτησης πρέπει να θεωρούνται ως η κάλλιστη λύση για διαφήμιση και προώθηση μιας Ιστοσελίδας . Μόλις γίνει σωστή εκχώρηση, μπορούν να στείλουν ένα ικανοποιητικό ποσό από επισκέπτες χωρίς κόπο. Αλλά η πληρωμή, φέρνει 2x,3x μέχρι και 10x καλύτερα αποτελέσματα.

Υπάρχουν εκατομμύρια Ιστοσελίδες, και κάθε μέρα προστίθενται περισσότερες. Όμως πολλές από αυτές δεν κάνουν σωστή χρήση των meta-tags. Τα meta-tags χρησιμοποιούνται από τα bots των μηχανών αναζήτησης για indexing της Ιστοσελίδας .

Ο ΔΕΚΑΛΟΓΟΣ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ :

1. Χρησιμοποιείτε τις πιο κοινές λέξεις-κλειδιά στον τίτλο. Οι περισσότερες μηχανές αναζήτησης το "μετρούν" ως καλύτερο και σας ανεβάζουν στο ranking.

2. Χρησιμοποιείτε τις λέξεις-κλειδιά στον πληθυντικό αριθμό (π.χ. "βιβλία", και όχι "βιβλίο") Έτσι, είτε κάποιος ψάχνει για "βιβλίο" ή "βιβλία", η Ιστοσελίδα σας θα εμφανιστεί και στις 2 περιπτώσεις.

3. Χρησιμοποιείτε κεφαλαία και πεζά γράμματα στις λέξεις-κλειδιά. (παράδειγμα, "βιβλία, Βιβλία, ΒΙΒΛΙΑ") Περισσότερα από 3 είδη γραφής είναι άχρηστα. Ωστόσο συχνά ορθογραφικά λάθη βοηθούν (π.χ. "αυγά, αβγά").

4. Ομαδοποιήστε τις λέξεις-κλειδιά. Οι χρήστες συνήθως ψάχνουν χρησιμοποιώντας εκφράσεις (π.χ. "πώληση βιβλίων").

5. Πρέπει να χρησιμοποιείτε meta-tags. Πολλές μηχανές αναζήτησης κάνουν

"indexing" μέσω των meta-tags. Αυτά τα τοποθετείτε ανάμεσα στο <head></head> στο HTML σας και δεν θα φαίνονται στην σελίδα σας.

6. Χρησιμοποιείτε συνδυασμούς από τις 10-15 καλύτερες λέξεις-κλειδιά. Ο εμπλουτισμός με λέξεις-κλειδιά στα κείμενα σας είναι ένα κόλπο για καλύτερη απόδοση.

7. Να εισάγετε στα alt tags των εικόνων σας λέξεις-κλειδιά.

8. Κάθε σελίδα σας θα πρέπει να καταχωρηθεί σε μηχανές αναζήτησης, όχι μόνο η αρχική της αλλά όλες ανεξαιρέτως οι σελίδες .

9. Προσοχή! Το keyword spamming είναι μια τεχνική που μερικοί χρησιμοποιούν για να κερδίζουν καλύτερο tracking. Αυτό επιτυγχάνετε βάζοντας keywords παντού, ξανά και ξανά. Μετά αλλάζουν το χρώμα του κειμένου και θέτουν ως χρώμα αυτό του background. Αν σας πιάσουν, πολλές μηχανές αναζήτησης θα σας απομακρύνουν από τους καταλόγους σας και θα κάνουν ban την Ιστοσελίδα σας.

10. Να ανανεώνετε συνεχώς. Θα πρέπει να ανανεώνετε συνεχώς τα περιεχόμενα της Ιστοσελίδας σας, έτσι ώστε οι μηχανές αναζήτησης να κάνουν συχνά re-indexing και να ανεβαίνει το ranking στους καταλόγους τους.

Παρακάτω παραθέτω μερικούς αλγόριθμους που χρησιμοποιεί 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;
```

Πλεονεκτήματα και Μειονεκτήματα

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 (webbased) 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.

7. ΜΗΧΑΝΕΣ ΠΟΛΛΑΠΛΗΣ ΑΝΑΖΗΤΗΣΗΣ

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.

<u>Βασικά Χαρακτηριστικά:</u>

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 ερευνώνται 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.

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

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 πλεονέκτημα 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.λπ.) της κάθε μηχανής ξεχωριστά.

8. ΟΙ ΠΙΟ ΓΝΩΣΤΕΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Ορισμένες από τις πιο δημοφιλείς Μηχανές Αναζήτησης είναι οι ακόλουθες:

Alta Vista

Μια από τις πιο περιεκτικές διαθέσιμες Μηχανές Αναζήτησης στο Δίκτυο. Επιτρέπει απλή ή εξειδικευμένη αναζήτηση σε HTML αρχεία καθώς και τη χρήση Boolean Operators (or, and, not). Η υπηρεσία "images" επιτρέπει την αναζήτηση σε μια μεγάλη βάση δεδομένων που περιλαμβάνει φωτογραφίες. Η υπηρεσία "translate" επιτρέπει αυτόματη μετάφραση των websites από μια γλώσσα σε μια άλλη.

NorthernLight

Μια από τις πιο περιεκτικές διαθέσιμες Μηχανές Αναζήτησης ίσου μεγέθους περίπου με την Alta Vista. Επιτρέπει απλή και εξειδικευμένη αναζήτηση με τη χρήση Boolean Operators και παραθέτει ως επί το πλείστον ποιοτικά αποτελέσματα. Ταξινομεί τα αποτελέσματα σε Folders, δυνατότητα που διευκολύνει το χρήστη στην έρευνα που διεξάγει.

Google

Τα αποτελέσματατα από τη Google περιλαμβάνουν όλους τους όρους αναζήτησης και ταξινομούνται ανάλογα με την εγγύτητα μεταξύ των όρων αναζήτησης. Η εντολή "I'm feeling lucky" μεταφέρει το χρήστη απευθείας στο website που βρίσκεται στη πρώτη θέση ανάλογα με την ταξινόμηση της Google. Κρατάει επίσης αρχεία ασφαλείας (back up) για τυχόν νεκρά links.

Excite

Επιτρέπει απλή ή εξειδικευμένη αναζήτηση και τη χρήση των Boolean Operators (or, and, not). Μετά την ολοκλήρωση της αναζήτησης προτείνει λέξεις για τον περιορισμό της έρευνας. Παρέχει πληροφορίες για σπορ, δελτία καιρού, δυνατότητα ανεύρεσης ατόμων, chat rooms, ωροσκόπια, ταξίδια και χάρτες, παιχνίδια, αυτοκίνητα, κλπ.

HotBot

Ευρετηριάζει περίπου110 εκατομμύρια sites (τοποθεσίες στο Δίκτυο).Επιτρέπει απλή ή εξειδικευμένη αναζήτηση σε θεματικούς καταλόγους, γκρουπ συζητήσεων, αναζήτηση επιχειρήσεων, ανεύρεση ατόμων και email διευθύνσεων, μικρές αγγελίες, πληροφορίες χρηματιστηρίου κλπ. Παρέχει ακόμη τη δυνατότητα έρευνας σε συγκεκριμένους τομείς (com ή org).

Go.com

Ευρετηριάζει 60 περίπου εκατομμύρια sites και προσφέρει ταχύτητα και ικανοποιητικά ποσοστά ακρίβειας στα αποτελέσματα της αναζήτησης. Επιτρέπει

έρευνα σε ξένη γλώσσα, ανάλογα με γεωγραφική τοποθεσία, καθώς και τη χρήση Boolean Operators.

Lycos

Από τις παλαιότερες Μηχανές Αναζήτησης , παρέχοντας όμως υψηλά σε σχετικότητα αποτελέσματα, με δυνατότητα για απλή ή εξειδικευμένη αναζήτηση, χρήση Boolean Operators και πολλαπλά links σε εικόνα και ήχο.

WebCrawler

Ευρετηριάζει 2 εκατομμύρια sites και είναι εύκολο στη χρήση. Η φόρμα των αποτελεσμάτων της έρευνας δίνει αρχικά 25 τίτλους, καθώς και τη δυνατότητα εμφάνισης περιλήψεων των κειμένων και το ποσοστό επιτυχίας της ίδιας της έρευνας.

Ελληνικές Μηχανές Αναζήτησης

In.gr

Ευρετηριάζει Ελληνικές τοποθεσίες ενώ ταυτόχρονα δίνει τη δυνατότητα παροχής προσωπικού λογαριασμού ηλεκτρονικού ταχυδρομείου, σύνδεση με το χρηματιστήριο, δελτίο καιρού για την Ελλάδα και τον υπόλοιπο κόσμο, καθώς και τη συμμετοχή σε forum συζητήσεων.

The Greek Indexer

Πρόκειται για Μηχανή Αναζήτησης που ευρετηριάζει Ελληνικές τοποθεσίες στο Διαδίκτυο.

FindLink

Το FindLink ευρετηριάζει τοποθεσίες αποκλειστικά με Ελληνικό περιεχόμενο. Παρέχει ακόμη σύνδεση με το Χρηματιστήριο, δελτία καιρού, κ. λ .π.

Phantis

Ελληνική Μηχανή Αναζήτησης που ευρετηριάζει 16.000 περίπου sites. Παρέχει θεματικό κατάλογο Ελληνικών νομών και πόλεων, πρόσβαση σε ειδήσεις στην Ελλάδα αλλά και στον κόσμο.

Meta Search Engines

Metacrawler

Παραθέτει τα αποτελέσματα της έρευνας σε μια μεγάλη λίστα, ταξινομημένα κατά φθίνουσα σειρά ανάλογα με την ακρίβεια που τα χαρακτηρίζει. Τα αξιολογεί βαθμολογώντας στην κλίμακα 1-1000 (τα ακριβέστερα ανάλογα με την αναζήτηση).

Dogpile

Στέλνει την έρευνα σε ένα αρκετά μεγάλο αριθμό από search engines και παραθέτει τα αποτελέσματα της έρευνας κατά συγκεκριμένη μηχανή αναζήτησης.

Metafind

Παραθέτει τα αποτελέσματα της έρευνας ταξινομημένα κατά μηχανή αναζήτησης, ή αλφαβητικά, ή κατά keyword.

Θεματικοί Οδηγοί (Subject Directories)

Yahoo

Πρόκειται για τον πιο περιεκτικό και τον πιο δημοφιλή θεματικό οδηγό του Internet. Καθώς το yahoo στηρίζει τις πληροφορίες που περιλαμβάνει σε υποδείξεις και υποβολές προτάσεων των ίδιων των χρηστών, ίσως δεν είναι τόσο περιεκτικό όσο μια search engine. Είναι όμως εύκολο στη χρήση και πολύ δημοφιλές στην κοινότητα των χρηστών του Διαδικτύου λόγω της πληθώρας των υπηρεσιών που προσφέρει.

The www Virtual Library

Πρόκειται για ένα θεματικό οδηγό του Internet που ευρετηριάζει sites ποιότητας. Παρέχει τη δυνατότητα έρευνας με θεματικές κατηγορίες αλλά και με λέξεις-κλειδιά.

Magellan

Παρέχει τη δυνατότητα έρευνας με λέξεις κλειδιά αλλά παράλληλα κάνει λογικούς συσχετισμούς για ακριβέστερα αποτελέσματα και ευκολία στη χρήση. Παραθέτει ανά 10 τα πιο σχετικά με την αναζήτηση αποτελέσματα.

Άλλες χρήσιμες διευθύνσεις

Hytelnet

Μέσω του Hytelnet είναι δυνατή η πρόσβαση σε on line καταλόγους βιβλιοθηκών (OPACs) που είναι διαθέσιμοι στο Internet.

Κατάλογος Ελληνικών OPAC

Πρόκειται για ένα κατάλογο των Ελληνικών Βιβλιοθηκών και των online καταλόγων τους (OPACs).

Greek explorer

Πρόκειται για μια μηχανή αναζήτησης που εντοπίζει πληροφορίες από Ελληνικά sites.

Tilenet

Πρόκειται για μηχανή αναζήτησης mailing lists και newsgroups.

9. ΓΛΩΣΣΑΡΙ

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 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 αποφασίζετε, 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 περιεχόμενα τους (συνδεόμενες σελίδες και έγγραφα, συνδεόμενα sites, κ.λπ.)

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 ευρύτερα είναι γνωστοί, αποτελεί την εναλλακτική λύση έναντι των mηχανών 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 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 περιεχόμενα τους (συνδεόμενες σελίδες και έγγραφα, συνδεόμενα sites, κ.λπ.).

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 περιεχόμενα τους (συνδεόμενες σελίδες και έγγραφα, συνδεόμενα sites, κ.λπ.).

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.

10. ΒΙΒΛΙΟΓΡΑΦΙΑ

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

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.

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).

Cheung D, W., Kao B., Lee J. (1998), Discovering user access patterns on the World Wide Web, Knowledge Based Systems.

Christophi C., Dikaiakos M. (2003), Automatic profile generation in eRACE, University of Cyprus.

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 .

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.

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Αναγνωστόπουλος, Ι. (2004),Νέες Μέθοδοι Για Ευφυή Ανάκτηση , Κατηγοριοποίηση Και Ταξινόμηση Δεδομένων Σε Πληροφορικά Διαδικτυακά Συστήματα, Εθνικό Μετσόβιο Πολυτεχνείο

Γεωργάκης, Κ. (2004) , Μελέτη Των Μηχανών Αναζήτησης Στο Διαδίκτυο Καθώς Και Των Τεχνικών Τους –Ανάπτυξη Ενός Μοντέλου –Προτύπου Ενιαίας Αναζήτησης , Πολυτεχνείο Κρήτης.

Καρασαββίδης , Σ. (2004), Προς Μια Μεθοδολογία Καταγραφής Και Ανάλυσης Επισκεψιμότητας Ιστοχώρων Σε Εφαρμογές Διαδικτύου, Πολυτεχνείο Κρήτης.

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.

INTERNET SITES

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/tsld00
1.
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/metasearc
h_
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

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ευκαιρία της εκπόνησης της πτυχιακής εργασίας μου θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα Καθηγητή μου κ. Αντωνιδάκη Μανώλη 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 πτυχιακή εργασία με τον καλύτερο δυνατό τρόπο.

Επίσης οφείλω ένα πολύ μεγάλο και θερμό ευχαριστώ στους γονείς μου για την αμέριστη συμπαράσταση και υποστήριξή τους τόσο ψυχολογική όσο και υλική.

Θα ήθελα τέλος να ευχαριστήσω όλους τους φίλους-συμφοιτητές μου που με βοήθησαν να ξεπεράσω τις δύσκολες καταστάσεις καθ' όλη τη διάρκεια των σπουδών μου.