Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδος

Σχολή Διοίκησης και Οικονομίας

Τμήμα Λογιστικής

Θέμα:

DEEP WEB VS SURFACE WEB.ΑΝΑΖΗΤΗΣΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟ ΔΙΑΔΙΚΤΥΟ

Φοιτητές: ΑΝΤΩΝΙΟΥ ΡΑΦΑΗΛ (ΑΜ 10715)
ΝΤΑΝΤΗΣ ΑΔΡΙΑΝΟΣ-ΜΙΧΑΗΛ (ΑΜ 10952)
ΛΙΑΛΙΑΤΣΗΣ ΧΡΗΣΤΟΣ (ΑΜ 10747)

Επιβλέπων καθηγητές: Φωτεινόπουλος Μιχάλης

Κωνσταντίνος Παξιμάδης

2018

Περιεχόμενα

ΠΕΡΙΛΗΨΗ 3
ABSTRACT 4
ΕΙΣΑΓΩΓΗ 5
ΚΕΦΑΛΑΙΟ 1ο: Προσδιοριστικά στοιχεία 6
1.1 Ορισμοί παγκόσμιο ιστός, μηχανές αναζήτησης. 6
1.1.1 Παγκόσμιος ιστός 6
1.1.1Μηχανές αναζήτησης 7
1.2Ιστορικήαναδρομή 9
1.3Μηχανές αναζήτησης 10
1.4Μηχανές Μετά-Αναζήτησης 25
1.5 Οφέλη από την χρήση μηχανών αναζήτησης 28
ΚΕΦΑΛΑΙΟ 2ο: Αναζήτηση, πάταξη και αποθήκευση δεδομένων 30
2.1 Αναζήτηση δεδομένων 30
2.2 Αποθήκευση δεδομένων 32
2.2.1 Ανεστραμμένοι δείκτες 35
2.3 Προοπτικό Ευρετήριο 39
2.4 Κατάταξη αποτελεσμάτων (Ranking) 40
2.4.1 Page Rank 41
2.4.2 Hyperlink-Induced Topic Search (HITS) 43
2.5 Τα είδη των μηχανών αναζήτησης 46
2.5.1 Βασικές λειτουργίες των μηχανών αναζήτησης 48
2.6 ΛεξικήΑλυσίδα 48
2.7 Αυτόματη Ανάθεση Κειμένων στη Θεματική Ιεραρχία 51
ΚΕΦΑΛΑΙΟ 3ο: Εξέλιξη διαδικτύου και αναζήτησης 57
3.1 Σημασιολογικές μηχανές αναζήτησης 57
3.1.1 Ακαδημαϊκές Βιβλιοθήκες και σημασιολογική αναζήτηση 57
3.1.2 Σημασιολογικά Μοντέλα 59
3.1.3 Σημασιολογικές πύλες 63
3.2 Τεχνολογίες web 2,deepweb 65
3.2.1 web 2 65
3.2.2 Deepweb 76

ΣΥΜΠΕΡΑΣΜΑΤΑ 81
ΒΙΒΛΙΟΓΡΑΦΙΑ 84

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αποτελεί μια προσπάθεια, περιγραφής, ανάλυσης και προσδιορισμού της σημαντικότητας του διαδικτύου μέσα από την παρουσίαση του σημαντικού ρόλου που διαδραματίζουν οι μηχανής αναζήτησης στην ευκολία αναζήτησης πληροφοριών για τους χρήστες. Υπάρχουν πολλές μηχανές αναζήτησης που η κάθε μια διακρίνεται από πλεονεκτήματα και μειονεκτήματα, παρόλα αυτά υπάρχουν παρά πολλές ώστε ο χρήστης να επιλέξει αυτήν που συμβάλλει αποτελεσματικά στις εργασίες που θέλει να επιτελέσει στο διαδίκτυο.

ABSTRACT

The present work is an effort, description, analysis and determination of the importance of the Internet through the presentation of the important role that the search engine plays in the ease of searching for information for the users. There are many search engines that each is distinguished by pros and cons, yet there are but many so the good to choose the one that contributes effectively to the tasks he wants to perform on the internet.

ΕΙΣΑΓΩΓΗ

Η παρούσα εργασία αποτελεί μια προσπάθεια προσέγγισης του ζητήματος που αφορά την αναζήτηση και διαχείριση πληροφορίας στο διαδίκτυο. Η προσπάθεια αυτή πραγματοποιήθηκε στα πλαίσια ολοκλήρωσης των σπουδών μας στο τμήμα λογιστικής στο Τεχνολογικό Ίδρυμα Δυτικής Ελλάδος. Η εργασία αποτελείται από τρείς ενότητες.

Στην πρώτη ενότητα αρχικά δίνεται ένας ορισμός του διαδικτύου και τι περιλαμβάνει ο παγκόσμιος ιστός, στην συνέχεια παρουσιάζεται μια ιστορική αναδρομή διαδικτύου και ακολούθως αναλύονται και προσδιορίζονται οι μηχανές αναζήτησης που υπάρχουν σε συνδυασμό με τις μηχανές μετά-αναζήτησης. Στην συνέχεια κρίθηκε σκόπιμο να προσδιοριστούν και να αναφερθούν τα οφέλη που προκύπτουν από την χρήση μηχανών αναζήτησης και η ενότητα ολοκληρώνεται με την παρουσίαση της ανατομίας μιας μηχανής αναζήτησης.

Στην δεύτερη ενότητα γίνεται αναφορά στις μορφές αναζήτησης και ειδικότερα στην κατηγοριοποίηση κειμένων, στον ειρμό και την συνάφεια, στους τύπους, στην λεξική συνάφεια, στην κυριαρχία της λεξικής συνάφειας, στην λεξική αλυσίδα, στην αυτόματη ανάθεση κειμένων στη θεματική ιεραρχία και τέλος παρουσιάζονται τα πλεονεκτήματα των στοχευόμενων μηχανών αναζήτησης.

Και η εργασία ολοκληρώνεται με την παρουσίαση των εξελίξεων που σχετίζονται με το διαδίκτυο και την αναζήτηση. Μέσα από την παρουσίαση των σημασιολογικών μηχανών αναζήτησης, τον προσδιορισμό των τεχνολογιών που αφορούν toweb 2, deepweb. Στην συνέχεια αναλύεται ο συμμετοχικός ιστός και η εργασία ολοκληρώνεται με την αναφορά και παρουσίαση στα σημασιολογικά δεδομένα.

ΚΕΦΑΛΑΙΟ 1ο: Προσδιοριστικά στοιχεία.

1.1 Ορισμοί παγκόσμιο ιστός, μηχανές αναζήτησης.

1.1.1 Παγκόσμιος ιστός

Ο Παγκόσμιος Ιστός είναι το μεγαλύτερο σε μέγεθος δημιούργημα που σχετίζεται με πληροφορίες που έχει κατασκευαστεί και λειτουργεί την παρούσα χρονική στιγμή, από τον άνθρωπο. Περιγράφεται ως ένα σύστημα διασυνδεδεμένων εγγράφων στα οποία τα άτομα χρήστες έχουν την δυνατότητα να έχουν πρόσβαση με την συμβολή του Διαδικτύου. Με την βοήθεια μιας εφαρμογής Web browser, ο χρήστης διαθέτει την δυνατότητα πρόσβασης σε ιστοσελίδες οι οποίες μπορεί να περιλαμβάνουν πληροφορίες σε μορφή κειμένου, εικόνων και βίντεο και μέσα από μια ακολουθία υπερσυνδέσμων που υπάρχουν στις ιστοσελίδες και να περιπλανείται σε αυτές.

Το 1989 δημιουργήθηκε ο Παγκόσμιος Ιστός από τον Sir Tim Berners-Lee στον Ευρωπαϊκό Οργανισμό Πυρηνικής Έρευνας και η χρήση του άρχισε το 1992. Από τότε, ο Berners-Lee έπαιξε ενεργό ρόλο στην καθοδήγηση της ανάπτυξης τεχνολογιών και προτύπων για τον Παγκόσμιο Ιστό. Ο Παγκόσμιος Ιστός ακολουθεί την αποκεντρωμένη δομή του Διαδικτύου και αυτό συντελεί ώστε τα δεδομένα να υπάρχουν σε πολλά και διαφορετικά σημεία, σε πολλαπλά αρχεία και διακομιστές. Αυτή είναι η δομή του Διαδικτύου, που σχεδιάστηκε έτσι ώστε να μπορεί να επιβιώσει μετά από έναν πυρηνικό πόλεμο, αλλά και να δημιουργεί προβλήματα σε ειρηνικές περιόδους κατά την χρήση του. Το σημαντικότερο πρόβλημα που προκύπτει στην αναζήτηση και εύρεση πληροφοριών, είναι η αποκεντρωμένη δομή των πληροφοριών που αποτελούσε εμπόδιο στη γρήγορη, αναζήτηση και εύρεση δεδομένων. Τα εμπόδιο αυτό ξεπεράστηκε και δεν είναι κατανοητό από τους σημερινούς χρήστες. Σε αυτό συντέλεσε η δημιουργία και εξέλιξη των τεχνολογιών που αφορούν τις μηχανές αναζήτησης (http://news.netcraft.com/archives/2008/12/24/december_2008_web_server_survey.html).

1.1.1 Μηχανές αναζήτησης

Η λειτουργία των μηχανών αναζήτησης είναι να αντιγράφουν τα περιεχόμενα που υπάρχουν στον Παγκόσμιο Ιστό σε μια εντελώς διαφορετική δομή από αυτήν στην οποία βρίσκονται στο Διαδίκτυο. Οι μηχανές αναζήτησης αποθηκεύουν πληροφορίες αλλά και τα περιεχόμενα μέσα από έναν τεράστιο όγκο πληροφοριών από ιστοσελίδες σε μια κεντρική βάση δεδομένων. Όλες οι ιστοσελίδες ανακτώνται από τον Παγκόσμιο Ιστό με την συμβολή ενός λογισμικού που καλείται Web Crawler και είναι ουσιαστικά ένας αυτοματοποιημένος φυλλομετρητής του ιστού ο οποίος επισκέπτεται το σύνολο των συνδέσμων που βρίσκονται στο δρόμο της αναζήτησης του. Τα περιεχόμενα της κάθε ιστοσελίδας ακολούθως υπόκεινται σε ανάλυση για να διαπιστωθεί ο τρόπος με τον οποίο είναι αναγκαίο η σελίδα να δεικτοδοτηθεί, δηλαδή να αποθηκευτεί με βάση διακεκριμένες λέξεις-κλειδιά με τις οποίες βρίσκεται σε συσχέτιση. Τα σημαντικότερο τμήμα για τη συσχέτιση των λέξεων-κλειδιών και της θεματικής περιοχής μιας ιστοσελίδας παρέχεται στο κείμενο που διακρίνει τον τίτλο, τις επικεφαλίδες των ενοτήτων όπως επίσης και τα ειδικά πεδία που καλούνται meta-tags τα οποία προσδιορίζονται από τον κατασκευαστή της ιστοσελίδας, και αυτός είναι και ο λόγος που δεν είναι εμφανή στους χρήστες. Μηχανές αναζήτησης, όπως για παράδειγμα το Google, αποθηκεύουν το σύνολο ή μέρος μιας ιστοσελίδας καθώς και κάποιες επιπλέον πληροφορίες που σχετίζονται με αυτήν, ενώ άλλες, όπως η AltaVista αποθηκεύουν κάθε λέξη κάθε ιστοσελίδας ξεχωριστά. Αυτό το αποθηκευμένο αντίγραφο της ιστοσελίδας καλείται cache και γίνεται χρήση κατά την διαδικασία της αναζήτησης.

Στην περίπτωση που κάποιος χρήστης εισάγει ένα ερώτημα με τη μορφή λέξεων-κλειδιών σε μια μηχανή αναζήτησης, η μηχανή αναλύει τη βάση δεδομένων της και γυρίζει πίσω σε αυτόν μια λίστα από ιστοσελίδες που θεωρεί ότι έχουν μεγαλύτερη σχετική συνάφεια περισσότερο με τους όρους του ερωτήματος. Το κάθε αποτέλεσμα στις περισσότερες των περιπτώσεων περιλαμβάνει τον τίτλο της σελίδας και ένα μικρό απόσπασμα είτε από την αρχή της σελίδας είτε τμήμα του κειμένου που περιλαμβάνει τις λέξεις-κλειδιά για την οποία έχει αποστείλει αίτημα αναζήτησης ο χρήστης.

Σχεδόν το σύνολο των μηχανών αναζήτησης κάνουν χρήση λογικών τελεστών AND, OR και NOT κατά την διαδικασία δημιουργίας ενός ερωτήματος. Ακόμα κάποιες από αυτές τις μηχανές αναζήτησης βασίζονται και στο proximity search, μέσω της οποίας ο χρήστης έχει την δυνατότητα να προσδιορίσει την μέγιστη απόσταση που μπορούν να διαθέτουν οι όροι του ερωτήματος εντός του κείμενου. Το πόσο χρήσιμη είναι μια μηχανή αναζήτησης προσδιορίζεται από την σχετικότητα των αποτελεσμάτων που επιστρέφει αυτή στο χρήστη μετά το ερώτημα του. Παρόλο που υπάρχουν εκατομμύρια ιστοσελίδες που να διαθέτουν κάποιες συγκεκριμένες λέξεις-κλειδιά ή φράσεις, κάποιες σελίδες είναι δυνατόν να είναι περισσότερο σχετικές, δημοφιλείς ή έγκυρες συγκριτικά με κάποιες άλλες. Με το μεγαλύτερο μέρος των μηχανών αναζήτησης να κάνουν χρήση των μεθόδων κατάταξης αυτών των αποτελεσμάτων έτσι ώστε οι «καλύτερες» ιστοσελίδες να επιστρέφονται πρώτες.

Οι τρόποι με τους οποίους αναφέρονται τα καλύτερα αποτελέσματα και η σειρά εμφάνισής τους διαφέρουν σημαντικά στις διαφορετικές μηχανές αναζήτησης. Οι μέθοδοι αυτές αλλάζουν ακόμα και με το πέρασμα του χρόνου, καθώς η χρήση του Διαδικτύου μεταβάλλεται και δημιουργούνται νέες τεχνολογίες. Λόγω του ότι οι περισσότερες μηχανές αναζήτησης είναι υπό την διαχείριση ιδιωτικών εταιριών, οι αλγόριθμοι που κάνουν χρήση καθώς και οι βάσεις δεδομένων τους έχουν αποθηκευτεί ως εταιρικά κρυφά στοιχεία δεν δημοσιοποιούνται στους χρήστες. Γενικά, τα επίπεδα λειτουργίας μιας μηχανής αναζήτησης διακρίνονται στα ακόλουθα:

- Ανακάλυψη (crawling)
- Αποθήκευση (indexing)
- Αναζήτηση (querying)
- Κατάταξη (ranking)

(Official Google Blog, 2017).

1.2 Ιστορική αναδρομή

Το πλήθος των ανακαλύψεων που αφορούν το Διαδίκτυο συνέβαλε σημαντικά στην τεραστία ανάπτυξη του και στο να μην έχει καμία σχέση με αυτό που ήταν όταν πρώτο δημιουργήθηκε. Επιπλέον πέραν των τεχνολογικών βελτιώσεων και ανακαλύψεων που εφαρμόστηκαν σε αυτό δημιουργήθηκε και πληθώρα ιστοσελίδων και το διαδίκτυο δεν αποτελούσε πια μια απλή πηγή πληροφόρησης, αλλά συνδεόταν με σημαντικές μεγαλύτερες επιχειρηματικές δραστηριότητες. Στο πρώτο στάδιο λειτουργιάς του αποτελούνταν από έναν αριθμό από Ftp (File transfer protocol) sites που χρήστες είχαν την δυνατότητα να κατεβάσουν ή να ανεβάσουν αρχεία. Παρόλα αυτά η αναζήτηση και εύρεση των αρχείων αυτών διακρινόταν από σημαντική δυσκολία και αυτό προέκυπτε από την ανάγκη γνώσης της ακριβής διεύθυνσης που ήταν τοποθετημένα. Η αναζήτηση αυτή διακρινόταν ως διαδικασία αρκετά δύσκολη, χρονοβόρα και απαιτούσε αρκετή υπομονή. Όλα αυτά, πριν ο φοιτητής, Alan Emtage [1990], δημιουργήσει το πρώτο εργαλείο αναζήτησης (SearchEngineTool). Η ανακάλυψη του ήταν ένας κατάλογος από τα αρχεία που υπήρχαν στο Internet και ονομάστηκε Archie (Archive). Το 1991 ένας ακόμα φοιτητής ο Mark McCahill συνειδητοποίησε ότι αφού μπορείς να ψάχνεις τα αρχεία στο Διαδίκτυο, τότε είχε την δυνατότητα να αναζητεί και κείμενα σε αυτό. Έτσι δημιούργησε το Gopher, ένα πρόγραμμα που κατηγοριοποιούσε το απλό κείμενο των αρχείων, και στην συνέχεια αποτέλεσαν τις ιστοσελίδες.

Η ανακάλυψη του Gopher, γέννησε την ανάγκη για προγράμματα που έχουν την δυνατότητα εντοπισμού πληροφοριών μέσα από τους καταλόγους του Gopher. Τα δεδομένα αυτά συντέλεσαν στην κατασκευή του προγράμματος Veronica (VeryEasyRodent-OrientedNet-wideIndextoComputerizedArchives) και το jughead (Jonzy'sUniversalGopherHierarchyExcavation and Display) τα οποία αναζητούν αρχεία και κείμενα που είχαν αποθηκευτεί στο Gopher σύστημα. Η λειτουργία αυτών των δυο προγραμμάτων διακρίνονταν από τις ίδιες διαδικασίες, δίδοντας την δυνατότητα στους χρήστες να ψάξουν τους καταλόγους με τις πληροφορίες με την χρήση λέξεων ή φράσεων. Η πρώτη πιο κοντινή απόπειρα ανάπτυξης λογισμικού που πλησίαζε τις σημερινές μηχανές αναζήτησης έγινε το 1993 από τον MathewGray και ονομάστηκε Wandex. Ήταν ένας webcrawler που δημιουργούσε και καταλόγους αλλά έψαχνε και τις ιστοσελίδες στο διαδίκτυο. Την πενταετία 1993 μέχρι το 1998 κατασκευάστηκαν όλες οι μεγάλες μηχανές αναζήτησης που ξέρουμε σήμερα (OfficialGoogleBlog, 2017).

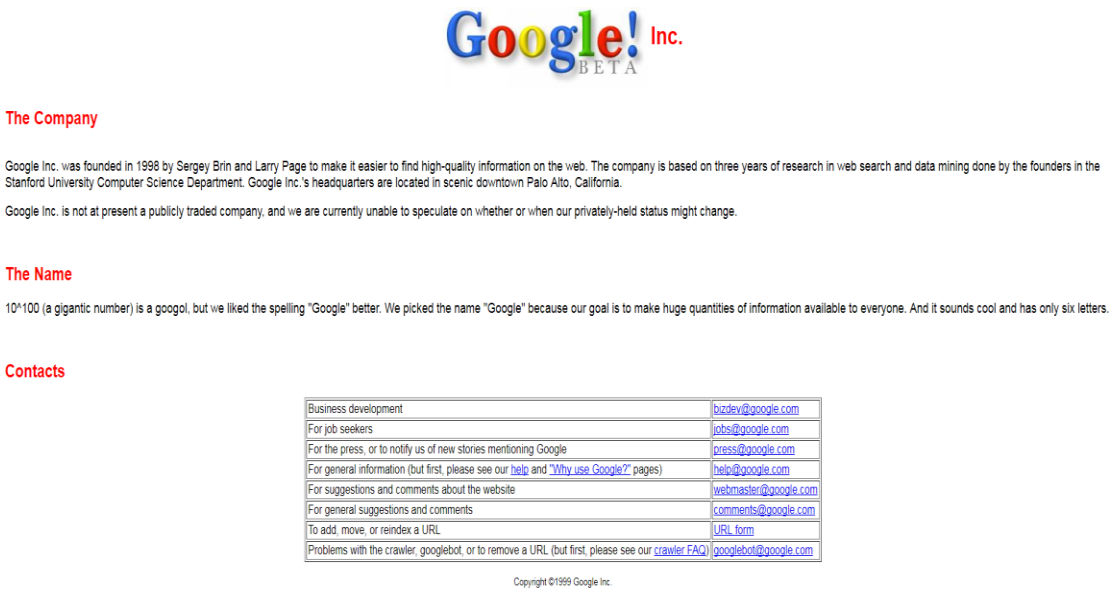
1.3 Μηχανές αναζήτησης

Οι μηχανές αναζήτησης των εμπορικών εταιρειών που απευθύνονται σε χρήστες-ιδιώτες είναι αρκετές και συνεχώς υπάρχει η προσπάθεια ανάδειξης νέων με εξελιγμένα εργαλεία και δυνατότητες. Από τις υπάρχουσες, οι μηχανές αναζήτησης της εταιρείας Google, το Mozilla και της Yahoo! Search αποτελούν τις πρώτες επιλογές των χρηστών στις προσπάθειες τους για είσοδο στο διαδίκτυο και στην αναζήτηση πληροφοριών.

Google

Για τους περισσότερους αποτελεί την μεγαλύτερη εταιρία διαδικτυακών υπηρεσιών. Ιδρύθηκε και ξεκίνησε την λειτουργία της Σεπτέμβριο του 1998 και αρχικός της σκοπός ήταν να οργανώσει όλες τις πληροφορίες του κόσμου και να τις προσφέρει σε παγκόσμιο επίπεδο. Το Google ξεκίνησε από μια κολεγιακή εργασία από τον Λάρρυ Πέιτζ και τον Σεργκέι Μπριν το 1996 σαν μια μηχανή αναζήτησης. Έκανε χρήση ενός αλγόριθμου ανάλυσης συνδέσμων ο οποίος ορίζει μια αριθμητική στάθμιση σε κάθε στοιχείο ενός συνόλου εγγράφων, όπως είναι το World Wide Web, με σκοπό να μετρήσει την ανάλογη σημασία του μέσα στο σύνολο. Τα αποτελέσματα του PageRank αποτελούν απόρροια της σημαντικότητας μιας σελίδας στο World Wide Web. Ένας σύνδεσμος υπερκειμένου σε μια σελίδα υπολογίζεται σαν έκφραση ενός χρήστη που έμεινε ευχαριστημένος από τα αποτελέσματα της. Το PageRank μιας ιστοσελίδας καθορίζεται κατ' επανάληψη και σχετίζεται από τον αριθμό και την τιμή του PageRank όλων των σελίδων που οδηγούν σε αυτήν. Μια σελίδα που σχετίζεται με πολλές σελίδες με υψηλό PageRank λαμβάνει η ίδια ένα υψηλό PageRank. Στην περίπτωση που δεν διακρίνονται σύνδεσμοι προς μια ιστοσελίδα δεν υπάρχει τιμή PageRank για αυτήν την σελίδα (http://el.wikipedia.org/wiki/Google).

Εικόνα 1.1Η ιστοσελίδα της Google το 1999.



Google! BETA Inc.

The Company

Google Inc. was founded in 1998 by Sergey Brin and Larry Page to make it easier to find high-quality information on the web. The company is based on three years of research in web search and data mining done by the founders in the Stanford University Computer Science Department. Google Inc.'s headquarters are located in scenic downtown Palo Alto, California.

Google Inc. is not at present a publicly traded company, and we are currently unable to speculate on whether or when our privately-held status might change.

The Name

10^100 (a gigantic number) is a googol, but we liked the spelling "Google" better. We picked the name "Google" because our goal is to make huge quantities of information available to everyone. And it sounds cool and has only six letters.

Contacts

Business development	bizdev@google.com
For job seekers	jobs@google.com
For the press, or to notify us of new stories mentioning Google	press@google.com
For general information (but first, please see our help and "Why use Google?" pages)	help@google.com
For suggestions and comments about the website	webmaster@google.com
For general suggestions and comments	comments@google.com
To add, move, or reindex a URL	URL form
Problems with the crawler, googlebot, or to remove a URL (but first, please see our crawler FAQ)	googlebot@google.com

Copyright ©1999 Google Inc.

Πηγή: https://web.archive.org/web/19990221202430/www.google.com/company.html

Yahoo!

Η Yahoo άρχισε σαν χομπύ κάποιον σπουδαστών που κατέληξε πια να είναι ένα παγκόσμιο brand, το οποίο έχει κατορθώσει να αποτελεί μια από τις σημαντικότερες μηχανές μέσο της οποίας επικοινωνούν οι άνθρωποι, βρίσκονται, αποκτούν πρόσβαση σε πληροφορίες και αγοράζουν προϊόντα και υπηρεσίες. Οι δύο ιδρυτές του Yahoo, ήταν ο David Filo και Jerry Yang, υποψήφιοι διδάκτορες ηλεκτρολογίας στο Πανεπιστήμιο του Στάνφορντ, ξεκίνησαν τον οδηγό τους σε ένα ρυμουλκόμενο της πανεπιστημιούπολης τον Φεβρουάριο του 1994 με στόχο να παρακολουθούν αυτά που τους ενδιαφέρουν στο διαδίκτυο. Κάποια στιγμή αντιλήφθηκαν ότι αφιέρωναν περισσότερο χρόνο στους καταλόγους των συνδέσεων που τους ενδιέφεραν σε σχέση με τις διδακτορικές τους διατριβές. Τελικά, οι λίστες του Τζέρι και του Ντέιβιντ έγιναν πολύ μεγάλες και δύσκολες. Έτσι αποφάσισαν να τις κατηγοριοποιήσουν και στην συνέχεια να αναπτύξουν υποκατηγορίες, πάνω σε αυτήν την υποκατηγοριοποίηση γεννήθηκε η ιδέα της δημιουργίας της Yahoo!.

Ο Τζέρι και ο Ντέιβιντ αντιλήφθηκαν αρκετά σύντομα ότι δεν ήταν οι μόνοι που έχουν ανάγκη από ένα εργαλείο για να βρουν χρήσιμες τοποθεσίες στο διαδίκτυο. Με το πέρασμα του χρόνου, εκατοντάδες άνθρωποι είχαν πρόσβαση στον οδηγό τους πολύ πιο μακριά από το ρυμολκούμενο του Στάνφορντ. Έτσι η ιδέα τους εξαπλώθηκε από τους φίλους τους και γρήγορα έγινε σημαντική, και του ακολούθησαν ένα μεγάλο πλήθος χρηστών στην στενά συνδεδεμένη κοινότητα του Διαδικτύου.

Η κυκλοφορία και ο ενθουσιασμός της υποδοχής που έλαβε το Yahoo, έκανε τους ιδρυτές της να προχωρήσουν στη δημιουργία μιας επιχείρησης. Τον Μάρτιο του 1995, συναντήθηκαν με δεκάδες επενδυτές κεφαλαίων της SiliconValley. Τελικά έρχονται σε επαφή με την Sequoia Capital, των οποίων οι πιο επιτυχημένες επενδύσεις ήταν η Apple Computer, η Atari, η Oracle και η Cisco Systems. Συμφώνησαν να χρηματοδοτήσουν το Yahoo τον Απρίλιο του 1995 με αρχική επένδυση ύψους περίπου 2 εκατομμυρίων δολαρίων.

Έτσι, το Yahoo! Inc. σήμερα είναι μια από τις σημαντικότερες εταιρείες παγκόσμιων επικοινωνιών, εμπορίου και μέσων επικοινωνίας στο διαδίκτυο που προσφέρει ένα ολοκληρωμένο δίκτυο επώνυμων υπηρεσιών σε περισσότερα από 345 εκατομμύρια άτομα σε μηνιαία βάση παγκοσμίως. Ο πρώτος οδηγός πλοήγησης στο διαδίκτυο, ο διαδικτυακός τόπος www.yahoo.com είναι από τους σημαντικότερους οδηγούς σε σχέση με την κυκλοφορία, τη διαφήμιση, την πρόσβαση των χρηστών και των επιχειρήσεων. Η εταιρία παρέχει επίσης σε απευθείας σύνδεση επιχειρηματικές υπηρεσίες που αποσκοπούν στην ενίσχυση της παραγωγικότητας και της παρουσίας των χρηστών του Yahoo! Αυτές οι υπηρεσίες περιλαμβάνουν το CorporateYahoo !, ένα δημοφιλές προσαρμοσμένο επιχειρηματικό portal ροής ήχου και βίντεο. Μπορεί να φιλοξενήσει και να διαχειριστεί καταστήματα και εργαλεία και υπηρεσίες ιστότοπου. Το παγκόσμιο δίκτυο Web της εταιρείας περιλαμβάνει 25 παγκόσμιες ιδιότητες. Με έδρα το Sunny vale της Καλιφόρνια, το Yahoo! έχει γραφεία στην Ευρώπη, την Ασία, τη Λατινική Αμερική, την Αυστραλία, τον Καναδά και τις Ηνωμένες Πολιτείες.
(http://archive.is/20120712130315/http://docs.yahoo.com/info/misc/history.html#selection-8.5-97.961).

Εικόνα 1.2: Κεντρικά γραφεία του Yahoo! Inc., στην Καλιφόρνια

Πηγή: https://www.britannica.com/topic/Yahoo-Inc

Bing

Για αυτήν παλαιότερες ονομασίες ήταν LiveSearch, Windows Live Search, MSN Search. Αποτελούσε μια τρέχουσα πολυγλωσσική μηχανή αναζήτησης Ιστού, που είχε δημιουργηθεί από την Microsoft. Την παρουσίαση της είχε πραγματοποιήσει ο SteveBallmer στις 28 Μαΐου 2009 στο *All Things Digital* συνέδριο στο ΣανΝτιέγκο. Οι βασικότερες μεταβολές σχετίζονταν με την λίστα των προτάσεων αναζήτησης σε πραγματικό χρόνο καθώς οι ερωτήσεις εισάγονται και μία λίστα σχετικών αναζητήσεων στηριζόμενη στη σημασιολογική τεχνολογία από την Powerset, που αγόρασε το 2008 η Microsoft αγόρασε. Η Bing συμπεριέλαβε ακόμα τη δυνατότητα *Save&Share* ιστορικών αναζήτησης μέσω των Windows Live ,SkyDrive,Facebook και ηλεκτρονικού ταχυδρομείου. Στις 29 Ιουλίου 2009, Microsoft και Yahoo! ανακοίνωσαν μια συμφωνία στην οποία η Bing θα τροφοδοτούσε την Yahoo! Search (http://en.wikipedia.org/wiki/Bing_search).

Εικόνα1.3: Λογότυπο της Bing

Πηγή: https://twitter.com/bing

Excite

Αποτελεί μια πύλη Διαδικτύου και μηχανή αναζήτησης Παγκόσμιου Ιστού. Η λειτουργιά της ξεκίνησε το 1994 ως Architext. Ήταν μία από τις σημαντικότερες "dotcom" "πύλες" στην δεκαετία του '90 και ένα από τα πιο εμπορικά σήματα τα οποία υπάρχουν στο διαδίκτυο. Τώρα πια έχει την δυνατότητα να παρέχει ποικίλες υπηρεσίες, συμπεριλαμβανομένης της αναζήτησης, web ηλεκτρονικό ταχυδρομείο, instant messaging, τα αποσπάσματα αποθεμάτων και μια διαμορφώσιμη από τον χρήστη της αρχικής σελίδας (http://www.excite.com/).

Εικόνα1.3: Λογότυπο της Excite

Πηγή: http://www.excite.com/

Ask.com

Αποτελεί μια μηχανή αναζήτησης που ξεκίνησε την λειτουργία της το 1996 από τον Garrett Gruener και τον David Warthen στο Μπέρκλεϋ της Καλιφόρνια. Το αρχικό λογισμικό της μηχανής αναζήτησης δημιουργήθηκε από τον Gary Chevskysky. Η Ask.com αποτελεί ιδιοκτησία της IAC/InterActiveCorp. Τον Σεπτέμβριο του 2001, η εταιρεία εξαγόρασε την Teoma Technologies και ξεκίνησε μια μεταβολή στους αλγόριθμους της Ask προς την αναζήτηση με φυσική γλώσσα. Το Ask Jeeves αποκτήθηκε από την IAC το Μάρτιο του 2005 για 1,85 δισεκατομμύρια δολάρια και μετονομάστηκε σε Ask.com το 2006.Τον Ιούλιο του 2010, η Ask.com επέστρεψε στις ρίζες της με την έκδοση beta των ερωτοαπαντήσεων (Q & A), η οποία χρησιμοποιεί ένα υβρίδιο τεχνολογίας αναζήτησης και ανθρώπινης αντίδρασης για την αντιμετώπιση πολύπλοκων ερωτήσεων ή ερωτήσεων με βάση τη γνώμη. Η εταιρεία ξεκίνησε ακόμα την εφαρμογής κινητής τηλεφωνίας το φθινόπωρο του 2010.

Το 2012, η Ask.com πραγματοποίησε δύο ακόμα εξαγορές ως μέρος μιας ευρύτερης στρατηγικής για να προσφέρει περισσότερο περιεχόμενο στην ιστοσελίδα του Ask.com. Στις 2 Ιουλίου 2012, το Ask.com αγόρασε την ανακάλυψη, NRelate, για ένα ποσό που δεν ανακοινώθηκε πότε. Και αφορούσε την απόκτηση εταιρείας συμβουλών και εμπειρογνωμόνων για πληροφορίες site site.co.uk, η οποία έκλεισε το Σεπτέμβριο του 2012 (https://www.crunchbase.com/organization/ask-com#/entity).

Εικόνα1.3: Κεντρικά γραφεία της Ask, στο Οκλαντ

Πηγή: https://en.wikipedia.org/wiki/Ask.com#/media/File:Askcomheadquarters.jpg

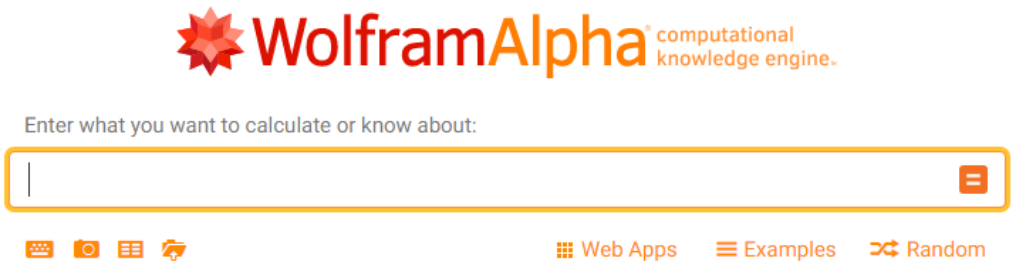
Wolfram Alpha

Είναι μια μηχανή «απάντησης» (answerengine) που αναπτύσσεται από την Wolfram Research. Αποτελεί μια υπηρεσία Διαδικτύου που απαντά στις πραγματικές ερωτήσεις άμεσα, με τον υπολογισμό της απάντησης από τα δομημένα δεδομένα, και όχι με την παροχή μιας λίστας εγγράφων ή ιστοσελίδων που να περιέχει την απάντηση, όπως μια μηχανή αναζήτησης. Ξεκίνησε την λειτουργιά της το Μάρτιο του 2009 από τον Stephen Wolfram, και κυκλοφόρησε στο κοινό στις 15 Μαΐου 2009.Στο κέντρο βρίσκεται η επαναστατική γλώσσα Wolfram, η οποία προσδιορίζει μια μοναδική σύγκλιση υπολογισμών και γνώσεων.

Το Wolfram Alpha χρησιμοποιείται από εκατομμύρια ανθρώπους καθημερινά στο διαδίκτυο, μέσω εφαρμογών για κινητά και ευφυών βοηθών και σε επιχειρήσεις,

το WolframAlpha αντιπροσωπεύει ένα από τα πιο σύνθετα και φιλόδοξα προγράμματα λογισμικού όλων των εποχών και είναι ένα σημαντικό πνευματικό και τεχνολογικό επίτευγμα.Επικεφαλής είναι ο Διευθύνοντας Σύμβουλος StephenWolfram(http://www.wolfram.com/company/background.html?source=nav).

Εικόνα1.3: Λογότυπο της Wolfram Alpha



Πηγή: http://www.wolframalpha.com/?source=frontpage-immediate-access

Yippy

Σε αρκετές περιπτώσεις το ζητούμενο δεν είναι η δημοφιλέστερη αναζήτηση, η οποία μπορεί να προσδιορίζεται στην βάση διαφορετικών και αρκετών παραγόντων που δεν σχετίζονται ούτε με την εγκυρότητα, ούτε με την πληρότητα της πληροφορίας την οποία ένα χρήστης αναζητά. Ο αχανείς κόσμος του διαδικτύου και συγκεκριμένες ιστοσελίδες, θησαυροί, αντιλαμβάνονται καλά το παιχνίδι και έχουν την δυνατότητα να "καλυφθούν" από τα αδιάκριτα μάτια αρκετών χρηστών.

Αυτό όμως δε συμβαίνει και για το Yippy, τη μηχανή αναζήτησης του "βαθέος ιστού" ή στην αγγλική γλώσσα "deepweb", που μέσα από την επεξεργασία και την παράγωγη συνδυαστικών αποτελεσμάτων συμβατικών μηχανών αναζήτησης προσφέρει ένα αποτέλεσμα σε βάθος για όποιο ζήτημα απασχολεί τον χρήστη. Η μηχανή παρέχει αποτελέσματα σε μορφή "cloud" και είναι πολύ πιθανό να προσφέρει ιστοσελίδες, που σε διαφορετική περίπτωση θα παρέμεναν θαμμένες και δε θα ήταν εύκολο να εντοπιστούν (http://www.yippyinc.com/company-0).

Εικόνα1.4: Λογότυπο της Yippy

Πηγή: https://yippy.com/

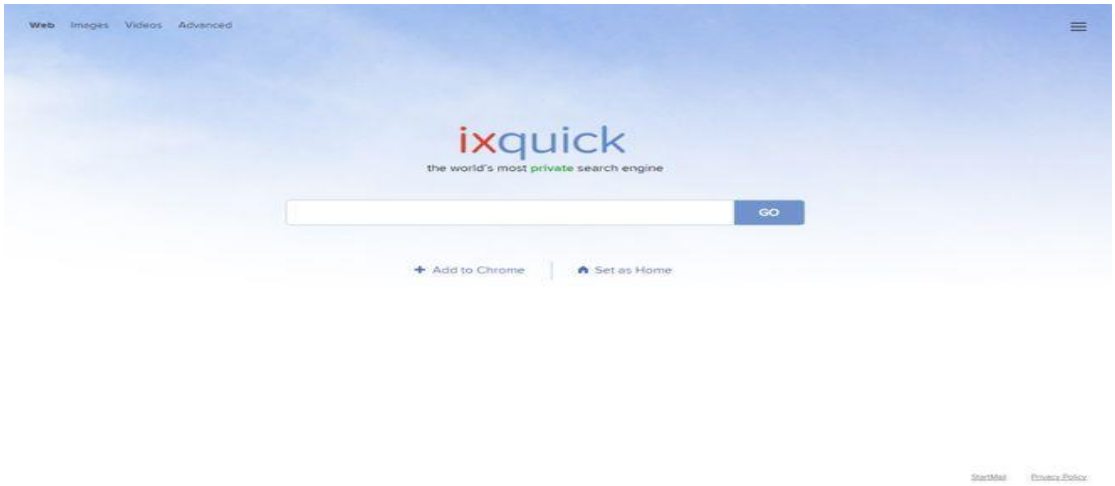
Kosmix

Αποτελεί μια μηχανή αναζήτησης-κατηγοριοποίησης που οργανώνει το Διαδίκτυο σε θεματικές σελίδες και προσφέρει την δυνατότητα στους χρήστες να μελετήσουν τον Παγκόσμιο Ιστό με βάση το θέμα, παρουσιάζοντας ένα ταμπλό των σχετικών βίντεο, φωτογραφιών, ειδήσεων, σχολίων, απόψεων και των συνδέσμων με θέματα που σχετίζονται με αυτό που αναζητά ο χρηστής. Ο Venky Harinarayan και ο Anand Rajaraman ίδρυσαν την Kosmix το 2005.

xquick

Αυτή η μηχανή αναζήτησης δίνει ιδιαίτερη έμφαση στα δεδομένα και ειδικότερα στα προσωπικά δεδομένα. Αυτό προκύπτει από το γεγονός ότι δεν ζητεί από το χρήστη στοιχεία για να δημιουργήσει Cookies όπως και δεν συλλέγεται στο ιστορικό προηγούμενων αναζητήσεων. Η μόνη επιλογή που έχει την δυνατότητα να πραγματοποιήσει κάποιος χρηστής της είναι η διαμόρφωση των προτιμήσεων του, και αυτό με την προϋπόθεση ότι μετά από τρεις μήνες αδράνειας διαγράφονται οριστικά (https://www.ixquick.com/eng/privacy-policy.html#hmb).

Εικόνα1.5: Λογότυπο της xquick

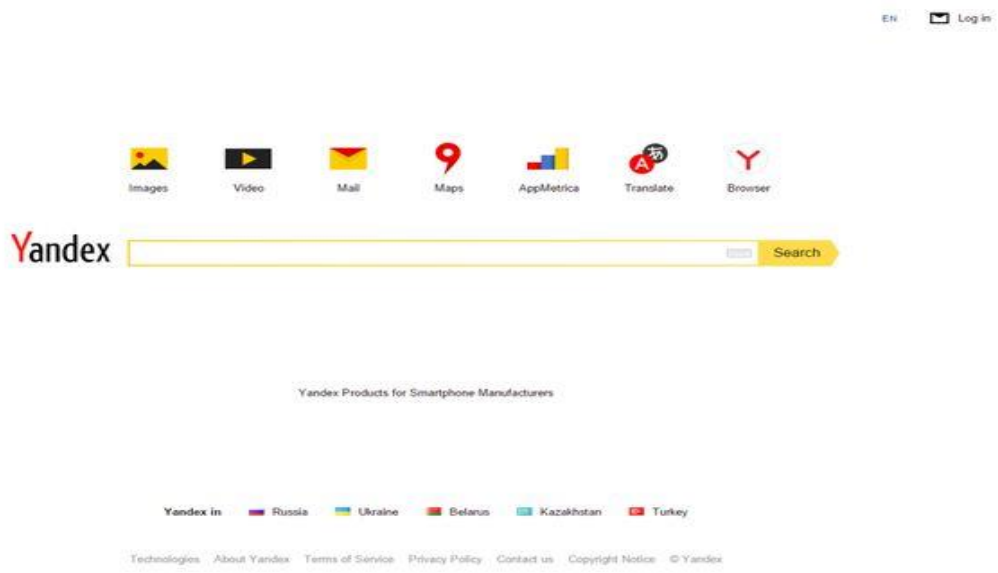


Πηγή: https://www.ixquick.com/

Yandex

Προσδιορίζεται ως μια ρωσική μηχανή αναζήτησης, το Yandex κατασκευάστηκε το 1997 και αποτελούσε κάτι σαν μια η ρωσική απάντηση στην αμερικανική Google. Εκτός ρωσικών συνόρων δεν είναι ιδιαίτερα γνωστή, όμως αποτελεί την πρώτη επιλογή των Ρώσων και χαρακτηριστικό είναι το στοιχείο ότι πραγματοποιεί 150 εκατ. αναζητήσεις την ημέρα. Διαθέτει σχεδόν όλα τα εργαλεία που έχει και κάποιος χρηστής της google, ενώ παρά την εθνική του ταυτότητα που την διακρίνει, προσφέρει ευρείες επιλογές σε αναζητήσεις σε ξένες γλώσσες. Αρκετοί είναι αυτή που την αναφέρουν ως καλύτερη εναλλακτική στην Google, χάρη στις πλούσιες πηγές που έχει στα εργαλεία της.

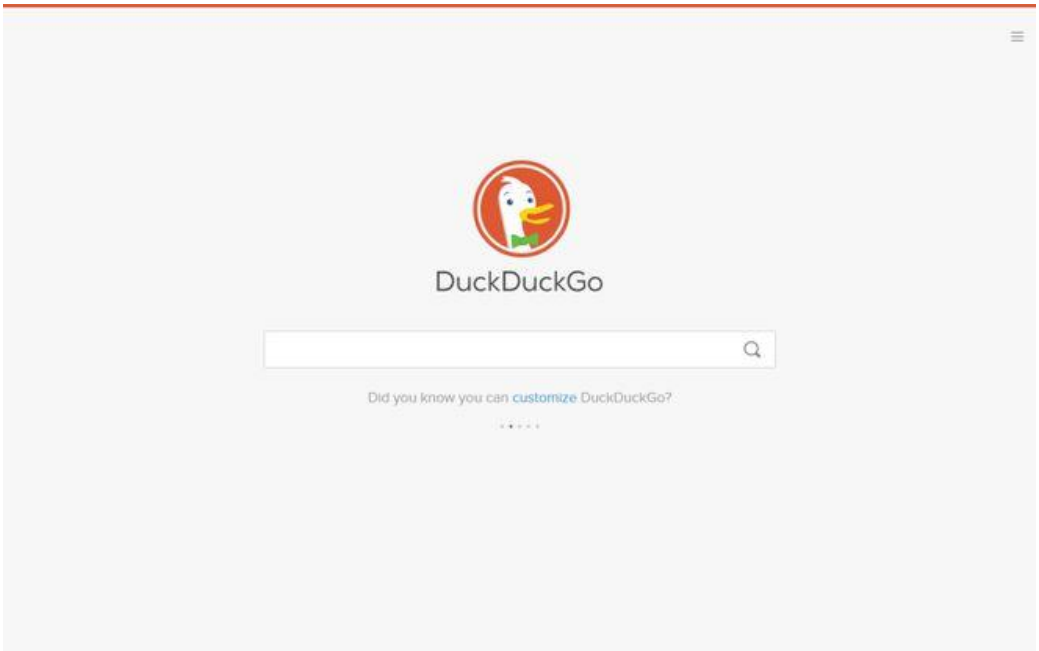
Εικόνα1.6: Λογότυπο της Yandex



DuckDuckGo.

Αποτελεί ιδανικό εργαλείο για "ιδιωτικές" αναζητήσεις, Η επιλογή της μηχανής που έχει στο λογότυπο της την πάπια, αποτελεί ιδανική μηχανή αναζήτησης για αυτούς που επιδιώκουν να διατηρήσουν την ανωνυμία τους στο διαδίκτυο. Το DuckDuckGo δεν χρησιμοποιεί προσωπικές πληροφορίες, ούτε λαμβάνει στοιχεία από προφίλ σε socialmedia, email και ιστορικό αναζήτησης για να προσφέρει πιο προσωποποιημένα αποτελέσματα, πρακτικές τις οποίες μεταχειρίζεται σε πολύ μεγάλο βαθμό η Google (https://duckduckgo.com/privacy)

Εικόνα1.7: Λογότυπο της DuckDuckGo



Πηγή: https://duckduckgo.com/

DogPile

Η επιστροφή μιας πρωτοπόρου μηχανής αναζήτησης στη δεκαετία του '90 το DogPile αποτελούσε μια ιδιαίτερα δημοφιλή επιλογή των χρηστών, πριν ακόμη την σαρωτική εμφάνιση της Google. Σήμερα, οι υπεύθυνοι της μηχανής προσπαθούν να κάνουν το comeback, προσφέροντας επίσης ενδιαφέροντες επιλογές σε φίλτρα και προτιμήσεις αναζήτησης (http://infospace.com/terms/privacy.html).

Εικόνα1.8: Λογότυπο της DogPile

Πηγή: http://www.dogpile.com/

Οι μηχανές αναζήτησης στην σημερινή εποχή προσδιορίζονται ως πολύπλοκα προγράμματα που προσφέρουν την δυνατότητα αναζήτησης από οποιαδήποτε πληροφορία από αρχεία, έγγραφα, εικόνες, video μέχρι λέξεις ή φράσεις που γίνεται χρήση στην καθημερινή ζωή του ανθρώπου. Οι ξένες Μηχανές Αναζήτησης, διακρίνονταν από σχετικά προβλήματα χαρακτήρων και σελίδων στην ελληνική γλωσσά και παρουσιάστηκε η ανάγκη ανάπτυξης ελληνικών Μηχανών Αναζήτησης, οι οποίες να μπορούν να αναγνωρίζουν ελληνικές ιστοσελίδες τόσο ξένες αλλά και σελίδες με ελληνικούς και ξένους χαρακτήρες ταυτόχρονα. Σήμερα όλες σχεδόν οι Μηχανές Αναζήτησης έχουν τη δυνατότητα να αναγνωρίζουν και να προσφέρουν στους χρήστες των ιστοσελίδων που περιέχουν χαρακτήρες των περισσότερων γλωσσών του κόσμου. Στην σημαντικότερες ελληνικές Μηχανές κατατάσσονται οι ακόλουθες(http://www.goonline.gr/ebusiness/specials/article.html?article_id=22):

- **Robby**: Αποτελεί την παλαιότερη από τις ελληνικές Μηχανές Αναζήτησης και χρησιμοποιεί το AltaVista για τις αναζητήσεις που πραγματοποιεί. Ταυτόχρονα προσφέρει, στους χρήστες ένα πλήθος άλλων υπηρεσιών, όπως

για παράδειγμα νέα, για τον καιρό, το χρηματιστήριο, το συνάλλαγμα, κ.α(http://www.robby.gr)

Εικόνα1.9: Λογότυπο της Robby

Πηγή: http://www.robby.gr/

- **Trinity**: χαρακτηρίζεται ως μια Μηχανή Αναζήτησης της ελληνικής πύλης Pathfinder. Παρέχει στους χρήστες τη δυνατότητα σύνθετης αναζήτησης αλλά δεν διαθέτει δικό της θεματικό κατάλογο, αλλά παραπέμπει στο θεματικό κατάλογο της Pathfinder (http://www.trinity.gr)

Εικόνα1.10: Λογότυπο της Trinity

Πηγή: http://www.trinity.gr

- **FORTHnet**: Η συγκεκριμένη μηχανή αναζήτησης κάνει χρήση της τεχνολογίας της AltaVista και προσφέρει στους χρήστες ταχύτατες και ακριβείς αναζητήσεις. Οι χρήστες της Μηχανής Αναζήτησης FORTHnet έχουν την δυνατότητα να επιλέξουν την αναζήτηση ειδησεογραφικών ιστοσελίδων, ακόμα και μέσα στις κατηγορίες του Forthnet directory. Επίσης, η FORTHnet παρέχει ευρετήριο για χρήσιμες συμβουλές για εύκολη καιγρήγορη χρήση της Μηχανής Αναζήτησης. (http://search.forthnet.gr).

Εικόνα1.11: Λογότυπο της Forthenet

Πηγή: http://search.forthnet.gr

- **Anazitisis**: Αποτελεί μια Μηχανή Αναζήτησης από την OTE net και προσφέρει στο χρήστη τη δυνατότητα σύνθετης αναζήτησης και αναλυτικές οδηγίες για την χρησιμοποίηση του. Αποτελεί την μοναδική Μηχανή Αναζήτησης που διακρίνεται για την ικανότητα της πάνω στις ιδιομορφίες της ελληνικής γλώσσας, καθώς επίσης διαθέτει δικό της θησαυρό λέξεων. (http://anazitisis.gr).

Εικόνα1.12: Λογότυπο της anazitis

Πηγή: http://anazitisis.gr

1.4 Μηχανές Μετά-Αναζήτησης

Προσδιορίζονται ως εφαρμογές που εκτελούν για λογαριασμό του χρήστη την αναζήτηση σε ένα πλήθος μηχανών και παρουσιάζουν αναδιοργανωμένα και συνοπτικά τα αποτελέσματα που προκύπτουν αφαιρώντας τις επαναλαμβανόμενες εγγραφές που προκύπτουν. Η διαδικασία αυτή εκτελείται με σκοπό την αξιοποίηση των δυνατοτήτων κάθε μηχανής, με τελικό στόχο ο χρήστης να εντοπίσει με μεγαλύτερη ευκολία τις πηγές που του είναι χρήσιμες.

Είναι βέβαιο πως με το πλήθος των ιστοσελίδων που υπάρχουν σήμερα στο διαδίκτυο, οι Μηχανές Αναζήτησης επιστρέφουν μετά από την αναζήτηση μιας πληροφορίας από τον χρήστη αρκετό υλικό. Στον περιορισμό αυτού του πλήθους και την εστίαση σε απολύτως περιεκτικά αποτελέσματα κατά την διάρκεια της αναζήτησης μιας πληροφορίας επιβάλλεται κανείς να λάβει υπόψη του τις λεγόμενες Μηχανές Μετά-Αναζήτησης (MetaSearchEngines). Οι Μηχανές Μετά-Αναζήτησης στην ουσία αποτελούν μηχανές αναζήτησης για τις Μηχανές Αναζήτησης και ουσιαστικά πραγματοποιούν ερωτήματα για ανάκτηση πληροφοριών σε αρκετές Μηχανές Αναζήτησης παράλληλα.

Ο τρόπος λειτουργίας τους ταυτίζεται με αυτών των απλών Μηχανών Αναζήτησης. Ο χρήστης πληκτρολογεί στη πλαίσιο εισαγωγής του ερωτήματος τις λέξεις-κλειδιά ή άλλες λέξεις που σχετίζονται το ζήτημα για το οποίο επιθυμεί την αναζήτηση της πληροφορίας. Με το πάτημα του κουμπιού για την έναρξη της αναζήτησης, η Μετα-Μηχανή αποστέλλει το ερώτημα του χρήστη παράλληλα και σε πολλές, διαφορετικές, απλές Μηχανές Αναζήτησης και συνεπώς στις βάσεις δεδομένων με web σελίδες αυτών. Με το πέρασμα μικρού χρονικού διαστήματος, η Μετά-Μηχανή επιστρέφει στο χρήστη τα αποτελέσματα που έχει συγκεντρώσει από όλες τις απλές Μηχανές Αναζήτησης στις οποίες μεταφέρει το ερώτημα του χρήστη. Μια πιο πολύπλοκη Μηχανή Μετά-Αναζήτησης προσφέρει στον χρήστη τη δυνατότητα να προσδιορίσει πολύπλοκες παραμέτρους με βάση τις οποίες θέλει να πραγματοποιηθεί η αναζήτηση πληροφορίας που σχετίζεται με το συγκεκριμένο θέμα που έχει εκδηλώσει ενδιαφέρον.

Η λειτουργία μιας τέτοιας μηχανής στηρίζεται και σε απλές Μηχανές Αναζήτησης. Ακόμα, όπως και στις απλές Μηχανές Αναζήτησης, είναι δυνατή στις Μηχανές Μετά-Αναζήτησης η χρήση των Boolean τελεστών AND, OR και NOT, καθώς και του τελεστή προσέγγισης NEAR, κατά την διαδικασία διατύπωσης των ερωτημάτων από το χρήστη. Οι Μηχανές Μετά-Αναζήτησης δεν έχουν δικές τους βάσεις δεδομένων με web σελίδες, όπως στην περίπτωση των απλών μηχανών αναζήτησης. Η διαδικασία που ακολουθούν είναι η ανάγνωση των ερωτημάτων των χρηστών στις βάσεις δεδομένων των απλών Μηχανών Αναζήτησης. Μια Μετά-Μηχανή Αναζήτησης χρειάζεται περισσότερο χρόνο για την εκτέλεση ενός ερωτήματος καθώς επιβάλλεται να διενεργήσει ελέγχους σε αρκετές Μηχανές Αναζήτησης που αφορούν αυτό το ερώτημα. Το σημείο στο οποίο διαθέτουν περισσότερα πλεονεκτήματα οι Μηχανές Μετά-Αναζήτησης έναντι των απλών Μηχανών Αναζήτησης είναι ότι σε αρκετές περιπτώσεις επιστρέφουν απαντήσεις σε σχετικά ασαφείς ερωτήσεις του χρήστη που μια απλή Μηχανή μπορεί να απωλέσει. Την δεδομένη χρονική στιγμή υπάρχουν τρεις τύποι Μηχανών Μετά-Αναζήτησης:

- Εργαλεία για ανάκτηση πληροφορίας σε αρκετές βάσεις που ζητά ο χρήστης μέσα σε αποτελέσματα αναζήτησης. Αυτά τα εργαλεία διακρίνονται για την χρήση τους και την καταλληλόλητα τους από ερευνητές που επιζητούν μια σε μεγαλύτερο βάθος για την αναζήτηση ανάκτηση πληροφοριών που σχετίζονται με κάποιο ζήτημα. Μέσα από αυτό προσφέρεται η δυνατότητα στο χρήστη να στοχεύσει σε συγκεκριμένη περιοχή, για παράδειγμα Αμερική από την οποία επιθυμεί να προκύψουν τα αποτελέσματα της αναζήτησης ή ακόμη και συγκεκριμένους δικτυακούς τόπους.
- Μηχανές Μετά-Αναζήτησης που διενεργούν πολύπλοκες αναζητήσεις, ενοποιούν τα αποτελέσματα καλά, δεν επιτρέπουν τις διπλό-εμφανίσεις αποτελεσμάτων και παρέχουν επιπρόσθετες επιλογές, όπως έξυπνη ταξινόμηση ή ομαδοποίηση κατά πεδία των αποτελεσμάτων της αναζήτησης. Η Savvy Search (http://savvy.search.com/) αναφέρεται ως μια από τις παλαιότερες Μηχανές μετά-αναζήτησης που συνυπολογίζεται σε αυτήν την κατηγορία. Παρέχει στο χρήστη τη δυνατότητα να προσδιορίζει τον αριθμό των αποτελεσμάτων που επιθυμεί να του επιστραφούν από κάθε μια από τις απλές Μηχανές Αναζήτησης που θα ερωτηθούν. Η Clusty (http://clusty.com/) είναι ακόμα μια μηχανή μετά-αναζήτησης που περιλαμβάνεται σε αυτή την

κατηγορία που παρουσιάζει τα αποτελέσματα στον χρήστη ανά συστάδες, δηλαδή μια ομάδα από παρόμοια ζητήματα που είναι συναφή με την αρχική ερώτηση του χρήστη.

- Μηχανές μετά-αναζήτησης που πραγματοποιούν αναζητήσεις σε αρκετά σημεία και επιστρέφουν τα αποτελέσματα χωρίς τις επιλογές που προαναφέρθηκαν. Σε αυτή την κατηγορία ανήκουν πολλές Μηχανές μετά-αναζήτησης. Σε αυτή την κατηγορία ανήκει η Dogpile (http://www.dogpile.com). Μεταφέρει το ερώτημα του χρήστη σε 25 απλές Μηχανές Αναζήτησης. Μερικές από αυτές είναι: Excite, Lycos, InfoSeek, WebCraweler, Thunderstone, PlanetSearch καιYahoo (http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html)

Τα πλεονεκτήματα αυτών των μηχανών προσκοπούν μέσα από την εξοικονόμηση χρόνου διότι η αναζήτηση του ιδίου στοιχείου σε κάθε μηχανή ξεχωριστά αποτελεί μια χρονοβόρα και επίπονη διαδικασία. Ακόμα ο χρήστης έχει την δυνατότητα να κάνει την ερώτησή του προς τη μετά-μηχανή, η οποία αναλαμβάνει να επαναδιατυπώσει την ερώτηση σε κατάλληλη μορφή για κάθε μηχανή αναζήτησης, ώστε να πραγματοποιηθεί η αναζήτηση χωρίς προβλήματα στην επεξεργασία.

1.5 Οφέλη από την χρήση μηχανών αναζήτησης

Μέσα από την χρήση των Μηχανών Αναζήτησης προσδιορίζονται αρκετά οφέλη τόσο για τους χρήστες του διαδικτύου όσο και για την κάθε εταιρία που διαθέτει έναν δικτυακό τόπο. Τα οφέλη που προκύπτουν για τον χρήστη από την ύπαρξη των Μηχανών Αναζήτησης πηγάζουν από το γεγονός ότι η αναζήτηση της πληροφορίας στο Διαδίκτυο πραγματοποιείται με αρκετά μεγάλη ευκολία, ανεξάρτητα από το αν το ζήτημα που αναζητά ο χρήστης είναι πολύπλοκο ή όχι. Ο χρήστης δεν είναι αναγκαίο να πηγαίνει από σελίδα σε σελίδα και από σύνδεσμο σε σύνδεσμο ώστε να μπορέσει να ανακτήσει την πληροφορία που χρειάζεται. Το μόνο αναγκαίο είναι να πληκτρολογήσει την αρχική σελίδα της Μηχανής Αναζήτησης, και έπειτα πάνω στο πεδίο που αυτή προσφέρει να πληκτρολογήσει τους όρους που περιγράφουν με όσο μεγαλύτερη προσέγγιση και περιεκτικότητα γίνεται το ζήτημα που τον απασχολεί και να αναμένει λίγα δευτερόλεπτα έως ότου η Μηχανή Αναζήτησης πραγματοποιήσει για λογαριασμό του την περιήγηση σε όλο το Web και του επιστρέψει σε μορφή λίστας αποτελεσμάτων όλες τις σχετικές με το συγκεκριμένο ζήτημα σελίδες που ανακάλυψε.

Οι διαδικασίες αυτές συντελούν στην ταχύτερη εξυπηρέτηση του χρήστη να εξυπηρετείται γρηγορότερα, ευκολότερα και πληρέστερα μέσα από την χρησιμοποίηση ενός πανίσχυρου εργαλείου. Για την εταιρία που έχει σελίδα στο διαδίκτυο τα οφέλη από την χρήση των Μηχανών Αναζήτησης προσδιορίζονται από ανάλογη σημαντικότητα. Αν φανταστούμε το σύνολο των χρηστών που κάνουν χρήση των Μηχανών Αναζήτησης στην καθημερινή τους ζωή για οποιοδήποτε ζήτημα είναι τεράστιος, τότε καταλαβαίνουμε ότι η παρουσία του δικτυακού τόπου της εταίρας στη λίστα αποτελεσμάτων που παράγει μια Μηχανής αναζήτησης σημαίνει αυτόματα τεραστία αύξηση των εν’ δύναμη χρηστών - πελατών που θα επισκεφτούν την ιστοσελίδα της και θα ενημερώνονται για την ύπαρξη αυτής. Έτσι μέσω των Μηχανών Αναζήτησης προκύπτει ένας καινούργιος τρόπος στην προσέλκυση, με ευκολότερο τρόπο και γρηγορότερη απόκριση περισσότερων πελατών - ενδιαφερόμενων για τα προϊόντα-υπηρεσίες της. Το σημαντικότερο όμως είναι ότι αυτό πραγματοποιείται χωρίς σημαντικό επιπρόσθετο κόστος για την ίδια. Συμπερασματικά αυτό που προκύπτει είναι ότι οι Μηχανές Αναζήτησης αποτελούν ένα πολύ δυνατό εργαλείο τόσο για τους χρήστες όσο και για τις εταίρες μέσα στον

αχανή κόσμο του Διαδικτύου (η-Επιχειρείν: Αφιερώματα: Εταιρική παρουσία στο Διαδίκτυο / Χρήσιμα εργαλεία / Οφέλη από τη χρήση των Μηχανών Αναζήτησης (http://www.goonline.gr/ebusiness/specials/article.html?article_id=232).

ΚΕΦΑΛΑΙΟ 2ο: Αναζήτηση, πάταξη και αποθήκευση δεδομένων

2.1 Αναζήτηση δεδομένων

Η αναζήτηση δεδομένων στο διαδίκτυο πραγματοποιείται στην ουσία όπως περιγράφηκε και στην πρώτη ενότητα με την μορφή υποβολής ερωτημάτων στις μηχανές αναζήτησης. Τα ερωτήματα γίνονται μέσα από την χρήση λέξεων-κλειδιών και δεν διακρίνεται να υπάρχει κάποια ιδιαίτερη σύνταξη ή γλώσσα ερωτημάτων στην οποία να γίνεται χρήση. Οι λέξεις-κλειδιά δεν είναι αναγκαίο να διαχωρίζονται με κόμματα, ενώ οι χαρακτήρες συν ή πλην δίπλα από κάποια λέξη δηλώνουν αν θα πρέπει οπωσδήποτε να υπάρχει, ή όχι, αυτή η λέξη στο περιεχόμενο όλων των ιστοσελίδων που επιστρέφονται ως απάντηση στο ερώτημα. Στην περίπτωση που οι όροι της αναζήτησης περιβάλλονται από εισαγωγικά, αυτό προσδιορίζει το γεγονός ότι αφορά η αναζήτηση μια ενιαία φράση η οποία θα πρέπει να είναι παρούσα στις ιστοσελίδες των αποτελεσμάτων. Οι τύποι των ερωτημάτων που κάνουν οι χρήστες στις μηχανές αναζήτησης προσδιορίζονται ακολούθως:

- **Πληροφοριακά ερωτήματα**: Ο χρήστης κάνει αναζήτηση γενικές πληροφορίες σε κάποιο θεματικό πεδίο. Τα ερωτήματα αυτού του είδους περιλαμβάνουν συνήθως από 1 έως 3 γενικούς όρους και είναι πιθανόν να επιστρέψουν χιλιάδες σχετικά αποτελέσματα στο χρήστη.
- **Ερωτήματα πλοήγησης**: Με ερωτήματα αυτού του ιδίους ο χρήστης αναζητεί κάποιον συγκεκριμένο ιστότοπο ή ιστοσελίδα και προσπαθεί να προσδιορίσει την ακριβή διεύθυνση μέσω της μηχανής αναζήτησης.
- **Ερωτήματα συναλλαγής**: Το ενδιαφέρον του χρήστη εστιάζει στο να πραγματοποιήσει μια συναλλαγή ή μια συγκεκριμένη ενέργεια, όπως για παράδειγμα το να πραγματοποιήσει μια αγοραπωλησία ενός συγκεκριμένου προϊόντος ή να κατεβάσει κάποιο πρόγραμμα.

Σε αυτό το σημείο κρίθηκε αναγκαίο να αναφερθούν κάποιες ενδιαφέρουσες στατιστικές που αφορούν μηχανές αναζήτησης από τους απλούς χρήστες:

- Το μέσο ερώτημα αποτελείται από 2έως 3 όρους, με αυξητικές τάσεις των όρων καθώς το ευρύ κοινό αποκτά μεγαλύτερη γνώση και ευχέρεια στην διαχείριση των μηχανών αναζήτησης.
- Κοντά στους μισούς χρήστες προσέχουν μόνο τις πρώτες δύο σελίδες των αποτελεσμάτων που τους εμφανίζει η μηχανή αναζήτησης.

- Το ποσοστό που κάνει χρήση λογικών τελεστών (AND, OR, NOT, +, -), στα ερωτήματα του είναι κάτω από 5%.
- Περίπου 30% των ερωτημάτων αφορούν ιδία επαναλαμβανόμενα από τον ίδιο χρήστη, ο οποίος επιλέγει το ίδιο αποτέλεσμα κάθε φορά. Στοιχείο που αποδεικνύει ότι αρκετοί χρήστες κάνουν χρήση των μηχανών αναζήτησης σαν ένα είδος σελιδοδείκτη για να ανακτήσουν παλαιότερες πηγές.

(Spink, at.el., 2001).

2.2 Αποθήκευση δεδομένων

Οι μηχανές αναζήτησης κάνουν χρήση ενός ευρετηρίου στη βάση δεδομένων τους για την αποθήκευση των ιστοσελίδων και των λέξεων-κλειδιών που έγιναν χρήση κατά την διάρκεια των ερωτημάτων των χρηστών. Ο στόχος της χρήσης του ευρετηρίου είναι ο περιορισμός του χρόνου που απαιτεί η μηχανή αναζήτησης για να προσφέρει απόκριση σε ένα ερώτημα. Χωρίς το ευρετήριο, το σύστημα θα έπρεπε να ανιχνεύσει την κάθε αποθηκευμένη ιστοσελίδα για τις λέξεις-κλειδιά που εισήγαγε ο χρήστης στο ερώτημα. Μια τέτοια διαδικασία, θα απαιτούσε αρκετά μεγάλο χρονικό διάστημα για να ολοκληρωθεί, ακόμα και στην περίπτωση μικρών δειγμάτων της τάξης των 10.000 ιστοσελίδων.

Από την άλλη πλευρά, με τη χρήση του ευρετηρίου, τα αποτελέσματα από 10.000 ιστοσελίδες είναι δυνατόν να ανασυρθούν σε λιγότερο από ένα δευτερόλεπτο. Βέβαια, ένα ευρετήριο απαιτεί μεγαλύτερο αποθηκευτικό χώρο από ότι θα απαιτούνταν για την αποθήκευση των ίδιων των δεδομένων ενώ η διαδικασία ανανέωσής του χαρακτηρίζεται πολύ χρονοβόρα.

Τα μειονεκτήματα μπορούν να χαρακτηριστούν αμελητέα από το μικρό χρονικό διάστημα εκτέλεσης των ερωτημάτων, ο οποίος τελικά κάνει ρεαλιστική τη λειτουργία και τη χρήση μιας μηχανής αναζήτησης. Τα είδη των ευρετηρίων που στα οποία γίνεται χρήση στις περισσότερες περιπτώσεις από τις σημαντικότερες εμπορικές μηχανές αναζήτησης είναι το ανεστραμμένο ευρετήριο και το προοπτικό ευρετήριο. Τα οποία θα αναλυθούν και θα παρουσιαστούν στην συνέχεια.

Πριν τον προσδιορισμό και την ανάλυση των ανεστραμμένων δεικτών, κρίνεται αναγκαίο πρώτα να προσδιοριστεί η διαδικασία κατασκευής αυτών των δεικτών. Για ακαδημαϊκούς και ερευνητικούς σκοπούς, αυτό μπορεί να χαρακτηρίζεται σχετικά απλό. Οι τυπικές συλλογές για την έρευνα ανάκτησης πληροφοριών είναι αρκετές και διαθέσιμες σε ένα σύνολο ποικίλων ειδών που μπορεί να είναι ένα απλό blog έως το κείμενο ειδήσεων. Για τους ερευνητές που ενδιαφέρονται να διερευνήσουν την ανάκτηση ιστού, υπάρχει η συλλογή ClueWeb09 που περιλαμβάνει 1 δισεκατομμύριο ιστοσελίδες σε δέκα γλώσσες οι οποίες προσδιοριστήκαν από το πανεπιστήμιο CarnegieMellon στις αρχές του 2009 (http://boston.lti.cs.cmu.edu/Data/clueweb09/).

Η απόκτηση πρόσβασης σε αυτές τις τυπικές συλλογές είναι στις περισσότερες περιπτώσεις πολύ απλή και γίνεται με την καταβολή ενός λογικού

τέλους και μεριμνώντας για την παραλαβή των δεδομένων. Για την αναζήτηση ιστού σε πραγματικό κόσμο, ωστόσο, δεν μπορεί κανείς απλώς να υποθέσει ότι η συλλογή είναι ήδη διαθέσιμη. Για την απόκτηση περιεχομένου ιστού είναι αναγκαία η ανίχνευση, η οποία είναι η διαδικασία διέλευσης του ιστού ακολουθώντας επανειλημμένα υπερσυνδέσμους και αποθηκεύοντας τις σελίδες που έχουν ληφθεί για μεταγενέστερη επεξεργασία.

Προσδιοριστικά, η διαδικασία χαρακτηρίζεται ως απλή στην κατανόηση: ξεκινάμε με τη συμπλήρωση μιας ουράς με μια λίστα σελίδων "σπόρων". Η σελίδα ανίχνευσης εμφανίζει τις σελίδες στην ουρά, εξάγει συνδέσμους από αυτές τις σελίδες για να προσθέσει στην ουρά, αποθηκεύει τις σελίδες για περαιτέρω επεξεργασία και επαναλαμβάνει. Στην πραγματικότητα, οι στοιχειώδεις ανιχνευτές ιστού είναι δυνατόν να γραφτούν σε μερικές εκατοντάδες γραμμές κώδικα. Ωστόσο, η αποτελεσματική και αποδοτική ανίχνευση ιστού χαρακτηρίζεται αρκετά πιο περίπλοκη. Στην συνέχεια προσδιορίζονται ορισμένα θέματα τα οποία πρέπει να αντιμετωπίσουν οι ανιχνευτές του πραγματικού κόσμου:

- Ένας ανιχνευτής ιστού επιβάλλεται να εξασκεί καλή "ετικέτα" και να μην επιβαρύνει τους διακομιστές ιστού. Για παράδειγμα, είναι συνηθισμένη η πρακτική να περιμένει κάποιος για κάποιο χρονικό διάστημα πριν από την επανειλημμένη αίτηση στον ίδιο διακομιστή. Προκειμένου να σέβονται αυτούς τους περιορισμούς διατηρώντας παράλληλα καλή απόδοση, ένας ανιχνευτής διατηρεί συνήθως πολλά νήματα εκτέλεσης που εκτελούνται ταυτόχρονα και διατηρεί ταυτόχρονα πολλές συνδέσεις TCPπου μπορεί να φτάσουν να είναι και χιλιάδες.
- Δεδομένου ότι ο ανιχνευτής έχει πεπερασμένο εύρος ζώνης και πόρους, είναι αναγκασμένος να δώσει προτεραιότητα στη σειρά με την οποία ανεβάστηκαν οι σελίδες. Οι αποφάσεις αυτές πρέπει να υλοποιούνται σε απευθείας σύνδεση και σε περιβάλλον αντιφάσεων, υπό την έννοια ότι οι αποστολείς ανεπιθύμητης αλληλογραφίας δημιουργούν ενεργές "συνδέσεις ζεύξης" και "παγίδες αράχνης", γεμάτες σελίδες ανεπιθύμητης αλληλογραφίας για να εξαπατήσουν έναν ανιχνευτή να παρουσιάσουν υπερβολικό περιεχόμενο για έναν συγκεκριμένο ιστότοπο.

- Οι περισσότεροι ανιχνευτές ιστού πραγματικού κόσμου είναι κατανεμημένα συστήματα που λειτουργούν σε συστοιχίες μηχανών, που στις περισσότερες περιπτώσεις είναι γεωγραφικά κατανεμημένες. Για να αποφευχθεί η λήψη σελίδων πολλαπλά και για να διασφαλιστεί η συνέπεια των δεδομένων, η μηχανή αναζήτησης στο σύνολό είναι αναγκαίο να διαθέτει μηχανισμούς για τον συντονισμό και την εξισορρόπηση φορτίου. Ακόμα επιβάλλεται να είναι ισχυρή όσον αφορά τις βλάβες του μηχανήματος, τις διακοπές δικτύου και τα διάφορα είδη σφαλμάτων.
- Οι αλλαγές στο περιεχόμενο του ιστού αλλά με διαφορετική συχνότητα ανάλογα με τον ιστότοπο και τη φύση του περιεχομένου. Ένας ανιχνευτή ιστού είναι αναγκαίο να μάθει αυτά τα πρότυπα ενημερώσεων για να εξασφαλίζει ότι το περιεχόμενο είναι λογικά τρέχον. Η απόκτηση της σωστής συχνότητας επανάληψης είναι δύσκολη: πολύ συχνή σημαίνει σπατάλη πόρων, αλλά όχι αρκετά συχνή συντελεί σε παλιό περιεχόμενο.
- Ο ιστός είναι γεμάτος από διπλά περιεχόμενα. Για παραδείγματα περιλαμβάνει πολλαπλά αντίγραφα ενός δημοφιλούς βιβλίου συνεδρίων, «καθρέφτες τοποθεσιών» που έχουν πρόσβαση συχνά, όπως η Wikipedia, και το περιεχόμενο του newswire που συχνά επαναλαμβάνεται. Το πρόβλημα επιδεινώνεται από το γεγονός ότι οι περισσότερες επαναλαμβανόμενες σελίδες δεν χαρακτηρίζονται ως ακριβείς αντίγραφα, αλλά σχεδόν διπλότυπα, ουσιαστικά η ίδια σελίδα αλλά με διαφορετικές διαφημίσεις, γραμμές πλοήγησης κ.α. Κατά τη διαδικασία ανίχνευσης είναι επιθυμητό να εντοπίζονται τα διπλότυπα και να επιλέγετε το καλύτερο υπόδειγμα για το χρήστη.
- Ο ιστός χαρακτηρίζεται ως πολύγλωσσος. Δεν υπάρχει εγγύηση ότι οι σελίδες σε μία γλώσσα συνδέονται μόνο με σελίδες στην ίδια γλώσσα. Για παράδειγμα, ένας καθηγητής στην Ασία μπορεί να διατηρεί τον ιστότοπό του στην μητρική του γλώσσα, αλλά να προσφέρει συνδέσμους προς δημοσιεύσεις στα αγγλικά. Ακόμα, αρκετές σελίδες περιλαμβάνουν ένα μείγμα κειμένου σε διαφορετικές γλώσσες. Από τη στιγμή που οι τεχνικές επεξεργασίας εγγράφων διαφέρουν στην βάση της από τη γλώσσα, είναι αναγκαίο να προσδιοριστεί η βασική γλώσσα σε μια σελίδα.

2.2.1 Ανεστραμμένοι δείκτες

Στη βασική του μορφή, ένας ανεστραμμένος δείκτης περιλαμβάνει καταλόγους καταχώρισης, που κάθε ένας συνδέεται με κάθε όρο που εμφανίζεται στη συλλογή. Η δομή ενός ανεστραμμένου δείκτη απεικονίζεται στο ακόλουθο σχήμα (σχήμα 2.1). Ένας κατάλογος αποσπάσεων αποτελείται από μεμονωμένες καταχωρίσεις, κάθε μία από τις οποίες αποτελείται από μια ταυτότητα εγγράφου και ένα μήνυμα ωφέλιμου φορτίου που προσδιορίζει την εμφάνιση του όρου στο έγγραφο. Το απλούστερο ωφέλιμο φορτίο είναι. . . τίποτα! Για απλή ανάκτηση boolean, δεν απαιτούνται επιπλέον πληροφορίες για την ανάρτηση εκτός από το id του εγγράφου, η ύπαρξη της ίδιας της απόσπασης προσδιορίζει την παρουσία του όρου στο έγγραφο.

Το πιο συνηθισμένο ωφέλιμο φορτίο, ωστόσο, είναι η συχνότητα των όρων (tf) ή ο αριθμός της πειθούς που εμφανίζεται στο έγγραφο. Τα πιο σύνθετα ωφέλιμα φορτία περιλαμβάνουν τις θέσεις κάθε εμφάνισης του όρου στο έγγραφο, για την υποστήριξη ερωτημάτων φράσης και βαθμολόγησης εγγράφων με βάση τη χρονική εγγύτητα, ιδιότητες του όρου, όπως εάν συνέβη στον τίτλο της σελίδας ή όχι, σχετικά με τις έννοιες σπουδαιότητας, ή ακόμα και τα αποτελέσματα της πρόσθετης γλωσσικής επεξεργασίας για παράδειγμα, υποδεικνύοντας ότι ο όρος είναι μέρος ενός ονόματος χώρου, για την υποστήριξη αναζητήσεων διευθύνσεων. Στο πλαίσιο του διαδικτύου, είναι χρήσιμος ο εμπλουτισμός της αναπαράστασης του περιεχομένου του εγγράφου, οι πληροφορίες κειμένου αντιπροσώπευσης (κείμενο που συνδέεται με υπερσυνδέσμους από άλλες σελίδες προς την εν λόγω σελίδα), αυτές οι πληροφορίες αποθηκεύονται συχνά και στο ευρετήριο. Στο παράδειγμα που φαίνεται στο σχήμα 2.1, βλέπουμε ότι:

ο όρος1 εμφανίζεται στα {d1, d5, d6, d11,. . .},

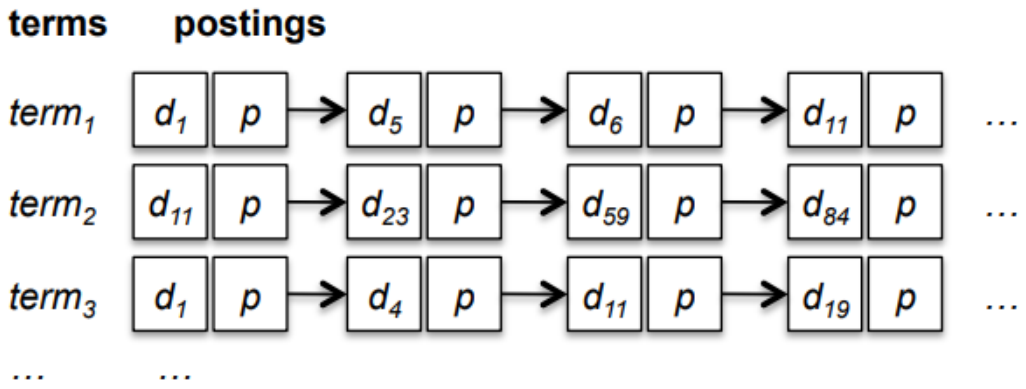
ο όρος2 εμφανίζεται στα {d11, d23, d59, d84,. . .}, και

ο όρος 3 εμφανίζεται στα {d1, d4, d11, d19,. . .}.

Σε μια πραγματική εφαρμογή, υπόθετε ότι τα έγγραφα είναι δυνατόν να αναγνωριστούν με ένα μοναδικό ακέραιο αριθμό που κυμαίνεται από 1 έως n, όπου n είναι ο συνολικός αριθμός των εγγράφων. Γενικά, οι καταχωρίσεις ταξινομούνται ανά

αναγνωριστικό εγγράφου, παρόλο που είναι επίσης δυνατές και άλλου είδους ταξινομήσεις.

Σχήμα 2.1: Απλή απεικόνιση ενός ανεστραμμένου δείκτη



πηγή: http://www.dcs.bbk.ac.uk/~dell/teaching/cc/book/ditp/ditp_ch4.pdf

Τα αναγνωριστικά του εγγράφου δεν έχουν εγγενή μεγάλη σημασία, παρόλο που η ανάθεση αριθμητικών αναγνωριστικών σε έγγραφα δεν χρειάζεται να είναι αυθαίρετη. Για παράδειγμα, οι σελίδες από τον ίδιο τομέα μπορεί να είναι αριθμημένες διαδοχικά. Εναλλακτικά, στις σελίδες που είναι υψηλότερες στην ποιότητα (με βάση, για παράδειγμα, στις τιμές PageRank) ενδέχεται να δοθούν μικρότερες αριθμητικές τιμές, ώστε να εμφανίζονται προς τα εμπρός σε μια λίστα καταχωρήσεων. Όπως και είναι τα πράγματα, μια δομή βοηθητικών δεδομένων είναι αναγκαία για να διατηρείται η χαρτογράφηση από τα αναγνωριστικά ακέραιου εγγράφου σε κάποια άλλη πιο χρήσιμη λαβή, όπως μια διεύθυνση URL.

Με την αποστολή ενός ερωτήματος από τον χρήστη, η ανάκτηση περιλαμβάνει τη λήψη λιστών καταλόγων που αφορούν τους όρους των ερωτημάτων και τη μετακίνηση των καταχωρίσεων για τον υπολογισμό του συνόλου αποτελεσμάτων. Στο απλούστερο (Boolean OR και τομή για boolean AND) στις λίστες αποσπασμάτων, οι οποίες μπορούν να υλοποιηθούν πολύ αποτελεσματικά αφού οι καταχωρίσεις ταξινομούνται με βάση την ταυτότητα του εγγράφου. Στη γενική περίπτωση, ωστόσο, πρέπει να υπολογίζονται οι βαθμολογίες του ερωτηματολογίου. Τα αποτελέσματα των μερικών εγγράφων αποθηκεύονται σε δομές

που ονομάζονται συσσωρευτές. Στο τέλος μόλις έχουν υποβληθεί σε επεξεργασία όλες οι καταχωρίσεις, τα έγγραφα κορυφής k εξάγονται ώστε να δώσουν μια ταξινομημένη λίστα αποτελεσμάτων για τον χρήστη. Φυσικά, υπάρχουν πολλές στρατηγικές βελτιστοποίησης για την αξιολόγηση του ερωτήματος (τόσο κατά προσέγγιση όσο και ακριβής) που μειώνουν τον αριθμό των καταχωρίσεων που πρέπει να εξετάσει ένας μηχανισμός ανάκτησης.

Το μέγεθος ενός ανεστραμμένου δείκτη ποικίλλει στην βάση του ωφέλιμου φορτίου που αποθηκεύεται σε κάθε απόσπαση. Εάν αποθηκευτεί μόνο η συχνότητα των όρων, ένας καλά βελτιστοποιημένος ανεστραμμένος δείκτης είναι δυνατόν να είναι το ένα δέκατο του μεγέθους της αρχικής συλλογής εγγράφων. Ένας ανεστραμμένος δείκτης που αποθηκεύει πληροφορίες θέσης θα ήταν εύκολα αρκετές φορές μεγαλύτερος από έναν που δεν το κάνει. Γενικά, είναι δυνατό να διατηρηθεί στη μνήμη ολόκληρο το λεξιλόγιο, ειδικά με την χρήση των τεχνικών όπως frontcoding. Ωστόσο, με εξαίρεση τις εμπορικές μηχανές websearch που διαθέτουν επαρκείς πόρους, οι 6 λίστες καταχωρήσεων είναι στις περισσότερες περιπτώσεις υπερβολικά μεγάλες για να αποθηκεύονται στη μνήμη και πρέπει να διαγράφονται από το δίσκο.

Σχήμα 2.2: Αλγόριθμος αντιστρέψιμης ευρετηρίασης αναφοράς

```
1: class Mapper
2:     procedure Map(docid n, doc d)
3:         H ← new AssociativeArray
4:         for all term t ∈ doc d do
5:             H{t} ← H{t} + 1
6:         for all term t ∈ H do
7:             Emit(term t, posting ⟨n, H{t}⟩)
1: class Reducer
2:     procedure Reduce(term t, postings [⟨n1, f1⟩, ⟨n2, f2⟩ . . .])
3:         P ← new List
4:         for all posting ⟨a, f⟩ ∈ postings [⟨n1, f1⟩, ⟨n2, f2⟩ . . .] do
5:             Append(P, ⟨a, f⟩)
6:         Sort(P)
7:         Emit(term t, postings P)
```

Πηγή: http://www.dcs.bbk.ac.uk/~dell/teaching/cc/book/ditp/ditp_ch4.pdf

Επομένως, η αξιολόγηση των ερωτημάτων συνεπάγεται απαραιτήτως με τυχαία πρόσβαση στο δίσκο και "αποκωδικοποίηση" των καταχωρήσεων. Μια σημαντική πτυχή του προβλήματος ανάκτησης είναι η οργάνωση εργασιών δίσκου έτσι ώστε να περιορίζονται πιθανές τυχαίες αναζητήσεις.

(http://www.dcs.bbk.ac.uk/~dell/teaching/cc/book/ditp/ditp_ch4.pdf).

2.3 Προοπτικό Ευρετήριο

Ένα προοπτικό ευρετήριο αποθηκεύει μια λίστα από διαφορετικές λέξεις που υπάρχουν σε οποιαδήποτε ιστοσελίδα, όπως διακρίνεται και στον ακόλουθο πίνακα (πινάκας 2.1). Η λογική πίσω από το προοπτικό ευρετήριο είναι ότι καθώς μια ιστοσελίδα επεξεργάζεται τους όρους τους οποίους περιλαμβάνει, είναι πιο εύκολο να αποθηκεύεται ο κάθε όρος απευθείας για κάθε ιστοσελίδα. Το στοιχείο αυτό παρέχει την δυνατότητα για την ασύγχρονη επεξεργασία των σελίδων. Ενώ παράλληλα ενημερώνει το κυρίως ευρετήριο της μηχανής αναζήτησης, πράγμα το οποίο επιλύει το πρόβλημα συνωστισμού που διακρίνεται όταν επιχειρείται παράλληλη επεξεργασία και ενημέρωση ενός ανεστραμμένου ευρετηρίου (Brinand Page, 1998).

Πινάκας2.1: Παράδειγμα προοπτικού ευρετηρίου

Ιστοσελίδες	Λέξεις-κλειδιά
URL_1	$KW_4, KW_2, KW_5, KW_7, KW_8, KW_{12}, KW_{15}$
URL_2	KW_2, KW_1, KW_3
URL_3	$KW_7, KW_9, KW_{10}, KW_{27}$

Πηγή: Μπάτζιος Α.,(2009),Εξόρυξη και διαχείριση δημασιολογικής πληροφορίας στον παγκόσμιο Ιστό, διδακτορική διατριβή, Θεσσαλονίκη 2009.

2.4 Κατάταξη αποτελεσμάτων (Ranking)

Οι μηχανές αναζήτησης κάνουν χρήση πολύπλοκων αλγορίθμων κατάταξης για να προσδιορίσουν τη σειρά με την οποία τα αποτελέσματα θα παρουσιαστούν στους χρήστες. Οι διαδικασίες αυτές στοχεύουν ώστε τα πρώτα αποτελέσματα να είναι τα πιο πρόσφατα, έγκυρα και αυτά που σχετίζονται περισσότερο με το ερώτημα. Αν βασιστούμε και στο γεγονός που παρουσιάστηκε σε προηγούμενη ενότητα ότι ένα πολύ μεγάλο ποσοστό χρηστών εξετάζει μόνο την πρώτη σελίδα των αποτελεσμάτων, διακρίνεται η σημαντικότητα αυτού του στοιχείου. Η τελική σειρά κατάταξης των αποτελεσμάτων δεν προσδιορίζεται από έναν και μοναδικό αλγόριθμο αλλά από ένα σύνολο αλγορίθμων, κάθε ένας από τους οποίους προσφέρει ένα ή περισσότερα σήματα που αφορούν την κατάταξη που πρέπει να έχει μια ιστοσελίδα για ένα ερώτημα.

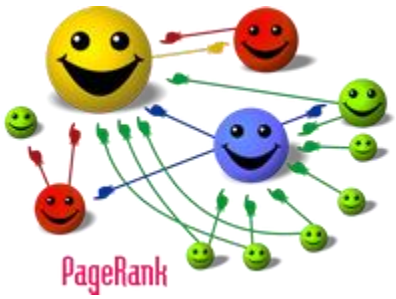
Σύμφωνα με επίσημες δηλώσεις της μηχανής αναζήτησηςGoogle, γίνεται χρήση περίπου 200 αναλογικών σημάτων για τον προσδιορισμό της σειράς κατάταξης των αποτελεσμάτων. Οι μεγάλες εμπορικές μηχανές αναζήτησης κρατούν τους ακριβείς αλγορίθμους και τις παραμέτρους τους ως εταιρικά μυστικά. Δεν σημαίνει ότι οι αλγόριθμοι αυτοί είναι εντελώς άγνωστοι καθώς οι πρώτες εκδόσεις τους ήταν ευρέως διαθέσιμες ως ερευνητικές ή ακαδημαϊκές εργασίες.

Έτσι, οι βασικότεροι αλγόριθμοι επάνω στους οποίους στηρίζονται οι μηχανές αναζήτησης για τον προσδιορισμό της σειράς κατάταξης των αποτελεσμάτων τους είναι οι PageRank, HITS, Hilltop και TrustRank, στου οποίους θα γίνει αναφορά ακολούθως.

2.4.1 PageRank

Το PageRank συναντάται και ακούγεται πολύ συχνά στο χώρο του Search Engine Optimization (SEO) είναι το PageRank με την συντόμευσή PR . Το Pagerank είναι μία αριθμητική τιμή, η οποία προσδιορίζει την σημαντικότητα και την σπουδαιότητα για μια ιστοσελίδα στο διαδίκτυο, στην βάση της αντικειμενικής κρίσης της μηχανής αναζήτησης Google. Το Pagerank έχει σημαντική επίδραση, σε συνδυασμό με κάποιους επιπλέον παράγοντες, στην κατάταξη και στα αποτελέσματα αναζήτησης. Για τον υπολογισμό και προσδιορισμό του Pagerank, η Google κάνει χρήση ενός πολύπλοκου αλγόριθμου με εκατομμύρια μεταβλητές και όρους και ο υπολογισμός του πραγματοποιείται με αδιάβλητη αυτοματοποιημένη διαδικασία.

Εικόνα 2.2: rage-raking στην google



Πηγή: http://www.webmasterslife.gr/search-engine-optimization/73.html

Βασική λογική της όμως στην αξιολόγηση είναι η σύνδεση όπως για παράδειγμα μιας Ιστοσελίδας Α με μία Ιστοσελίδα Β. Θεωρεί τον σύνδεσμο αυτό ότι αποτελεί μια θετική ψήφος για την Ιστοσελίδα Β από την Ιστοσελίδα Α**.** Συνεπώς οι διασυνδέσεις των ιστοσελίδων και εσωτερικά και εξωτερικά συνυπολογίζονται με θετικό τρόπο, διότι όσο πιο πολλές συνδέσεις έχει μία ιστοσελίδα από άλλες, καθώς και η σπουδαιότητά τους, θεωρούνται πολλές θετικές ψήφοι, άρα η ιστοσελίδα χαρακτηρίζετε σημαντική και σπουδαία. Εάν δηλαδή ο ψήφοι προς την ιστοσελίδα προκύπτουν από σημαντικές ιστοσελίδες μεγάλης σπουδαιότητας, η αξία η οποία προκύπτει είναι ακόμα μεγαλύτερη.

Εικόνα 2.2 : page raking στην google

Πηγή: http://www.quertime.com/article/67-google-pagerank-10-and-9-websites-you-must-know/

Η κλίμακα διαβάθμισης του Pagerank είναι από το 1 έως το 10 (PR1 - PR10). Όσο μεγαλύτερο Pagerank προσλαμβάνει μια ιστοσελίδα, τόσο υψηλότερη θέση λαμβάνει στην κατάταξη και στα αποτελέσματα αναζήτησης (SERP). Αυτός δεν είναι όμως ο μόνος παράγοντας στον οποίο κάνει χρήση η Google για να ταξινομήσει τις σελίδες, αλλά αποτελεί και αυτός σημαντικό στοιχειό.

Ακόμα ένας σημαντικός παράγοντας αξιολόγησης είναι το υλικό το οποίο υπάρχει στις σελίδες. Η Google αναλύει όχι μόνο τα meta tags αλλά συνολικότερα το περιεχόμενο της ιστοσελίδας. Συνυπολογίζοντας αρκετές παραμέτρους, όπως το μέγεθος και χρωματισμό των γραμματοσειρών, τους τίτλους, τις παραγράφους, τις λέξεις και τη συνάφειά τους με το κείμενο αλλά και το υπόλοιπο περιεχόμενο των άλλων ιστοσελίδων.

Συνηθίζεται να αναφέρεται ότι “Content is the king” για να γίνει σαφές ότι η μεγάλη σημασία του περιεχομένου αλλά και η ποιότητα του, στην αξιολόγηση μίας ιστοσελίδας από την Google αλλά και τις υπόλοιπες μηχανές αναζήτησης (http://www.webmasterslife.gr/search-engine-optimization/73.html).

2.4.2 Hyperlink-Induced Topic Search (HITS)

Η αναζήτηση η οποία προκύπτει μέσα από υπερσύνδεσμους (HITS) προσδιορίζεται από έναν αλγόριθμο ανάλυσης συνδέσμων ο οποίος συμβάλει στην αξιολόγηση ιστοσελίδων που είναι πιο διαδεδομένες ως Hubs οι οποίες αναπτύχθηκαν από τον Jon Kleinberg και αποτελούσαν τον πρόδρομο του PageRank.

Η ιδέα για το Hubs and Authorities προέκυψε από μια διορατικότητα στη κατασκευή των ιστοσελίδων με την έναρξη διαμόρφωσης του Διαδικτύου. Στην ουσία ορισμένες ιστοσελίδες, γνωστές ως κόμβοι, μεταχειριστήκαν ως σημαντικοί κατάλογοι που δεν ήταν στην πραγματικότητα έγκυροι στις πληροφορίες που παρείχαν, αλλά χρησιμοποιήθηκαν ως συλλογές ενός γενικότερου καταλόγου πληροφοριών που βοηθούσαν τους χρήστες άμεσα στο να οδηγούσαν άλλες αξιόπιστες σελίδες. Με άλλα λόγια, ένας καλός κόμβος αντιπροσώπευε μια σελίδα που «έδειχνε» πολλές άλλες σελίδες και μια καλή αρχή αντιπροσώπευε μια σελίδα που συνδέονταν με πολλούς διαφορετικούς κόμβους. Συμπερασματικά θα μπορούσαμε να πούμε δύο κύριες τιμές για μια σελίδα είναι:

1. Αρχή σελίδας, η οποία εκτιμά την αξία του περιεχομένου της σελίδας.
2. Τιμή διαύλου σελίδας, η οποία υπολογίζει την αξία των συνδέσεών της σε άλλες σελίδες.

Αρχικά ανακτά το σύνολο των αποτελεσμάτων στο ερώτημα αναζήτησης, έτσι ώστε ο υπολογισμός να εκτελείται μόνο σε αυτό το σύνολο αποτελεσμάτων και όχι σε όλες τις ιστοσελίδες.

Ο αλγόριθμος εκτελεί μια ροή επαναλήψεων, που προσδιορίζονται από δύο βασικά βήματα:

- **Ενημέρωση Αρχής:** Ενημέρωση της βαθμολογίας αρχής κάθε κόμβου ώστε να είναι ίση με το άθροισμα των απολεσθέντων βαθμών κάθε κόμβου που δείχνει σε αυτό. Δηλαδή, ένας κόμβος δέχεται έναν υψηλό βαθμό βαθμολογίας με τη σύνδεση με σελίδες που προσδιορίζονται ως Hubs για πληροφορίες.
- **Ενημέρωση διανομέα:** Ενημερώνεται η βαθμολογία κάθε κόμβου για να είναι ίση με το άθροισμα των Score of κάθε κόμβου που δείχνει. Δηλαδή, ένας

κόμβος λαμβάνει ένα υψηλό αποτέλεσμα κόμβου συνδέοντας τους κόμβους που χαρακτηρίζονται ως αρχές επί του θέματος.

Η βαθμολογία Hub και η βαθμολογία αρχής για έναν κόμβο προσδιορίζονται στην βάση του παρακάτω αλγόριθμου:

1. Εκκίνηση με κάθε κόμβο που διαθέτει βαθμολογία πλήθους και βαθμός εξουσίας.
2. Εκτέλεση του κανόνα ενημέρωσης αρχής.
3. Υλοποίηση του κανόνα
4. Αναβάθμιση Hub . Κανονικοποιήση των τιμών διαιρώντας κάθε βαθμολογία Hub με το άθροισμα των τετραγώνων όλων των βαθμολογιών Hub, και ακολούθως διαιρώντας κάθε βαθμολογία αρχής με το άθροισμα των τετραγώνων όλων των αποτελεσμάτων της αρχής.
5. Επαναλάβετε από το δεύτερο βήμα όπως είναι απαραίτητο.

Στην συνέχεια υπάρχει μια σύγκριση μεταξύ αλγορίθμων αναζήτησης σελίδας και αλγορίθμων επαγόμενων από την υπερσύνδεση (HITS) Υπερσύνδεση που προκαλείται από το θέμα:

Το HITS στηρίζεται σε δύο τιμές ποιότητας των "Ενημέρωση Αρχής" και "Ενημέρωση Hub". Η ενημερωμένη έκδοση της αρχής υπολογίζεται από το σύνολο των συνδέσμων κόμβων που επικοινωνούν με τον ιστότοπο αρχής και η ενημέρωση Hub υπολογίζεται από τον αριθμό των ιστοτόπων αρχής που συνδέονται με τον ιστότοπο του Hub. Το συνολικό αποτέλεσμα HITS στηρίζεται στη σχέση μεταξύ αυτών των δύο τιμών. Στην πραγματικότητα υπολογίζει δύο βαθμολογίες ανά έγγραφο. Το HITS λειτουργεί σε μικρά υποσυστήματα που αντιπροσωπεύουν μια σύνδεση μεταξύ των ιστοτόπων Hub και Authority. Στο HITS, η αύξηση του βάρους αρχής διογκώνει το βάρος του κόμβου των τοποθεσιών. Ακόμα το HITS υπολογίζει το σκορ χωρίς την ευρετηρίαση. Επίσης το HITS διαθέτει μια ειδική χρήση σε ιστοσελίδες με σχετική ανάλυση συγκεκριμένα.

PageRank:

Το PageRank βασίζεται σε ένα σύνολο διαφόρων παραγόντων, και ιδίως στον αριθμό των συνδέσμων ποιότητας. Οι σύνδεσμοι ποιότητας προσδιορίζονται ως εκείνοι οι σύνδεσμοι που σχετίζονται με την εξειδικευμένη τοποθεσία του ιστότοπου και τοποθετούνται σε ιστότοπους υψηλής επισκεψιμότητας. Έτσι το PageRank λαμβάνει υπόψη του κυρίως μία βαθμολογία ανά έγγραφο. Η σελίδα Rank λειτουργεί σε ένα μεγάλο web Graph που στηρίζεται σε όλους τους προηγουμένους συνδέσμους και παράγοντες συνάφειας. Στο RankPage, ο σύνδεσμος προηγούμενης ποιότητας στην υψηλή ιστοσελίδα PR αυξάνει την τάξη της ιστοσελίδας. Επίσης το PageRank υπολογίζει τη βαθμολογία μετά τη διαδικασία ευρετηρίασης. Ακόμα το PageRank μπορεί να κάνει χρήση πολλών παρόντων, όπως ο βαθμός Street (κατάταξη των τόπων εκτός από τους ιστότοπους με βάση τις επισκέψεις του πληθυσμού). Ομοίως, η κατάταξη PageRank μπορεί να χρησιμοποιηθεί από πολλαπλά περιβάλλοντα από ινστιτούτα έως ανιχνευτές μηχανών αναζήτησης.

Τόσο το HITS όσο και το RankPage διαθέτουν το πλεονέκτημα όπως το να μπορούν να εφαρμόσουν διαφορετικά σενάρια. Η σελίδα Rank είναι πιο δημοφιλής λόγω του ότι διαθέτει την δυνατότητα χρήσης σε πολλαπλά περιβάλλοντα διαφορετικά από την αναζήτηση ιστού. Το HITS είναι αρκετά χρήσιμο λόγω της ιδιαίτερης εστίασής του στην κατηγοριοποίηση ιστοσελίδων Hub και αρχών (http://complexnt.blogspot.gr/2012/04/hyperlink-induced-topic-search-hits.html).

2.5 Τα είδη των μηχανών αναζήτησης

Οι διαφοροποιήσεις στις μηχανές αναζήτησης είναι πολλές, μια από αυτές προσδιορίζεται στην βάση των αποτελεσμάτων που προσφέρουν στο χρηστή τους. Το στοιχείο αυτό μας προσφέρει την δυνατότητα του διαχωρισμού σε μηχανές αναζήτησης γενικού ενδιαφέροντος και σε αυτές που καλούνται ως στοχευόμενες. Σαν βασική τους διαφορά περιγράφεται, το στοιχείο ότι οι στοχευόμενες μηχανές αναζήτησης, εστιάζουν σε ειδικευμένα ζητήματα. Διενεργούν μια προσπάθεια στο να εντοπίζουν, να βρουν και να καταγράψουν όσες περισσότερες ιστοσελίδες μπορούν για μια συγκεκριμένη θεματική, επισκεπτόμενες ένα περιορισμένο αριθμό δικτυακών τόπων που περιλαμβάνουν το συγκεκριμένο ζήτημα. Εν' αντίθεση με αυτές που είναι γενικού ενδιαφέροντος, οι οποίες διενεργούν μια προσπάθεια να καταγράψουν όσο το δυνατόν μεγαλύτερο τμήμα των ιστοσελίδων του διαδικτύου, άσχετος θεματικής.
Επίσης παρά το γεγονός ότι οι μηχανές αναζήτησης δεν έχουν πολλά χρόνια λειτουργίας από την πρώτη τους εμφάνιση, η διαρκής ανάπτυξη της τεχνολογίας, αλλά και ο ανταγωνισμός που υπάρχει στο κλάδο, έχουν συμβάλει στην διαφοροποίηση σε ακόμα δυο κατηγορίες. Έτσι έχουμε τις μηχανές αναζήτησης πρώτης και δεύτερης γενιάς. Πρώτης γενιάς, είναι αυτές που συσχετίζουν και παρουσιάζουν τα αποτελέσματα, με βάση το ποσοστό συνάφειας τους, ενώ δεύτερης γενιάς είναι οι μηχανές που έχουν την δυνατότητα εμφάνισης και ιεράρχησης των αποτελεσμάτων με ποικίλους τρόπους, όπως είναι η ιεράρχηση των αποτελεσμάτων σύμφωνα με την δημοτικότητα τους, σύμφωνα με το είδος ή τον τύπο των τεκμηρίων, η ακόμα και να δεχτούν ερωτήσεις σε φυσική γλώσσα και να παράγουν αποτελέσματα που έχουν προσδιοριστεί σε προγενέστερο χρόνο.

Ακόμη μια άλλη, ξεχωριστή κατηγορία στις μηχανές αναζήτησης, αποτελούν οι Μετά-μηχανές. Αυτές σε σύγκριση με τις κανονικές, κάνουν χρήση του λογισμικού αράχνης για την κατασκευή της βάσης δεδομένων τους, οι μετά-μηχανές δεν διαθέτουν δικό τους ευρετήριο, αλλά λαμβάνουν τα αποτελέσματα τους από τα ευρετήρια άλλων μηχανών αναζήτησης. Έτσι σε κάθε αναζήτηση που πραγματοποιείται, στέλνουν τις λέξεις-κλειδιά ταυτοχρόνως σε μια σειρά προκαθορισμένων υπηρεσιών αναζήτησης. Ο μηχανισμός αναζήτησης παραμένει λίγο χρόνο στο ευρετήριο κάθε βάσης και επιστρέφει ένα σύνολο αποτελεσμάτων από κάθε βάση.

Asnicar F. and Tasso C. June 1997. ifWeb: A Protoype of User Model-Based Intelligent Agent for Documentation Filtering and Navigation in the World Wide Web. In Proceedings of the 6th International Conference on User Modeling.

Η ιδέα της μετά-μηχανής μπορεί να είναι πάρα πολύ καλή, όμως η υλοποίηση της δεν προσφέρεται πάντοτε τα προσδοκώμενα αποτελέσματα. Αν και υπάρχει εξοικονόμηση χρόνου μέσα από την χρήση των εν' λόγω υπηρεσιών, παρόλα αυτά έχει αποδειχθεί σε αρκετές περιπτώσεις ότι τα αποτελέσματα δεν ικανοποιούν πάντοτε τους χρήστες. Αυτό πηγάζει από το γεγονός ότι επιστρέφουν ένα συγκεκριμένο ποσοστό αποτελεσμάτων από κάθε μηχανή αναζήτησης, με αποτέλεσμα αυτό να μην πραγματοποιείται με τον όρθρο τρόπο, η επιλογή της πληροφορίας που είναι αναγκαίο να αντληθεί.

2.5.1 Βασικές λειτουργίες των μηχανών αναζήτησης

Ασφαλώς και διακρίνονται διαφορές στο τρόπο λειτουργίας του συνόλου των μηχανών αναζήτησης που υπάρχουν, με βάση το είδος τους, όμως όλες τους διενεργούν τρεις βασικές λειτουργίες :

1. Αναζητούν και συλλέγουν συγκεκριμένες ιστοσελίδες του διαδικτύου, με κύριο κριτήριο σημαντικές σημασιολογικά λέξεις- εκφράσεις.
2. Παράγουν και διατηρούν ένα ευρετήριο με τις λέξεις και την τοποθεσία που τις εντόπισαν (URL).
3. Δίνουν την δυνατότητα στους χρήστες να αναζητήσουν λέξεις, ή συνδυασμό λέξεων στο ευρετήριο της μηχανής αναζήτησης.

(Barzilay, 1997).

2.6 Λεξική Αλυσίδα

Η συνάφεια παράγεται στο κείμενο από την κατάλληλη επιλογή λέξεων. Η λεξική συνάφεια προκύπτει μέσα από την χρησιμοποίηση των λέξεων, οι οποίες με κάποιο τρόπο συνδέονται με λέξεις στις οποίες έγινε χρήση νωρίτερα.

Λεξική Αλυσίδα

Οι Halliday και Hasan ταξινόμησαν την λεξική συνάφεια σε δύο κύριες κατηγορίες - *επανάληψη (reiteration) και παράθεση (collocation).*Η επανάληψη προκύπτει όταν ένα στοιχείο του κειμένου , φέρνει στο μυαλό την σημασία ενός προαναφερθέντος στοιχείου. Η επανάληψη προκύπτει μέσα από την χρησιμοποίηση επαναλήψεων λέξεων, συνωνύμων και υπονύμων. Για παράδειγμα, η λέξη *"μηχανή"* επαναλαμβάνεται τρεις φορές. Η σχέση συνωνυμίας φαίνεται από το ζευγάρι *"μικροϋπολογιστής"* και *"προσωπικός υπολογιστής"*. Όμοια η λέξη *"εξοπλισμός"'* είναι πιο γενική από την *"μηχανή"* (σχέση υπονυμίας).

Η παράθεση αναφέρεται σε λέξεις, οι οποίες τείνουν να εμφανίζονται στο κείμενο. Σε αυτές υπάρχει η δυνατότητα να διαχωριστούν σε δύο κατηγορίες: συστηματικές και μη-συστηματικές. Η συστηματική σημασιολογική σχέση δημιουργείται από τις σχέσεις μετωνυμίας και αντωνυμίας. Οι μη συστηματικές σχέσεις είναι δύσκολο να προσδιοριστούν. Τέτοιες παραθέσεις περιγράφουν ζητήματα, που τείνουν να εμφανίζονται μαζί σε παρόμοιες καταστάσεις ή περιβάλλοντα στον πραγματικό κόσμο.

Κυριαρχία της Λεξικής Συνάφειας

Οι Halliday και Hasan έλεγξαν την συχνότητα των διαφορετικών τύπων συνάφειας με κείμενα ποικίλα σε ύφος. Τα αποτελέσματα που πρόεκυψαν, αναφέρουν ότι η λεξική συνάφεια είναι η κυριαρχούσα κατηγορία - κατέχει πάνω από το 40% των μηχανισμών συνάφειας. Αυτά τα συμπεράσματα υποστηρίζονται από τον Hoey, ο οποίος διενέργησε ερευνητικές διεργασίες σε εφτά διαφορετικούς τύπους κειμένου. Ο Hoey ισχυρίζεται ότι η λεξική συνάφεια είναι ο κρίσιμος παράγοντας, ο οποίος προσδίδει σε ένα κείμενο το χαρακτηρισμό κείμενο ευνόητο. Ο ισχυρισμός, ότι η ανάλυση συνάφειας μόνο με την χρήση της λεξικής συνάφειας είναι πιθανόν να

επιφέρει σημαντικά αποτελέσματα βασίζεται σε αυτά τα αποτελέσματα από τους Halliday, Hasan και Hoey.

Τέλος ως λεξική αλυσίδα καλείται η λεξική συνάφεια που πιθανόν να υπάρχει όχι μόνο μεταξύ δύο όρων, αλλά και μεταξύ ακολουθιών από σχετικές λέξεις - το τελευταίο καλείται λεξική αλυσίδα. Οι λεξικές αλυσίδες είναι δυνατόν να επιμηκύνονται μεταξύ προτάσεων και διαφορετικών τμημάτων κειμένου. Ο όρος εισήχθη από τους Halliday και Hasan το 1976 (Halliday and Hasan,1976) και επεκτάθηκε στην μετέπειτα εργασία του Hasan (Hasan, 1984).

2.7 Αυτόματη Ανάθεση Κειμένων στη Θεματική Ιεραρχία

Ακολούθως περιγράφεται, ο τρόπος με τον οποίο γίνεται χρήση στο υφιστάμενο εκτεταμένο σημασιολογικό δίκτυο για την αυτόματη ανάθεση των ιστοσελίδων στις αντίστοιχες θεματικές ενότητες. Η διαδικασία που συντελείται στο μοντέλο κατηγοριοποίησης για την οργάνωση των ιστοσελίδων σε θεματικές ενότητες ακολουθεί τις παρακάτω φάσεις. Δεδομένου ενός συνόλου κειμένων, αρχικά επεξεργάζεται το κάθε κείμενο και εφαρμόζεται με το σημασιολογικό δίκτυο που αναπτύσσεται για να εξήχθη από το περιεχόμενο τις λέξεις αυτές που μεταφέρουν πληροφορία σχετικά με τη θεματολογία της.

Ακολούθως προσδιορίζονται οι σημασιολογικές σχέσεις του δικτύου μας και συνδέονται οι θεματικές λέξεις κάθε ιστοσελίδας σε μια αντίστοιχη λεξική αλυσίδα. Στην συνέχεια, εφαρμόζεται το μοντέλο κατηγοριοποίησης στις λεξικές αλυσίδες των αντίστοιχων ιστοσελίδων διαδικασία κατά την οποία το μοντέλο κατηγοριοποίησης αντιστοιχίζει τις έννοιες που αντιπροσωπεύουν οι θεματικές λέξεις των αλυσίδων στις έννοιες της οντολογίας, και υπολογίζει αλγοριθμικά τη θεματική κατηγορία της οντολογίας που αντιπροσωπεύει πληρέστερα το περιεχόμενο κάθε ιστοσελίδας. Τέλος, ο αλγόριθμος θεματικής κατηγοριοποίησης, αναθέτει κάθε κείμενο στην αντίστοιχη θεματική κατηγορία της οντολογίας, της οποίας οι έννοιες καλύπτουν καλυτέρα το περιεχόμενο του ανάλογου κειμένου. Στην συνέχεια περιγράφονται, τα βήματα που ακολουθεί ο αλγόριθμος θεματικής κατηγοριοποίησης:

- Εξόρυξη θεματικών λέξεων από τα κείμενα.
- Κατασκευή λεξικών αλυσίδων για τις θεματικές λέξεις του κάθε κειμένου.
- Αποσαφήνιση των λέξεων της κάθε λεξικής αλυσίδας.
- Αντιστοίχιση των θεματικών λέξεων της κάθε λεξικής αλυσίδας στις έννοιες της οντολογίας.
- Υπολογισμός του βαθμού συσχέτισης ανάμεσα στις θεματικές λέξεις κάθε κειμένου και των θεματικών κατηγοριών της οντολογίας.
- Κατηγοριοποίηση του κάθε κειμένου στην κατηγορία με το μέγιστο βαθμό συσχέτισης.

Εξόρυξη και Αποσαφήνιση Θεματικών Όρων

Η βασική θεώρηση στην προσέγγιση που υιοθετείται για την κατηγοριοποίηση κειμένων σε μια οντολογία θεματικών κατηγοριών, είναι πως ο υπολογισμός της θεματικής συσχέτισης ενός κειμένου με κάποια κατηγορία της οντολογίας, σχετίζεται με τη λεξική συνοχή (*lexical cohesion*) στο κείμενο, δηλ., από το αν σημαντικός αριθμός λέξεων του κειμένου σχετίζεται με την ίδια θεματική κατηγορία. Για να αποτυπωθεί η προαναφερθείσα ιδιότητα στις ιστοσελίδες του Παγκόσμιου Ιστού αλλά και γενικότερα στα κείμενα γενικότερα, υιοθετήθηκε η τεχνική των λεξικών αλυσίδων (MorrisandHirst, 1991, HirstandSt-Onge, 1998, Barzilay, 1997) και πως κατασκευάστηκε το περιεχόμενο της κάθε ιστοσελίδας μια αλληλουχία σημασιολογικά συσχετισμένων λέξεων, επονομαζόμενη ως λεξική αλυσίδα (*lexical chain*). *Μια λεξική αλυσίδα είναι μια ακολουθία συσχετιζόμενων λέξεων, οι οποίες ενδέχεται να εκτείνονται μέσα στο κείμενο σε αποστάσεις τόσο μικρές μεταξύ τους όσο και η απόσταση γειτονικών λέξεων, ή τόσο μεγάλες ώστε να καλύπτουν ολόκληρο το κείμενο.*

Κάθε λεξική αλυσίδα διακρίνεται από πλήρη ανεξαρτησία από τη γραμματική δομή του κειμένου που αναπαριστά, με αποτέλεσμα η αλυσίδα να αποτελεί ουσιαστικά μια λίστα σημασιολογικά συσχετιζόμενων λέξεων, οι οποίες περιγράφουν τμήμα της συνοχικής δομής ενός κειμένου Stairmand and Black, 19969.

Το υπολογιστικό μοντέλο που υιοθετήθηκε για την κατασκευή των αντίστοιχων λεξικών αλυσίδων για το περιεχόμενο των κειμένων είναι παρόμοιο με αυτό που έκαναν χρήση οι (Barzilay and Elhadad, 1999) για την αναπαράσταση κειμένων σε λεξικές αλυσίδες. Ειδικότερα, για την αναπαράσταση ενός κειμένου υπό τη μορφή λεξικής αλυσίδας, πρωτεύων προϋπόθεση είναι η μορφολογική επεξεργασία του κειμένου, με στόχο να προσδιοριστούν μορφολογικά οι λέξεις που περιέχονται όντος αυτού. Ο μορφολογικός χαρακτηρισμός των λέξεων ενός κειμένου σχετίζεται ουσιαστικά στην ανάθεση ετικετών στις λέξεις του κειμένου, οι οποίες προσφέρουν πληροφορία σχετικά με το μέρος του λόγου στο οποίο ανήκει κάθε λέξη. Στη συνέχεια, ακολουθεί η αναπαράσταση του κειμένου υπό τη μορφή λεξικής αλυσίδας, διαδικασία η οποία καταγράφει τρία παρακάτω βήματα:

- Επιλογή ενός συνόλου υποψήφιων θεματικών όρων από ένα κείμενο.
- Αποσαφήνιση των υποψήφιων όρων και διαπίστωση των σημασιολογικών συσχετίσεων μεταξύ του κάθε υποψήφιου όρου και

στους υπόλοιπους υποψήφιους θεματικούς όρους, διαμέσου ενός σημασιολογικού δικτύου.

- Σε περίπτωση που διαπιστωθεί η ύπαρξη σημασιολογικής σχέσης ανάμεσα στους δύο υποψήφιους όρους, η τοποθέτηση αυτών σε μια κοινή ομάδα λέξεων και δήλωση της σχέσης που τις συσχετίζει.

Επιλογή Θεματικών Λέξεων

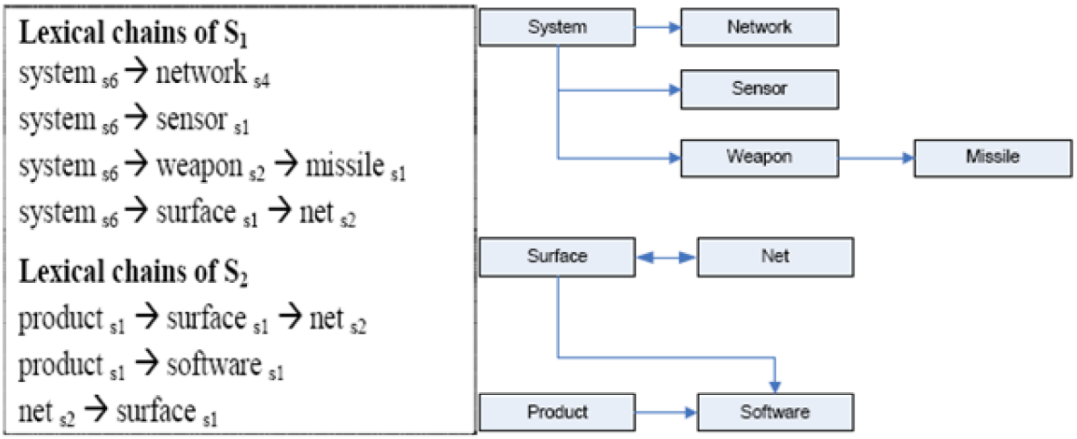
Λαμβάνοντας υπόψη πως το μεγαλύτερο μέρος της πληροφορίας αναφορικά με το θεματικό περιεχόμενο ενός κειμένου μεταφέρεται μέσω των ουσιαστικών που περιέχονται σε αυτό [Kazmanetal., 1996], επιλέγουμε τις υποψήφιες θεματικές λέξεις για κάθε κείμενο, εξετάζοντας μόνο τα ουσιαστικά του κειμένου και αγνοούμε κατά την επιλογή μας λέξεις που ανήκουν σε άλλες γραμματικές κατηγορίες. Πιο συγκεκριμένα, αρχικά επιλέγουμε τις υποψήφιες θεματικές λέξεις κάθε κειμένου από τις λέξεις που είναι μορφολογικά χαρακτηρισμένες με τις ετικέτες *Noun* (ουσιαστικό) ή *ProperNoun* (κύριο όνομα). Στη συνέχεια, προχωρούμε στη μορφολογική κανονικοποίηση (lemmatization) των υποψήφιων λέξεων, διαδικασία κατά την οποία καθεμιά από τις υποψήφιες λέξεις ανάγεται στον πρώτο κλιτικό της τύπο, που για την περίπτωση των ουσιαστικών είναι η ονομαστική ενικού. Κατ' αυτόν τον τρόπο καταλέγεται να επιλέγεται από κάθε κείμενο ένα σύνολο λέξεων (δηλ., τα ουσιαστικά και τα κύρια ονόματα του κειμένου), οι οποίες είναι υποψήφιες να αποτελέσουν τις θεματικές λέξεις της αντίστοιχης αλυσίδας για το κείμενο.

Αποσαφήνιση και Σύνδεση των Θεματικών Λέξεων σε Αλυσίδες

Ακολούθως γίνεται η αρχικοποίηση, μιας αλυσίδα για κάθε υποψήφια θεματική λέξη και αναζητούνται οι τελευταίες σ' ένα σημασιολογικό δίκτυο λημμάτων. Ειδικότερα, αναζητούμε την ύπαρξη κάθε υποψήφιας λέξης μεταξύ των λέξεων του σημασιολογικού δικτύου. Η σειρά με την οποία αναζητούνται οι υποψήφιες λέξεις στο σημασιολογικό δίκτυο ακολουθεί τη σειρά με την οποία γίνεται η εμφάνιση των αντίστοιχων λέξεων στο υπό μελέτη κείμενο. Σε περίπτωση που διαπιστωθεί η ύπαρξη μιας υποψήφιας λέξης μεταξύ του δικτύου, μελετάται η πιθανότητα αυτή να σχετίζεται σημασιολογικά με την αμέσως προηγούμενη λέξη του ίδιου κειμένου που εντοπίστηκε στο σημασιολογικό δίκτυο.

Οι διαδοχικές λέξεις που σχετίζονται μεταξύ τους σημασιολογικά στο δίκτυο λημμάτων, τοποθετούνται στην ίδια ομάδα (αλυσίδα) λέξεων. Η σειρά με την οποία ενσωματώνονται οι λέξεις σε κάποια ομάδα, ακολουθεί τη σειρά με την οποία εξετάζονται οι λέξεις στο σημασιολογικό δίκτυο. Ακόμα, κάθε λέξη μέσα σε μια ομάδα σχετίζεται αμέσως με την αμέσως προηγούμενη λέξη της ομάδας, ακολουθώντας τη σχέση που αφορά τις υφιστάμενες λέξεις στο σημασιολογικό δίκτυο. Στο ακόλουθο σχήμα παρουσιάζεται ένα παράδειγμα οργάνωσης των ουσιαστικών: *system, network,sensor, weapon, missile, surface, net, software* ενός κειμένου σε λεξικές αλυσίδες (αριστερά στο σχήμα), όπου η έννοια της κάθε λέξης στο σημασιολογικό δίκτυο υποδηλώνεται από την ετικέτα.

Σχήμα 2.1. Παράδειγμα λεξικής αλυσίδας



Οι σημασιολογικές σχέσεις του δικτύου που μελετώνται στο διάστημα δημιουργίας των αλυσίδων για τις λέξεις ενός κειμένου, είναι οι σχέσεις συνωνυμία, μερωνυμίας, ολωνυμίας, υπωνυμίας, υπερωνυμίας, αντωνυμίας και η σχέση επανάληψης. Οι διαδοχικές λέξεις του κειμένου που σχετίζονται μεταξύ τους στο δίκτυο λημμάτων τοποθετούνται στην ίδια ομάδα λέξεων, δημιουργώντας κατ' αυτόν τον τρόπο μια λεξική αλυσίδα.

Η επιλογή των σχέσεων συνωνυμίας, μερωνυμίας, ολωνυμίας, υπωνυμίας, υπερωνυμίας και αντωνυμίας από το σύνολο των σημασιολογικών σχέσεων που κωδικοποιούνται στο δίκτυο λημμάτων, βασίζεται στα αποτελέσματα των Morris και Hirst (1991), οι οποίοι προσδιόρισαν ότι η λεξική συνοχή ενός κειμένου περιγράφεται

από την ύπαρξη μιας εκ των προαναφερθέντων σχέσεων ανάμεσα στις λέξεις του κειμένου.

Ακόμα, η επιλογή της σχέσης επανάληψης για τη σύνδεση επαναλαμβανόμενων λέξεων βασίζεται στα συμπεράσματα των Barzilay και Elhadad ,1999), οι οποίοι έδειξαν πως η πιο ισχυρή ένδειξη της συνεκτικότητας στο περιεχόμενο ενός κειμένου είναι η επανάληψη των ίδιων λέξεων εντός αυτού.

Δεδομένης της πολυσημίας που περιγράφει τα στοιχεία των τις λέξεων των φυσικών γλωσσών, βασικός παράγοντας για την αποτελεσματική αναπαράσταση ενός κειμένου υπό τη μορφή μια λεξικής αλυσίδας, αποτελεί η προγενέστερη αποσαφήνιση των υποψήφιων θεματικών λέξεων του κειμένου. Η αποσαφήνιση των υποψήφιων θεματικών λέξεων προσδιορίζεται ως βασική αναγκαία, εφόσον η ίδια λέξη είναι δυνατόν να εμφανίζεται παραπάνω από μία φορές στο σημασιολογικό δίκτυο. Στο σημείο αυτό κρίνεται αναγκαίο να επισημανθεί πως η οργάνωση των λημμάτων του σημασιολογικού δικτύου προκύπτει με βάση την έννοια που το καθένα από αυτά εκπροσωπεί.

Άρα, κάθε λέξη εμφανίζεται τόσες φορές εντός του σημασιολογικού δίκτυο, όσες και οι διαφορετικές σημασίες που είναι δυνατόν να έχει στην αντίστοιχη φυσική γλώσσα. Με παρόμοιο τρόπο, το είδος της σημασιολογικής σχέσης που σχετίζει κάποια λέξη με τις υπόλοιπες λέξεις του δικτύου προσδιορίζεται τόσο από τη σημασία της ίδιας της λέξης όσο και από τις σημασίες των υπολοίπων λέξεων του δικτύου.

Αλγόριθμος Θεματικής Κατηγοριοποίησης

Για την ανάθεση ιστοσελίδων στις αντίστοιχες θεματικές κατηγορίες της οντολογίας, ο αλγόριθμος θεματικής κατηγοριοποίησης συλλέγει τις θεματικές λέξεις στις αλυσίδες κάθε ιστοσελίδας και τις αναζητά στις έννοιες της οντολογίας. Ειδικότερα, ο αλγόριθμος θεματικής κατηγοριοποίησης κάνει προσπάθεια να αντιστοιχίσει κάθε θεματική λέξη μιας αλυσίδας στον αντίστοιχο κόμβο της οντολογίας που αντιπροσωπεύει την έννοια της υφιστάμενης θεματικής λέξης. Όπως αναφέρθηκε και παραπάνω σε προηγούμενες παραγράφους, οι θεματικές λέξεις στις αλυσίδες ενός κειμένου είναι αποσαφηνισμένες, εξασφαλίζοντας με αυτόν τον τρόπο έτσι πως οποιαδήποτε θεματική λέξη αντιστοιχίζεται σε έναν κόμβο της οντολογίας.

Έχοντας προσδιορίσει τις έννοιες της οντολογίας που αντιστοιχίζονται στις θεματικές λέξεις της αλυσίδας ενός κειμένου, ο αλγόριθμος θεματικής κατηγοριοποίησης προελαύνει τις ιεραρχίες των παραπάνω κόμβων, ακολουθώντας τις σχέσεις υπερωνυμίας των τελευταίων. Η παραπάνω διαδικασία υλοποιείται επαναληπτικά και τελειώνει όταν ο αλγόριθμος θεματικής κατηγοριοποίησης συναντήσει για κάθε θεματική λέξη έναν κόμβο (έννοια) που να είναι επισημειωμένος με μια ειδική θεματική κατηγορία της οντολογίας, κατηγορίες τις οποίες επιστρέφει στο μοντέλο κατηγοριοποίησης.

Για κείμενα με πολύ εξειδικευμένη θεματολογία, είναι δυνατόν ο αλγόριθμος θεματικής κατηγοριοποίησης να εντοπίζει μία και μοναδική ειδική κατηγορία στην οντολογία, την οποία και επιστρέφει στο μοντέλο κατηγοριοποίησης. Παρόλα αυτά, δεδομένης αφενός της σποραδικότητας του περιεχομένου του Παγκόσμιου Ιστού και αφετέρου του πλήθους των υποκατηγοριών της οντολογίας μας, είναι πιθανόν να μην αντιπροσωπεύουν όλες οι θεματικές λέξεις μιας αλυσίδας μία κοινή ειδική κατηγορία. Σ’ αυτή την περίπτωση, ο αλγόριθμος θεματικής κατηγοριοποίησης επιστρέφει πίσω στο μοντέλο όλες τις υποκατηγορίες της οντολογίας, στις οποίες αντιστοιχούν οι θεματικές λέξεις μιας αλυσίδας. Για να διαλέξει ο αλγόριθμος κατηγοριοποίησης μεταξύ πολλαπλών θεματικών κατηγοριών, την κατηγορία εκείνη που ανταποκρίνεται με μεγαλύτερη ακρίβεια στο περιεχόμενο ενός κειμένου συνολικά, κάνει χρήση μιας μέτρησης που προσδιορίστηκε για να περιγράψει η συσχέτιση των κειμένων με κάθε θεματική κατηγορία.

ΚΕΦΑΛΑΙΟ 3ο: Εξέλιξη διαδικτύου και αναζήτησης

3.1 Σημασιολογικές μηχανές αναζήτησης

3.1.1 Ακαδημαϊκές Βιβλιοθήκες και σημασιολογική αναζήτηση

Οι ακαδημαϊκές βιβλιοθήκες αποτελούν κύρια πηγή οργανωμένης γνώσης, ενώ ταυτόχρονα υποστηρίζουν το εκπαιδευτικό και ερευνητικό έργο που πραγματοποιείται στα ακαδημαϊκά Ιδρύματα. Οι βασικές υπηρεσίες που παρέχονται στις ακαδημαϊκές βιβλιοθήκες ακαδημαϊκών βιβλιοθηκών είναι: ο δανεισμός, ο διαδανεισμός, η πρόσκτηση υλικού, η καταλογογράφηση, η διακίνηση, η αποθήκευση και προστασία του υλικού, η αναζήτηση, η εύρεση, καθώς και η δυνατότητα πρόσβασης και διάθεσης του πληροφοριακού υλικού. Την παρούσα χρονική στιγμή, το πληροφοριακό υλικό των ακαδημαϊκών βιβλιοθηκών δεν υπάρχει μόνο σε έντυπη, άλλα και σε ψηφιακή μορφή. Η ραγδαία εξέλιξη της Πληροφορικής και των Επικοινωνιών έχει συντελέσει σε τεράστιες μεταβολές στη μορφή και στο υλικό των ακαδημαϊκών βιβλιοθηκών. Οι ψηφιακές ακαδημαϊκές βιβλιοθήκες προσφέρουν μέσα από το Διαδίκτυο τη δυνατότητα πρόσβασης και ανάκτησης σημαντικού όγκου πληροφοριών. Η τεχνολογική αυτή εξέλιξη έχει σημαντική επίδραση και παρουσιάζεται ως σημαντική πρόκληση και προβλήματα που σχετίζονται με την περιγραφή, ταξινόμηση, οργάνωση, αξιοπιστία και εμπιστευτικότητα των προσφερόμενων υπηρεσιών πληροφοριών. Σαφώς, η ετερογένεια των δεδομένων είναι το βασικότερο πρόβλημα, διότι τα διάφορα πληροφοριακά συστήματα των βιβλιοθηκών κάνουν χρήση διαφορετικών μεταδεδομένων για την αναπαράσταση των βιβλιογραφικών τους δεδομένων (SDT JP, 1999), (http://www-digUb.stanford.edu/Jestbed/doc2/SDLIP/).

Τα μεταδεδομένα είναι απαραίτητα για την ορθή περιγραφή, οργάνωση, αποθήκευση, αναζήτηση, εύρεση και ανάκτηση των προσφερόμενων πληροφοριών, διότι οι πληροφορίες στο Διαδίκτυο προσδιορίζονται ως ανομοιογενείς. Σε πολλές ακαδημαϊκές βιβλιοθήκες, χρήσιμες ακαδημαϊκές πηγές του Διαδικτύου - καταλογογραφούνται και εισάγονται στους καταλόγους τους. Ειδικότερα, ορισμένες βιβλιοθήκες προσφέρουν στις βιβλιογραφικές τους εγγραφές σε αρχεία μεταδεδομένων, έτσι ώστε οι εγγραφές των φυσικών τους συλλογών υλικού να είναι

προσπελάσιμες από αυτές των ψηφιακών βιβλιοθηκών. Ο προσδιορισμός των μεταδεδομένων επηρεάζεται από:

1) Τον τύπο του πληροφοριακού υλικού που φιλοξενεί σε μια βιβλιοθήκη.

2) Τις απαιτήσεις αναζήτησης,

3) τις ανάγκες πληροφόρησης που απαιτούνται που προέρχονται από τους χρήστες.

Στην δεδομένη χρονική στιγμή έχουν δημιουργηθεί διάφορα πρότυπα μεταδεδομένων με ένα σύνολο επίπεδων πολυπλοκότητας ως προς την περιγραφή των δεδομένων όπως για παράδειγμα το Dublin Core Metadata Initiative, το MARC, το PRISM, το UNIX for books κ.α. Ωστόσο, δεν έχει ακόμη προσδιοριστεί ένα διεθνές πρότυπο μεταδεδομένων για τις ακαδημαϊκές βιβλιοθήκες. Κάθε αρχιτεκτονική μεταδεδομένων στηρίζεται στην περιγραφή των πεδίων του καταλόγου όπως για παράδειγμα του συγγραφέας, του τίτλου, της ημερομηνίας, του εκδότης καθώς και της χρήσης λεξιλογίων που ελέγχονται και που δίδουν πρόσβαση στους καταλόγους των βιβλιοθηκών.

Οι ακαδημαϊκές βιβλιοθήκες κάνουν χρήση πολύπλοκων δομών περιγραφής των βιβλιογραφικών τους δεδομένων. Παρόλα αυτά, οι αρχιτεκτονικές μεταδεδομένων παρέχουν μόνο ένα ελάχιστο ποσοστό αυτής της πολυπλοκότητας, γεγονός που συντελεί σε ικανοποιητικά επίπεδα λειτουργικότητας και χρηστικότητας των μεταδεδομένων στο Διαδίκτυο (Paepckeetal.. 1998).

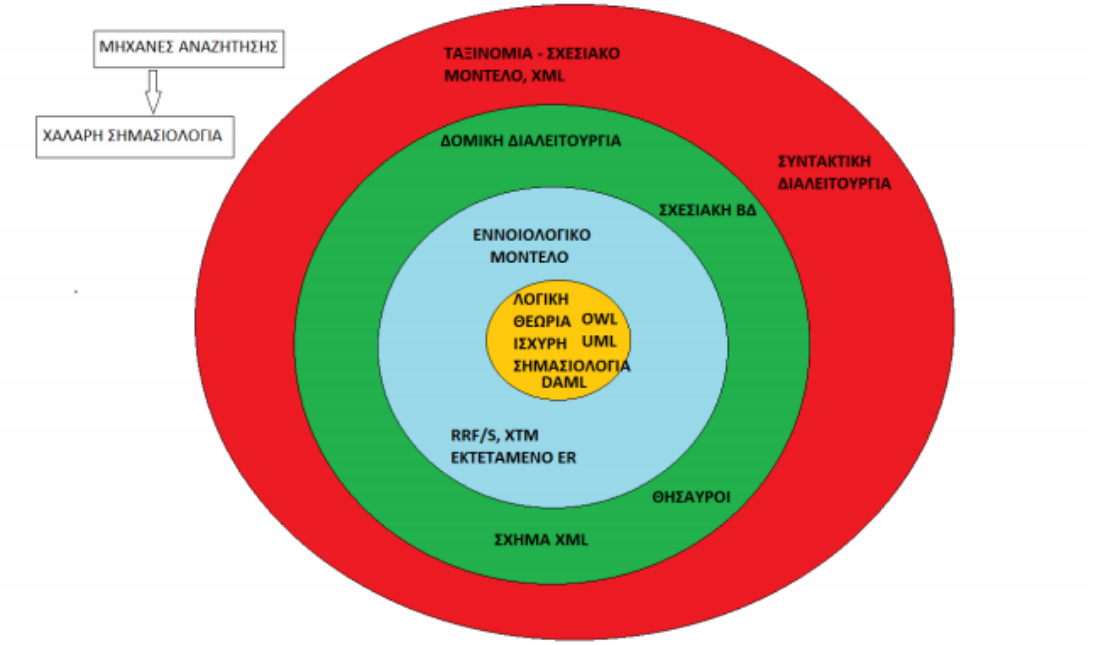
Στην παρούσα εργασία περιγράφονται και αναλύονται οι μορφές και οι μηχανές αναζήτησης που σχετίζονται με τις τεχνολογίες του Σημασιολογικού Ιστού που προάγουν τη διαλειτουργικότητα, τη δυνατότητα επαναχρησιμοποίησης και διαμοιρασμού του ψηφιακού υλικού των ακαδημαϊκών βιβλιοθηκών. Τεκμηριώνεται η άποψη ότι ο Σημασιολογικός Ιστός χαρακτηρίζεται ως μία πλατφόρμα για σημασιολογικές μηχανές αναζήτησης, διαμεσολαβητές πληροφορίας και έξυπνους πράκτορες λογισμικού. Οι δυνατότητες αυτές του Σημασιολογικού Ιστού συντελούν σε μεγάλο βαθμό στην ανάπτυξη του τρόπου λειτουργίας των ακαδημαϊκών βιβλιοθηκών.

3.1.2 Σημασιολογικά Μοντέλα

Ενώ έχουν πραγματοποιηθεί σημαντικά βήματα στη σύνδεση εντελώς ανόμοιων πηγών δεδομένων μέσα από την χρησιμοποίησε και χρήση εκλεπτυσμένων λύσεων ενδιάμεσου λογισμικού, προχωρημένων πρωτοκόλλων ανταλλαγής δεδομένων και κοινών πρότυπων λεξιλογίου, υπάρχει ακόμη μια παράληψη συσχετισμών στο επίπεδο της σημασιολογίας. Ένα από τα σημαντικότερα ζητήματα είναι η πραγματική τιθάσευση της δύναμης της πληροφορίας και τον καλύτερο εξοπλισμό μας για να μπορεί ο χρηστής να ανταποκριθεί στις προκλήσεις, είναι η έλλειψη κατανόησης του τι προσδιορίζει η πληροφορία, και πώς χρησιμοποιείται σε ένα σύστημα, σε σχέση με κάποιο άλλο. Ακόμα, η σημασιολογία αποτελεί κλάδο της γλωσσολογίας που καταπιάνεται με τη μελέτη της σημασίας, τις αλλαγές στη σημασία και τις αρχές που περιγράφουν τη σχέση ανάμεσα στις προτάσεις ή λέξεις και τις σημασίες τους, ενώ η σημασιολογία της πληροφορίας είναι η σημασιολογική αναπαράσταση των συστημάτων, των δεδομένων, των εγγράφων ή των παραγόντων μας. Η σημασιολογία της πληροφορίας αναπαριστά τα οργανωτικά και πολιτισμικά πλαίσια ως ενσωματωμένα εντός των οργανωτικών αποστολών, ιεραρχιών, λεξιλογίων, ροής εργασίας, και προτύπων εργασίας. Η ίδια έννοια μπορεί να προσδιορίζει με διαφορετικούς όρους, όπως για παράδειγμα σε ένα σύστημα μπορεί να κάνει την εμφάνιση του «τιμή» και σε ένα άλλο «κόστος». Ο ίδιος όρος είναι δυνατόν να έχει διαφορετικές σημασίες. Τα σημασιολογικά μοντέλα προσδιορίζουν τους σημασιολογικούς συσχετισμούς ανάμεσα σε διαφορετικούς όρους και έννοιες. Οι όροι και οι έννοιες αποτελούν δυο διαφορετικά στοιχεία. Οι όροι είναι λέξεις ή φράσεις. Οι έννοιες είναι η σημασία πίσω από τους όρους που περιγράφουν τη σημασιολογία. Οι όροι, επομένως, δρουν ως ταμπέλες για τις έννοιες. Είναι δυνατόν να υπάρξει υπάρχει η έννοια του Ατόμου στην οποία όλοι οι όροι «άτομο», «κόσμος», «άνθρωπος» κτλ αναφέρονται. Υπάρχουν αρκετές μορφές σημασιολογικών μοντέλων. Για να βοηθήσουμε τη συζήτησή μας για τα σημασιολογικά μοντέλα, κάνοντας χρήση στην εικόνα «το φάσμα της οντολογίας» για να δείξουμε το «φάσμα της οντολογίας». Το φάσμα της οντολογίας (βλ . Σχήμα 3.1) παρουσιάζει την αλληλουχία μοντέλων, που κλιμακώνονται από κάτω αριστερά προς τα πάνω δεξιά, από μοντέλα με μικρότερη εκφραστική σημασιολογία («χαλαρή»

σημασιολογία) σε μοντέλα με αυξητικά περισσότερο εκφραστική σημασιολογία («ισχυρή» σημασιολογία»).

Σχήμα 3.1. Το φάσμα της οντολογίας



Η πρόοδος από κάτω αριστερά προς τα πάνω δεξιά αποτελεί ακόμα αύξηση της ποσότητας της δομής για το μοντέλο, με τα εκφραστικότερα σημασιολογικά μοντέλα να έχουν την περισσότερη δομή. Συμπεριλαμβάνουμε παραπλεύρως του φάσματος κάποιους τύπους μοντέλων και γλωσσών που είτε γνωρίζουμε είτε έχουμε ακουστά γι' αυτές: όπως για παράδειγμα το μοντέλο σχεσιακής βάσεως δεδομένων και η XML βρίσκονται κάτω αριστερά μετακινούμενοι προς τα πάνω και δεξιά έχουμε το Σχήμα XML, τα μοντέλα Σχέσεως-Οντότητας, το XTM (το πρότυπο θεματικού χάρτη της XML - XML Topic Map standard), την RDF/S, την UML (Ενοποιημένη Γλώσσα Απεικόνισης - Unified Modeling Language), την OWL (γλώσσα οντολογίας του Ιστού - Web Ontology Language) και μέχρι τη λογική πρώτης τάξης (τον κατηγορηματικό λογισμό - Predicate Calculus) και πιο πάνω.

Στην ουσία, το φάσμα μεγαλώνει και πέραν εκείνων των μοντέλων που μας ενδιαφέρουν. Μια από τις απλούστερες μορφές του σημασιολογικού μοντέλου είναι η ταξινομία. Οι ταξινομίες περιγράφουν απλά ως οι δομές στις οποίες γίνεται χρηση για να οργανωθεί η πληροφορία. Όταν ο κόσμος αναφέρεται στις ταξινομίες, συνήθως καταλαβαίνει ιεραρχικές δομές, όπως αυτές που φτιάχνουμε στις βιολογικές

επιστήμες. Από την οπτική μιας επιστήμης της πληροφορίας, οι ταξινομίες είναι δυνατόν να εμπλέκουν σε ένα συνδυασμό διάφορων τύπων δομών. Είναι δυνατόν να είναι επίπεδες δομές, ιεραρχίες, δικτυακές, πλεγματικές δομές ή πολύπλευρες ταξινομίες. Κάθε ένα από αυτά τα είδη δομών είναι χρήσιμο και ως ένα διαφορετικού είδος διαχείρισης της πληροφορίας και σκοπός πρόσβασης. Όλα τους είναι σημαντικά για την υποστήριξη των περίπλοκων σημερινών λύσεων για την πληροφορία και προσδιορίζουν ζωτικά συστατικά των περίπλοκων σημερινών συστημάτων της πληροφορίας (Bedford, 2004). Συμπεριλαμβανομένου και το θησαυρό, το εννοιολογικό μοντέλο και τη λογική θεωρία στην εικόνα «το φάσμα της οντολογίας», επειδή αυτά δρουν ως «ενδιάμεσοι σταθμοί» της διογκούμενης πολυπλοκότητας και του σημασιολογικού πλούτου καθώς πηγαίνουμε προς τα πάνω στο οντολογικό φάσμα.

Ένας θησαυρός είναι πιο περίπλοκος από μια ταξινομία, επειδή η σχέση γονέα-παιδιού που έχει προσδιοριστεί με συνέπεια ως «ευρύτερη από / στενότερη από», όπως για παράδειγμα ένας μητρικός κόμβος έχει μια ευρύτερη σχέση με τους θυγατρικούς του κόμβους ένας θυγατρικός κόμβος έχει μια στενότερη απ' ότι η σχέση με το μητρικό του κόμβο. Υπάρχουν και σχέσεις υπαγωγής, έτσι ένας γονέας υπάγει ένα παιδί. Παρόλα αυτά, σε ένα θησαυρό, οι κόμβοι δεν είναι απλά ταξινομήσεις όπως είναι σε μια ταξινομία, ούτε είναι «τάξεις» ή «έννοιες» όπως είναι σε ένα εννοιολογικό μοντέλο ή μια λογική θεωρία. Οι κόμβοι είναι μάλλον «όροι», δηλαδή λέξεις ή φράσεις, και οι όροι αυτοί έχουν στενότερες από ή ευρύτερες από σχέσεις ανάμεσά τους. Ένας θησαυρός είναι ουσιαστικά σχέσεις μεταξύ όρων από δομημένους κατά μία ταξινομία. Ένας θησαυρός περιλαμβάνει και άλλες σημασιολογικές σχέσεις ανάμεσα στους όρους, όπως τα συνώνυμα. Το σημασιολογικό μοντέλο στο οποίο οι σχέσεις καλούνται και διαφοροποιούνται σαφώς ονομάζεται οντολογία. Στην εικόνα «το φάσμα της οντολογίας», τόσο τα εννοιολογικά μοντέλα όσο και οι λογικές θεωρίες έχουν την δυνατότητα να θεωρηθούν ως οντολογία. Επειδή οι σχέσεις περιγράφονται, δεν υπάρχει πλέον ανάγκη αυστηρής δομής που καθορίζει τις σχέσεις. Το μοντέλο ουσιαστικά αποτελείται από ένα δίκτυο συνδέσεων, με κάθε σύνδεση να έχει έναν ανεξάρτητο συσχετισμό από κάθε άλλη σύνδεση. Η μεταβλητότητα αυτή προσφέρει φοβερή ευελιξία στην αντιμετώπιση εννοιών, επειδή πολλοί εννοιολογικοί τομείς δεν έχουν την δυνατότητα να εκφραστούν επαρκώς με μια ταξινομία. Σε μια ταξινομία ή ένα

θησαυρό, εμφανίζεται ένα πλήθος από ανωμαλίες και αντιφάσεις, επιβάλλοντας έτσι αβάστακτους συμβιβασμούς. Ακόμα, η μετακίνηση ανάμεσα σε ανόμοιες έννοιες συχνά έχει ανάγκη από εύθραυστους συνδετικούς μηχανισμούς που είναι δύσκολο να συντηρούνται ή να μεγαλώσουν (Πούλος, 2015: 42-44).

3.1.3 Σημασιολογικές πύλες

Ο Παγκόσμιος Ιστός αποτελείται από έναν τεράστιο όγκο πληροφορίας και έχει φοβερή επιτυχία τόσο σε όρους διαθέσιμης πληροφορίας όσο και σε ρυθμό ανάπτυξης του αριθμού των ατόμων που κάνουν χρήση και μεταχειρίζονται αυτά τα δεδομένα. Στο οποίο εισέρχονται πάνω από ένα δισεκατομμύριο χρήστες και υπάρχουν πάνω ένα δισεκατομμύριο έγγραφα. Η επιτυχία πρόεκυψε στο μεγαλύτερο ως αποτέλεσμα της απλότητάς που το διακρίνει, η οποία παρείχε στους προγραμματιστές, στους παρόχους πληροφορίας και στους χρήστες εύκολη πρόσβαση σε αυτό το υλικό. Παρ' όλα αυτά, η ίδια η απλότητά του και η τεραστία διάδοση του είναι και αυτά που συμβάλουν σε σημαντικό βαθμό σοβαρά εμπόδια για την περαιτέρω ανάπτυξη του Παγκόσμιου Ιστού. Το συνολικότερο πρόβλημα στο να εντοπίσει κάνεις πληροφορίες στον Παγκόσμιο Ιστό πηγάζει από το γεγονός ότι αναζητήσεις είναι ανακριβείς και στο ότι οι λέξεις που χρησιμοποιούνται επιστρέφουν πολλές χιλιάδες αποτελέσματα. Ακόμα οι χρήστες έρχονται αντιμέτωποι με ζητήματα ότι είναι αναγκασμένοι να διαβάσουν το έγγραφο που επιστρέφεται από την αναζήτηση προκειμένου να προσδιορίσουν αν είναι αυτό που έψαχναν. Αυτοί οι περιορισμοί κάνουν την εμφάνιση τους φυσικά στις υπάρχουσες Δικτυακές Πύλες που στηρίζονται σε αυτήν την τεχνολογία, προσδιορίζοντας αυτή την αναζήτηση, την πρόσβαση, την εξαγωγή, τη μετάφραση και την επεξεργασία της πληροφορίας μια δύσκολη και χρονοβόρα διαδικασία. Για τους προαναφερθέντες λόγους οι τεχνολογίες του σημασιολογικού ιστού είναι δυνατόν να καλυτερεύσουν αισθητά τη διαδικασία διανομής της πληροφορίας ξεπερνώντας τα ζητήματα που εμφανίζουν οι σημερινές Δικτυακές Πύλες. Στην βάση αυτών των στοιχείων, οι Δικτυακές Πύλες που στηρίζονται στις τεχνολογίες του Σημασιολογικού Ιστού και είναι δεδομένο ότι θα αποτελούν την επόμενη γενιά Δικτυακών Πυλών και ονομάζονται Σημασιολογικές Πύλες (ΣΠ). Σημασιολογικές πύλες είναι οι ακόλουθες:

Mindswap

Οι δημιουργοί αυτής της Πύλης πιστεύουν ότι αυτή είναι η πρώτη ιστοσελίδα που έκανε χρήση στο Σημασιολογικό Ιστό υλοποιείται με τεχνολογίες συμβατές με OWL. Η ιστοσελίδα αυτή διενεργεί μια προσπάθεια να επιδείξει τις αυξημένες

δυνατότητες που προσφέρουν σε μια ιστοσελίδα οι τεχνολογίες του Σημασιολογικού Ιστού. Παρ' όλα αυτά η ιστοσελίδα αυτή δεν δείχνει τις τεχνολογικές δυνατότητες που παρέχει η εφαρμογή τεχνολογιών του Σημασιολογικού Ιστού: για παράδειγμα ούτε η αναζήτηση ούτε το μενού κάνουν εφαρμογή σε τεχνολογίες του Σημασιολογικού Ιστού.

Οι Πύλες της Karlsruhe

Το AIFB στο Πανεπιστήμιο της Καρλσρούης έχει δημιουργήσει μια από τις αρχικές Σημασιολογικές Πύλες. Στοχεύοντας στο να αποτελέσει την πλατφόρμα για ανταλλαγή πληροφορίας και συνεργασία για την κοινότητα KA2 (Knowledge Annotation initiative of the Knowledge Acquisition community). Για αυτό το λόγο κατασκευάστηκε μια Οντολογία ως βάση για να προσδιοριστούν δικτυακά έγγραφα για την κοινότητα των χρηστών που ενδιαφέρονται για την ανάκτηση γνώσης ώστε να γίνει εφικτή η «έξυπνη» πρόσβαση σε αυτά. Η Πύλη αυτή όμως δεν υποστηρίζεται πλέον και να δε συντηρείται. Μολονότι η Πύλη στηρίζεται συνολικά σε Οντολογία, οι λειτουργικότητες που περεχεί είναι πολύ απλές και σε επίπεδο πρόσβασης στην πληροφορία από τον τελικό χρήστη είναι σε πολύ μικρότερο επίπεδο από αντίστοιχες συμβατικές Πύλες.

OntoWeb Portal

Η OntoWeb αποτελεί μια Δικτυακή Πύλη για μια κοινότητα ατόμων, από τα ακαδημαϊκά ιδρύματα και τη βιομηχανία, που παρουσιάζουν ενδιαφέρον για το Σημασιολογικό Ιστό. Η Πύλη αυτή είχε κατασκευαστεί σαν μέρος του προγράμματος OntoWeb της Ευρωπαϊκής Ένωσης (Gerberat. el.,2008).

3.2 Τεχνολογίες web 2,deepweb

3.2.1 web 2

Ένα από τα σημαντικότερα ζητήματα στο χώρο του διαδικτύου και της πληροφορικής τα τελευταία χρόνια είναι η εξέλιξη που έχει επέλθει στο παραδοσιακό Web 1.0 και σε αυτό που έχει επικρατήσει και καλείται Web 2.0. Η χρήση του διαδικτύου σχετίζεται με ολοένα και περισσότερες ανθρώπινες δράσεις και η σημασία της εξέλιξης του είναι ιδιαιτέρως σημαντική. Οι χρήστες, ακόμα και αν είναι ιδιώτες, εταιρίες, οργανισμοί, εκπαιδευτικά ιδρύματα, κλπ. Με το πέρασμα του χρόνου ενημερώνονται για τα στοιχεία και τις τεχνολογίες που συνιστούν το Web 2.0 και επωφελούνται από τα πλεονεκτήματά του. Στην συνέχεια αναλύονται τα χαρακτηριστικά του Web 2.0 και την επίδραση τους (oreilly.com/web2/archive/what-is-web-20.html).

Το ακριβές νόημά του όρου δεν έχει κλείσει ακόμα προς αντιπαράθεση και μερικοί ειδικοί, συμπεριλαμβανομένου του Tim Berners-Lee, έχουν αμφισβητήσει ανά διαστήματα τον όρο και το κατά ποσό έχει κάποιο πραγματικό νόημα. Ανάμεσα σε άλλα, το Web 2.0, έχει λοιδορηθεί ότι αποτελεί εφεύρεση του μάρκετινγκ. Μάλιστα επιχείρημα που αποτέλεσε το ότι οι τεχνολογίες της οποίες χρησιμοποιεί το διαδίκτυο αναβαθμίζονται διαρκώς και πως η προηγούμενη μέρα πριν το Web 2.0 δεν είναι μακριά. Ακόμα θεωρήθηκε πως με την ίδια λογική, σε ελάχιστο διάστημα μετά την εφεύρεση του Web 2.0 θα έπρεπε να εμφανιστεί το Web 2.1. Είναι όμως αδύνατο να εκφραστεί η τεχνολογική ιδιότητα του internet ακριβώς με έναν αριθμό μιας και το αμάλγαμα τεχνολογιών που είναι αυτή τη στιγμή χρησιμοποιούνται στο διαδίκτυο είναι αχανώς πολυδιάστατο. Ακόμα, λειτουργίες επικοινωνίας με τον χρήστη υπήρχαν εδώ και πολλά χρόνια, όπως για παράδειγμα η χρήση σελίδων χρήστη (homepages), τα φόρα, τα chat (IRC) και άλλα. Αυτό που μπορεί πάντως να αναφερθεί με βεβαιότητα είναι ότι άλλαξε η ευκολία χρήσης των διαδικτυακών εφαρμογών. Αν και ο όρος Web 2.0 προσφέρει την αίσθηση ότι αποτελεί μια καινούργια έκδοση Web, εν τέλει δεν αφορά κάποιο καινούργιο πρωτόκολλο αλλά αναφέρεται στις μεταβολές του τρόπου που αξιοποιούνται οι ήδη υπάρχουσες τεχνολογίες και στον τρόπο που οι σχεδιαστές πληροφοριακών συστημάτων καθώς και οι χρήστες κάνουν χρήση στο διαδίκτυο.

Για να μπορέσει κάνεις να παράγει συμπεράσματα είναι δυνατόν να παρατηρήσει την κατάσταση που επικρατούσε λίγα χρόνια πριν στο διαδίκτυο. Ο χρήστης απλά επισκέπτεται ιστοσελίδες χωρίς να έχει αρκετές δυνατότητες δημιουργίας πληροφοριών. Χαρακτηριστικό είναι το παράδειγμα πως πριν μερικά χρόνια αποτελούσε κατόρθωμα όταν κάποιος κατάφερνε να κατασκευάσει ένα βίντεο στο web ενώ την δεδομένη χρονική στιγμή αρκούν μόλις μερικά δευτερόλεπτα για να ανεβάσει ένας αρχάριος χρήστης βίντεο στο Youtube και σε παρόμοιες σελίδες. Γενικότερα, τα τελευταία χρόνια έχουν διενεργηθεί κοσμοϊστορικές μεταβολές στο διαδίκτυο στην βάση του web 2. Σταδιακά οι χρήστες ξεκίνησαν από μόνοι τους να παρουσιάζουν τις ανάγκες τους για κοινωνική δικτύωση, αυτό συνέβαλε στην δημιουργία αρκετών υπηρεσιών οι οποίες έχουν ως σημείο αναφοράς τον ίδιο τον χρήστη, προσφέροντας του την δυνατότητα να συμμετέχει ο ίδιος στην ανάπτυξη του περιεχομένου και στη σχεδίαση των διαδικτυακών εφαρμογών.

Ο όρος Web 2.0 λοιπόν, είναι δυνατόν να χρησιμοποιηθεί για να περιγράψει την δεύτερη γενιά υπηρεσιών διαδικτύου που βασίζεται στην δυνατότητα των χρηστών να διαμοιράζονται πληροφορίες και να συνεργάζονται online. Ο χρήστης δεν αποτελεί πια έναν απλό ως θεατή, έναν πελάτη, έναν καταναλωτή αλλά συμμετέχει ενεργά, και συχνά αλτρουιστικά στην διαμόρφωση και διαχείριση των πληροφοριών του διαδικτύου. Χρήστες από διαφορετικές κουλτούρες έχουν την δυνατότητα πια να επικοινωνούν δίχως να είναι αναγκαίες οι εξειδικευμένες γνώσεις σε ζητήματα υπολογιστών και δικτύων. Ο αρχικός παθητικός ρόλος παρουσίασης των πληροφοριών διαρκώς μεταβάλλεται. Έννοιες όπως διαδραστικότητα, δυναμικό περιεχόμενο, συνεργασία, συνεισφορά και socialcomputing διαδραματίζουν πια πρωτεύοντα ρόλο και αρκετοί υποστηρίζουν ότι μια τεχνολογική και κοινωνική επανάσταση είναι παρούσα και εξελίσσεται μπροστά στα μάτια μας (www.broadband.gr/opencms/sites/Broadband/News/news071228c/).

Το web 2.0 ξεπερνά τα όρια της περιορισμένης σε έναν υπολογιστή πλατφόρμας. Ο χρήστης έχει την δυνατότητα να δρα στον Παγκόσμιο Ιστό όπως δρούσε μέχρι τώρα στον υπολογιστή του. Οι ειδικοί αναφέρονται σε έναν νέο τρόπο σχεδίασης των ιστοσελίδων ο οποίος θα στηρίζεται κατά βάση στην δράση του χρήστη και θα του προσφέρει την δυνατότητα να μεταβάλει τόσο το περιβάλλον της σελίδας όσο και να παρέμβει στο περιεχόμενό των πληροφοριών που διαθέτει. Πολλές από τις διαδράσεις που προσδιορίζουν την λειτουργία του web 2.0, είναι ήδη γνωστές από αρκετές

ιστοσελίδες όπως το facebook, το YouTube κ.ά. Ορισμένες εκφράσεις διάδρασης είναι η αναζήτηση (search), η προσθήκη ετικετών (tagging), η παράθεση/επεξεργασία συνδέσμων (linking) ή το authoring όπως λειτουργεί σε πολλά wiki, όπου οι χρήστες έχουν την δυνατότητα να δημιουργούν, να επεξεργάζονται ή να διαγράφουν πληροφορίες. Ακολούθως σε αφαιρετικό επίπεδο θα διακριθούν κάποια από τα χαρακτηριστικά του web 2.0 που θα συμβάλουν στην καλύτερη αντίληψη του ορού και τις έννοιας

- Το διαδίκτυο και όλες οι συσκευές που συνδέονται με αυτό, αφορούν μια παγκόσμια πλατφόρμα επαναχρησιμοποιούμενων υπηρεσιών και δεδομένων, τα οποία προκύπτουν κατά βάση από τους ίδιους τους χρήστες και σχεδόν στο μέγιστο των περιπτώσεων διακινούνται ελεύθερα.
- Αρκεί ένας φυλομετρητής ώστε να "τρέξει" μια web 2.0 εφαρμογή, η οποία λειτουργεί ανεξαρτήτως συσκευής πρόσβασης όπως για παράδειγμα Η/Υ, PDA2, κινητό τηλέφωνο και λειτουργικού συστήματος. Μόνη προϋπόθεση αποτελεί η ύπαρξη δικτύου που να συνδέει στο διαδίκτυο.
- Λογισμικό, περιεχόμενο και εφαρμογές ανοιχτού κώδικα (opensource).
- Χρήση κυρίως “ελαφριάς” τεχνολογίας που σχετίζεται με τα πρωτόκολλα, τις γλώσσες προγραμματισμού, τις διεπαφές χρήστη, ενώ περιγράφεται και μια τάση για απλότητα στον προγραμματιστικό σχεδιασμό τους.
- Πολυμεσικές και διαδραστικές διεπαφές χρήστη (Rich Internet Applications-RIA), δυναμικό περιεχόμενο, ιστοσελίδες που μεταβάλουν αποκλειστικά και μόνο το περιεχόμενό που αλλάζει (τεχνολογία Ajax).
- Διαρκής και άμεση ανανέωση των δεδομένων και του λογισμικού που επιβάλλεται να προσαρμόζεται διαρκώς στις απαιτήσεις των χρηστών.
- Προώθηση του δημοκρατικού χαρακτήρα του διαδικτύου, με τους χρήστες να έχουν καθοριστικό ρόλο.
- Υιοθέτηση της τάσης για αποκέντρωση των δεδομένων, υπηρεσιών και προτύπων.
- Δυνατότητα κατηγοριοποίησης του περιεχομένου από το χρήστη με σημασιολογικές έννοιες για ευκολότερη και καλύτερη αναζήτηση της πληροφορίας.
- Δυνατότητα για ανοιχτή επικοινωνία, ανάδραση, διάχυση πληροφοριών, άμεση συγκέντρωση και αξιοποίηση της γνώσης των χρηστών για διάφορα ζητήματα.

- Αμφίδρομη επικοινωνία του χρήστη με επιχειρήσεις ή οργανισμούς που είναι δυνατόν να έχει σαν αποτέλεσμα την υιοθέτηση κατευθύνσεων και τη λήψη αποφάσεων.

Κατηγορίες και παραδείγματα Web 2.0 εφαρμογών

Το web 2.0 οφείλει την ύπαρξή του σε εφαρμογές, υπηρεσίες, εργαλεία και λειτουργίες που προσδιορίζονται από καινοτομίες και ευκολίες τις οποίες επιζητούν οι χρήστες, για αυτό όταν αυτές δημιουργήθηκαν υποδεχτήκαν από αυτούς με μεγάλη αποδοχή και γνωρίζουν τεραστία διάδοσης. Παρακάτω, περιγράφονται μερικές από τις βασικότερες κατηγορίες Web 2.0 εργαλείων:

- **Blogs:** Τα ιστολογία (blogs) ουσιαστικά είναι ιστοσελίδες που περιλαμβάνουν απόψεις, πληροφορίες, προσωπικές καταχωρήσεις, συνδέσεις με άλλες διευθύνσεις, φωτογραφίες, κ.α. Οι καταχωρήσεις είναι τοποθετημένες με χρονολογική σειρά και ξεκινούν με την άποψη ή το σχόλιο του δημιουργού τους για ένα ζήτημα όπως για παράδειγμα πολιτική, επιστήμη, κοινωνικά, καθημερινότητα κλπ. Η διάδοσή τους πηγάζει κατά κύριο λόγο κυρίως στο ότι παρέχουν τη δυνατότητα σε όποιον χρήστη επιθυμεί να γράψει το σχόλιό του, ανοίγοντας έτσι ένα δημόσιο διαδικτυακό διάλογο με πιθανούς αποδέκτες το σύνολο των χρηστών. Στις αρχές του 2008 μετρήθηκαν πάνω 112.000.000 blogs στο σύνολο του διαδικτύου με τα στατιστικά της μηχανή αναζήτησης της Technorati. Λόγω αυτής της δημοτικότητας, της αίσθησης κοινωνικοποίησης μεταξύ των συμμετεχόντων και της επίδρασης που διαθέτουν ακόμη και εκτός διαδικτύου, προσδιορίζονται από πολλούς ως καινούργιο κοινωνικό φαινόμενο. Χαρακτηριστικά ο Rodzvilla (2002), "τα blogs είναι πολυμεσικά και εύκολα στη χρήση websites που μέσα από τη χρονολογική τους δομή και τις αρχειοθετικές τους δυνατότητες λειτουργούν ως εξατομικευμένα και διασυνδεδεμένα φίλτρα του web δημιουργώντας μια νέα online δημόσια σφαίρα που γύρισε το web πίσω στον κόσμο". Κάποια παραδείγματα του web 2.0 εργαλείων δίδον τη δυνατότητα δημιουργία και

την φιλοξενία ιστολογίων είναι τα: Blogger, Edublogs, LiveJournal, Tumblr και Posterous.

- **Wikis:** Τα wikis αποτελούν ιστοσελίδες με περιεχόμενο το οποίο δημιουργεί και χωρίζει ο χρήστης με απλό τρόπο, σε αντίθεση με τις κοινές ιστοσελίδες τις οποίες έχει τη δυνατότητα να μεταβάλει αποκλειστικά και μόνο ο ιδιοκτήτης – διαχειριστής. Κάθε φορά που ο χρήστης μεταβάλει κάτι στη σελίδα, η προηγούμενη έκδοσή της εξακολουθεί να είναι διαθέσιμη. Τα wikis διαθέτουν αρκετή φήμη σαν μέσο συλλογικής εργασίας πάνω σε κάποιο ζήτημα. Προσφέρουν τη δυνατότητα στα μέλη μιας ομάδας χρηστών, να καταθέτουν ισότιμα τη συμβολή τους για την παραγωγή ενός κοινού έργου που αναρτάται σε έναν δικτυακό τόπο όπως για παράδειγμα μια μικρή σχολική έρευνα, παραγωγή σημειώσεων, ανταλλαγή ιδεών για ένα αντικείμενο συζήτησης κ.α. Ο κάθε χρήστης που λαμβάνει μέρος στη συγγραφή κάποιου έργου προσθέτει την προσωπική του γνώση η οποία μπορεί να διαβαστεί από το σύνολο των χρηστών. Ακόμη και μέσα σε εταιρίες, οργανισμούς, υπηρεσίες, κ.α. η χρήση των wikis ως σελίδες αναφοράς της προόδου των εργασιών, συμβάλει στην ενημέρωση των εργαζομένων για ό,τι συμβαίνει στην εταιρία. Χαρακτηριστικό παράδειγμα wiki είναι η Wikipedia, που αποτελεί μια διαδικτυακή εγκυκλοπαίδεια στην οποία υπάρχουν πάνω από πέντε εκατομμύρια άρθρα με ορισμούς και πληροφορίες σε παρά πολλές γλώσσες. Η σύνταξή της πραγματοποιείται από τους χρήστες, αφού οποιοσδήποτε μπορεί να γράψει ένα καινούργιο άρθρο ή να προσθέσει κάτι στα ήδη υπάρχοντα. Η δημοτικότητά της αυξάνει διαρκώς και βάσει του αριθμού επισκέψεων βρίσκεται μέσα στα δέκα δημοφιλέστερα sites σε παγκόσμιο επίπεδο. Μερικά παραδείγματα web 2.0 εργαλείων που παρέχουν υπηρεσίες δημιουργίας και φιλοξενίας wiki είναι το Wetpaint, το Wikispaces, το Foswiki και το Mediawiki.

- **Mash-ups:** Ο όρος προέρχεται από τη μουσική βιομηχανία και γίνεται χρήση του για να περιγράψει τον συνδυασμό των φωνητικών ενός τραγουδιού με τη μουσική υπόκρουση από κάποιον άλλο. Κάτι παρόμοιο συμβαίνει και με το mashup, που αφορά τον συνδυασμό και τη χρήση δεδομένων και εφαρμογών από διαφορετικές ιστοσελίδες σε μία. Τα mash-

ups υλοποιούνται διαμέσου "ανοιχτών" διεπαφών προγραμματισμού (openAPIs'–Application Programming Interfaces) και έχουν ως σκοπό την βελτίωση της λειτουργικότητας των ιστοσελίδων. Για παράδειγμα, σε ιστοσελίδες ενοικίασης σπιτιών, με την ενσωμάτωση χαρτών από μία υπηρεσία όπως η GoogleMaps, έχει την δυνατότητα να προσφέρει στο χρήστη την ακριβή τοποθεσία των σπιτιών ώστε να παρέχεται ειδικότερη πληροφόρηση. Ορισμένα Web 2.0 εργαλεία, που αναλαμβάνουν την δημιουργία mash-ups και προσαρμόζονται στις ανάγκες του κάθε χρήστη, αποτελούν τα: iGoogle, Pageflakes και Netvibes.

- **Micro-blogging**: Τα μικρο-ιστολόγια είναι κοινωνικές πλατφόρμες blogging που σου επιτρέπουν να έρθεις σε επαφή και να αλληλεπιδράσεις με άλλα μέλη. Ο όρος μικρο-blogging εστιάζει στο ότι ο χρήστης καλείται να δημοσιοποιήσει την κατάστασή του μέσα σε 140 χαρακτήρες κειμένου, δίχως την χρήση εικόνων ή άλλων πολυμέσων. Για τον micro-blogger είναι πολύ πιο εύκολο να γράψει κάτι, από το να προετοιμάσει ένα blog post για το Wordpress ή το Blogger. Επίσης, μία άλλη διαφορά του microblogging, είναι ότι προσφέρει στους χρήστες τη δυνατότητα να κάνουν post χρησιμοποιώντας διάφορα μέσα, όπως υπολογιστές, κινητά τηλέφωνα με υποστήριξη SMS ή Wi-Fi, messengers και email. Αυτή η πληθώρα επιλογών, σε συνδυασμό με το μικρό μέγεθος των μηνυμάτων, έχει καταστήσει τις microblogging πλατφόρμες πολύ δελεαστικές, ιδιαίτερα σε χρήστες που θέλουν να εκφραστούν ανά πάσα στιγμή για οτιδήποτε μπορεί να θεωρούν αξιόλογο προς αναφορά. Μέσα από τα μικρο-ιστολόγια μπορούμε λοιπόν να στέλνουμε σύντομα μηνύματα με τις σκέψεις, τις δραστηριότητες, τις ερωτήσεις μας ή οτιδήποτε άλλο θέλουμε, τα οποία θα λαμβάνουν όσοι έχουν επιλέξει να μας ακολουθούν μέσω της υπηρεσίας. Κατά αντίστοιχο τρόπο μπορούμε να βλέπουμε τα μηνύματα όσων έχουμε επιλέξει να ακολουθούμε. Παραδείγματα web 2.0 εργαλείων που δραστηριοποιούνται στο χώρο του micro-blogging είναι το Tweeter, το Gravity, το Cirip, το plinky, το Jaiku και το Pownce.

- **Rss:** Ο όρος RSS πηγάζει από τα αρχικά των αγγλικών λέξεωνReallySimpleSyndication το οποίο αποτελεί ένα format ανταλλαγής περιεχομένου που στηρίζεται στην γλώσσα XML. Τα RSS feeds, παρέχουν

τη δυνατότητα στους χρήστες να λαμβάνουν νέες πληροφορίες από διάφορες ιστοσελίδες, τη στιγμή που αυτές αναρτούνται, χωρίς να είναι αναγκαίο να τις επισκεφθούν. Το RSS είναι δηλαδή ένας νέος τρόπος ενημέρωσης για νέα, εξελίξεις και γεγονότα. Είναι γεγονός πως το διαδίκτυο αποτελείται πλέον από δισεκατομμύρια σελίδες οι οποίες περιέχουν άπειρο πλούτο πληροφοριών που είναι σχεδόν αδύνατο για τον οποιονδήποτε να έχει την δυνατότητα να παρακολουθεί διαρκώς ότι νεότερο συμβαίνει στον κόσμο ή στο πεδίο που τον ενδιαφέρει. Στο πρόβλημα αυτό ήρθαν να δώσουν τη λύση τα RSS feeds. Με το Rss ο χρήστης έχει την δυνατότητα να διακρίνει πότε ανανεώθηκε το περιεχόμενο των δικτυακών τόπων που τον ενδιαφέρουν, λαμβάνοντας κατευθείαν στον υπολογιστή του τους τίτλους των τελευταίων ειδήσεων και των άρθρων κατευθείαν μόλις αυτά γίνουν διαθέσιμα χωρίς να είναι απαραίτητο να επισκέπτεται καθημερινά τους αντίστοιχους δικτυακούς τόπους. Μάλιστα η ενημέρωση είναι δυνατόν να προκύψει και μέσω της φορητής συσκευής του χρήστη όπως για παράδειγμα το κινητό τηλέφωνο, PDA, κ.α. Με αυτό τον τρόπο η σχέση με το διαδίκτυο γίνεται πιο άμεση. Ορισμένα web 2.0 εργαλεία που ειδικεύονται στην παροχή Rss feeds είναι το Feedburner, το RapidFeeds, το FeedJournal και το GoogleAlerts.

- **Social Bookmarking:** Το socialbookmarking (κοινωνική επισήμανση) περιγράφει τον τρόπο με τον οποίο οι χρήστες του διαδικτύου διαμοιράζονται, σχολιάζουν, αναζητούν, διαχειρίζονται και οργανώνουν επισημάνσεις για διάφορες ιστοσελίδες. Οι χρήστες αποθηκεύουν τις επισημάνσεις-προτιμήσεις τους σε σελίδες και τις διαμοιράζονται με άλλους χρήστες. Οι επισημάνσεις μπορούν να διαμοιραστούν δημόσια ή σε συγκεκριμένα ιδιωτικά δίκτυα. Ο όρος social bookmaring προέκυψε από το tagging, δηλαδή την δυνατότητα χαρακτηρισμού με σημασιολογικές λέξεις (tags), ιστοσελίδων, φωτογραφιών, κειμένων και γενικά οποιουδήποτε διαδικτυακού υλικού. Κατά την διάρκεια του tagging, κάθε ιστοσελίδα προσδιορίζεται με περιγραφικές ετικέτες από του χρήστες χωρίς να απαιτείται καμία μορφή ιεραρχικής οργάνωσης. Το τελικό προϊόν αυτής της οργάνωσης ετικετών καλείται “folksonomy.” Η αξία αυτού του εξωτερικού συστήματος οργάνωσης προέρχεται από το γεγονός ότι τα

άτομα κάνουν χρήση του δικού τους λεξικού για να αποδώσουν νοήματα που έχουν εντοπίσει στη συγκεκριμένη σελίδα και τα οποία είναι δυνατόν να μην περιγράφονται ρητά μέσα σε αυτήν. Τα άτομα δηλαδή δεν κατηγοριοποιούν τις ιστοσελίδες άμεσα αλλά έμμεσα αφού περιγράφουν τρόπους με τους οποίους είναι δυνατόν να συνδεθούν αργότερα τα διαφορετικά στοιχεία. Η σύνδεση των ιστοσελίδων μεταξύ τους δε τους επιβαρύνει γνωστικά κατά τη κατασκευή του συνδέσμου. Σήμερα, πολλά εκατομμύρια χρήστες έχουν δημιουργήσει επισημάνσεις σε εκατοντάδες εκατομμύρια ιστοσελίδες. Αποτέλεσμα του social bookmarking είναι το ότι από τη μία οι χρήστες οργανώνουν τα δεδομένα τους πολύ καλύτερα και από την άλλη κοινωνικοποιούνται, γνωρίζοντας τις επιλογές των άλλων ατόμων που έχουν κοινά ενδιαφέροντα με αυτούς. Μερικά παραδείγματα Web 2.0 εργαλείων που δραστηριοποιούνται στον χώρο του social bookmarking είναι το Diigo, το Delicious, το Stumbleupon, το CiteuLike και το Zibaba.

(https://learn20.wikispaces.com/Web+2.0%28%CE%BA%CE%B5%CE%AF%CE%BC%CE%B5%CE%BD%CE%BF%29).

- **Podcasting:** Η λέξη 'Podcast' ανακηρύχθηκε «Λέξη του Έτους 2005» από του εκδότες του NewOxford American Dictionary καθώς ξεκίνησε να γίνεται χρήση της ευρύτατα, λόγω της ευκολίας εγγραφής και αναπαραγωγής των mp3 αρχείων που προσέφερε σε όλους τους χρήστες τη ικανότητα δημιουργίας και αναπαραγωγής podcast με μια σύνδεση στο διαδίκτυο. Το Podcasting λοιπόν αποτελεί μια πρακτική της δημιουργίας αρχείων ήχου που προσφέρονται online με τρόπο τέτοιο που το λογισμικό αναγνωρίζει τα καινούρια αρχεία και τα κατεβάζει αυτόματα. Για να "παίξουν" τα podcasts δεν είναι αναγκαίο το iPod ή κάποια φορητή συσκευή αναπαραγωγής mp3. Κάθε νέο podcast αναφέρεται ως επεισόδιο (episode) και πολλά επεισόδια μαζί που έχουν τη μορφή μιας σειράς αναφέρονται ως κανάλι (channel). Τα podcasts στις περισσότερες περιπτώσεις «κατεβαίνουν» αυτόματα στις κινητές συσκευές αναπαραγωγής ήχου ή τους προσωπικούς υπολογιστές και παρέχουν ροές (feeds) με ενημερώσεις για τις νέες δημοσιεύσεις. Παραδείγματα Web 2.0

εργαλείων που ασχολούνται με το Podcasting είναι τα: voicethread, podhawk, podcast και audacity.

- **Social Networks:** Ως Social Network μπορεί να θεωρηθεί οποιοδήποτε site προσφέρει στους επισκέπτες, μέσω μιας πλατφόρμας, την δυνατότητα δημιουργίας προφίλ και αλληλεπίδρασης με άλλους χρήστες μέσω «κοινωνικών συνδέσεων» εντός ενός χώρου ηλεκτρονικής κοινότητας. Με το όρο φιλία εννοούμε την σύνδεση των προφίλ των χρηστών, με την οποία «ξεκλειδώνονται» κάποια χαρακτηριστικά της πλατφόρμας όπως η ενημέρωση του ενός για τις ανανεώσεις προφίλ του άλλου ή η εμφάνιση προσωπικών φωτογραφιών. Εκτός από φίλους στα social networking sites μπορεί κανείς να συναντήσει και τον όρο θαυμαστές (fans-followers). Ο όρος αυτός εκφράζει την μονόδρομη σχέση επικοινωνίας μεταξύ των χρηστών του δικτύου. Θεωρητικά, οι δυνατότητες αλληλεπίδρασης είναι άπειρες και συνήθως περιορίζονται από τον χαρακτήρα που θέλει να εκφράσει το κάθε socialnetworkingsite. Γενικά, τα social networks μπορούν να κατηγοριοποιηθούν σε δύο βασικές ομάδες, τα κάθετα social networks που περιλαμβάνουν χρήστες-μέλη με κοινά ενδιαφέροντα και κοινούς στόχους και τα οριζόντια social networks, που αποτελούνται από μέλη με διαφορετικά ενδιαφέροντα που συνήθως έχουν ως σκοπό απλά να έρθουν σε επικοινωνία μεταξύ τους, να γνωριστούν και να αλληλεπιδράσουν. Ορισμένα παραδείγματα social networking site αποτελούν τα facebook, myspace, hi, LinkedIn και το zokem. Επιπλέον, αξίζει να σημειωθεί πως έχουν επίσης δημιουργηθεί εργαλεία, όπως το Ning και το Elgg, που δίνουν την δυνατότητα στον χρήστη να αναπτύξει ο ίδιος εύκολα και γρήγορα το δικό του social networking site.

Βέβαια,εκτός από τις κατηγορίες που αναφέρθηκαν πιο πάνω υπάρχουν και άλλες σημαντικές κατηγορίες web 2.0 εργαλείων όπως chat, co-writing, concept mapping, conferencing, course development, file hosting, image processing, microblogging, mushing up, personal file sharing, podcasting-sound, gaming, presentation, quiz development, recommendation, screen casting, video tools, web site creation και work organization (skull.gr/blog/web-20).

Web 2.0 Τεχνολογίες

Στην συνέχεια γίνεται αναφορά στις βασικότερες τεχνολογίες στις οποίες γίνετε χρήση από το Web 2.0 και το διαφοροποιούν ως προς τον τρόπο λειτουργίας και παρουσίασης των ιστοσελίδων σε σχέση με το προγενέστερο Web:

- Πλούσια και διαδραστικά interfaces χρηστών (Rich Internet Applications-RIA) που κάνουν χρήση της τεχνολογία Flash, Javascript, κλπ και την Ajax, που αντιπροσωπεύει την τάση του Web 2.0 για όσο το δυνατόν μεγαλύτερη εκμετάλλευση του δικτύου. Αντί να φορτώνεται το σύνολο της σελίδα, ανανεώνονται μόνο τα δεδομένα που μεταβάλλουν όσο ο χρήστης βρίσκεται ή επανέρχεται σε αυτή όπως για παράδειγμα, στο Gmail ο υπολογισμός του διαθέσιμου αποθηκευτικού χώρου μεταβάλλεται σε πραγματικό χρόνο ενώ από όλη τη ιστοσελίδα αλλάζει μόνο την δεδομένη πληροφορία.

- Χρήση CSS (Cascading Style Sheets) για να διαχωριστούν τα δεδομένα πληροφορίας από τα δεδομένα μορφοποίησης σε μια ιστοσελίδα. Αυτό, εκτός από την οικονομία στο εύρος ζώνης του δικτύου, παρέχει και την δυνατότητα στον τρόπο παρουσίασης των δεδομένων, αφού ο χρήστης βλέπει τα δεδομένα με βάση το CSS που ο ίδιος διαθέτει όπως για παράδειγμα τα ίδια δεδομένα ανάλογα με το CSS μπορούν να παρουσιαστούν σε οθόνη υπολογιστή, απευθείας σε εκτυπωτή, σε μορφή ανάγνωσης για τυφλούς ή και να μετατραπούν σε φωνή και με χρήση κατάλληλου λογισμικού. Χρήση σημασιολογικών δεδομένων και microformats για την περιγραφή των δεδομένων που υπάρχουν στις ιστοσελίδες. Με αυτό τον τρόπο τα δεδομένα κατηγοριοποιούνται και η αναζήτησή τους πραγματοποιείται ευκολότερα και αποδοτικότερα.

- Χρήση RSS feeds ή και Atom (παραπλήσια τεχνολογία) με τα πλεονεκτήματα που προαναφέρθηκαν στην προηγούμενη ενότητα.

- Χρήση ανοικτού λογισμικού όπως για παράδειγμα Linux σαν λειτουργικό σύστημα, Apache σαν Web server, MySQL σαν βάση δεδομένων και PHP, Pearl, Python, όπως γλώσσες προγραμματισμού.

- “Ελαφρά” πρωτόκολλα δικτύου REST και SOAP που κάνουν χρήση από απλές HTTP εντολές (get, post, put, κλπ) για ανάκτηση δεδομένων από τους servers.

- Αρχιτεκτονικές SOA (Service Oriented Architecture) που δίδουν την δυνατότητα στο διαμοιρασμό και την επαναχρησιμοποίηση υπηρεσιών-εφαρμογών από διαφορετικά προγράμματα λογισμικού και SaaS (Software as a Service) όπου οι εφαρμογές είναι εγκατεστημένες σε κεντρικό server του δικτύου με σκοπό οι χρήστες να κάνουν χρήση τους μέσω browser, ανεξαρτήτως Η/Υ, τόπου, και χρονικής στιγμής.

(www.netschoolbook.gr/epimorfosi/mashup.html)

3.2.2 Deep web

Με την έναρξη μιας αναζήτησης στο Google, το Mozzila, Bing και αλλού υπάρχουν απειράριθμα αποτελέσματα που δεν εμφανίζονται στο χρήστη, είναι ένα πολύ μικρό μέρος όλων εκείνων των πληροφοριών που υπάρχουν στο web με την συντριπτική τους πλειοψηφία να παραμένει άγνωστη στους τελικούς χρηστές. Για να γίνει αντιληπτό το περίφημο «βαθύ» διαδίκτυο, το Deep Web, είναι όλο εκείνο το δίκτυο από διασυνδεδεμένα συστήματα που αποτελεί το web και αυτό που δεν βλέπουν οι κοινοί χρήστες, είναι μέχρι και 500 φορές μεγαλύτερο από αυτό που χρησιμοποιείται να εμφανίζετε στις μηχανές αναζήτησης. Η καλύτερη αναπαράσταση μάλιστα του web, είναι αυτή με την μορφή ενός παγόβουνου. Όπου το web στο οποίο γίνετε χρήση το περισσότερο που είναι το ορατό μέρος πάνω από την επιφάνεια της θάλασσας, και το Deep Web, όλος ο πάγος που βρίσκεται από κάτω.

Τι υπάρχει στο Deep Web;

Όπως αναφέρθηκε σε προηγούμενα κεφαλαία, οι μηχανές αναζήτησης εμφανίζουν αποτελέσματα κάνοντας χρήση κάποιων αλγόριθμων που «βάζουν σε λίστες» της ιστοσελίδες και λέγονται crawlers. Οι crawlers όμως δεν ανακαλύπτουν τα πάντα. Υπάρχουν «κρυφοί» πόροι στο διαδίκτυο που χονδρικά, κατατάσσονται στις ακόλουθες κατηγορίες.

- Δυναμικό περιεχόμενο: δυναμικές σελίδες στις οποίες μπορεί ένας χρήστης να έχει πρόσβαση μόνο μέσα από φόρμες στις οποίες συμπληρώνει στοιχεία.
- Μη συνδεδεμένο περιεχόμενο: σελίδες που δεν συνδέονται με άλλες σελίδες. Έτσι, τα crawlers που κάνουν χρήση οι μηχανές αναζήτησης, δεν μπορούν να τις «βρουν» από άλλες σελίδες που εξετάζουν.
- Private Web: ιστοσελίδες που είναι αναγκαίο να γίνει login από το χρήστη με username και password.
- Contextual Web: είναι οι σελίδες εκείνες, το περιεχόμενο των οποίων προσαρμόζεται ανάλογα με τον τρόπο που έχει ένας χρήστης πρόσβαση σε αυτό. Παραδείγματος χάριν, οι σελίδες εκείνες που, αν κάποιος διαθέτει πρόσβαση σε αυτές με μία διεύθυνση IP από την Ελλάδα, βλέπετε διαφορετικό περιεχόμενο από το αν θα επισκεπτόσασταν την ίδια σελίδα από μία IP των ΗΠΑ.

- Περιεχόμενο περιορισμένης πρόσβασης: ιστοσελίδες που περιορίζουν την πρόσβαση στο περιεχόμενό τους με τεχνικούς τρόπους όπως για παράδειγμα: Robots Exclusion Standards, CAPTCHAS και άλλες.
- Scriptedcontent: σελίδες που είναι διαθέσιμες μόνο από συνδέσμους που προκύπτουν από Java Script καθώς και περιεχόμενο που κατεβάζεται από Web servers μέσω Flash π.χ.
- Non-HTML/textcontent: περιεχόμενο κειμένου που διαθέτει κάποια κωδικοποίηση σε αρχεία multimedia ή συγκεκριμένα formats που δεν είναι δυνατόν να διαβάσουν οι μηχανές αναζήτησης.
- Οτιδήποτε δεν κάνει χρήση του πρότυπου HTTP/HTTPS.

Είναι δεδομένο ότι όσοι κάνουν χρήση του διαδικτύου θα έχουν συναντήσει ιστοσελίδες και περιεχόμενο το οποίο δεν είναι σε δημόσια έκθεση, ή όλοι θα έχουν συνειδητοποιήσει στη ζωή τους ότι στο διαδίκτυο πίσω από μία εφαρμογή, ή ένα βίντεο που βλέπει, υπάρχει μία υποδομή, μόνο το αποτέλεσμα της οποίας βλέπετε εσείς στον browser του.

Πώς γίνεται η αναζήτηση στο Deep Web;

Το κλειδί για να εισέρθει κάνεις σε αυτό το παράλληλο web είναι το Tor. Το Tor, είναι ουσιαστικά ένα δίκτυο από μηχανήματα εθελοντών που παρέχει την δυνατότητα ανώνυμης περιήγησης στο διαδίκτυο εφόσον ουσιαστικά, η πληροφορία που στέλνεται ή λαμβάνεται και γίνεται χρήση του από τον χρηστή, περνά από διάφορα στάδια κρυπτογράφησης και διάφορες διαδρομές, μέχρι τα δεδομένα να καταλήξουν στον προορισμό που ο χρήστης επιδιώκει, ή να πάνε στον προορισμό που αυτός έχει επιλέξει.

Η τυχαιότητα της διαδρομής που θα ακολουθήσει η πληροφορία του χρήστη, είναι εκείνη που ουσιαστικά εξασφαλίζει, τόσο την ανωνυμία στην περιήγησή του, όσο και την δυσκολία σε κάποιον κακόβουλο να παρακολουθήσει τις ηλεκτρονικές διαδρομές και της αναζητήσεις που αυτός κάνει. Η ίδια τυχαιότητα όμως και η πρόσβαση στην «υποδομή» του διαδικτύου είναι εκείνη που έχει κατασκευάσει και την τεράστια πηγή πληροφοριών στο Deep Web (https://blog.torproject.org/).

Το Tor αποτελεί την πύλη για το Deep Web καθώς το δίκτυο έχει την δυνατότητα να λειτουργεί σε όλες τις πλατφόρμες. Είτε όμως γίνετε χρήση του

λογισμικού Tor ή της ιστοσελίδα του δικτύου, θα πρέπει ο χρήστης να λαμβάνει υπόψη και μία άλλη εναλλακτική, το Tails OS, ένα λειτουργικό που έχει την δυνατότητα να γίνει χρήση του ανά πάσα στιγμή και σε οποιονδήποτε υπολογιστή, εφόσον είναι bootable είτε από DVD (ιδανικά), είτε αν το έχει ο χρήστης φορτώσει και το κρατά μαζί του σε ένα memory stick.
https://thehackernews.com/2012/05/what-is-deep-web-first-trip-into-abyss.html

Τι μπορεί να βρει κάποιος στο Deep Web;

Είναι αναγκαίο να αντιληφθεί κάποιος ότι η περιήγησή του στο Deep Web δεν έχει ως εμπειρία καμία σχέση με αυτό που απολαμβάνετε στο συμβατικό διαδίκτυο που συνήθως δραστηριοποιείται. Και ο βασικός λόγος είναι διότι δεν υπάρχει κατηγοριοποίηση στις πληροφορίες, οπότε, δεν λειτουργεί με τόσο απλοϊκό τρόπο η αναζήτηση, όπως την έχετε συνηθίσει στο Google, και στις άλλες μηχανές αναζήτησης. Ακόμη, θα πρέπει να ξεχάσει τις «καταλήξεις» που γνώριζε στις ιστοσελίδες, όπως τα .com, .gr, .gov, .org. Τα περισσότερα domains στο Deep Web έχουν την κατάληξη .onion και πάρα πολλά από τα urls θα διαπιστώσει πως δεν έχουν καμία λεκτική συνοχή.

Αν θέλετε να περιηγηθεί στο Deep Web, καλό είναι να ξεκινήσει με σελίδες που περιέχουν λίστες με περιεχόμενο στο Deep Web, όπως το Hidden Wiki παραδείγματος
χάριν http://3suaolltfj2xjksb.onion/hiddenwiki/index.php/Main_Page ή άλλες σελίδες που υπάρχουν για να του προτείνουν «χρήσιμους» συνδέσμους.

Στο Deep Web μπορεί ένα χρήστης να συναντήσει τα πάντα. Από αγορές ναρκωτικών, όπως για παραδειγμαSilkRoad μέχρι οποιαδήποτε παράνομη και παράδοξη συναλλαγή μπορεί κάποιος να φανταστεί. Πωλούνται τα πάντα, από πλαστές ταυτότητες, κακόβουλο λογισμικό και όπλα, έως δολοφονίες και κλεμμένα gadgets, ενώ σε αρκετά από τα listings μπορεί κάποιος να παρατηρήσει την ένδειξη CP (ChildPornography) ή PD (pedophile) που αποτελεί στοιχειό αποφυγής για τους αναζητητές πληροφοριών. Αν χρειάζεστε βοήθεια για να συνειδητοποιήσετε το εύρος των πραγμάτων που μπορεί κανείς να αγοράσει στο Deep Web, θα πρέπει να φανταστεί την εξάπλωση του ηλεκτρονικού εμπορίου και τι μπορεί να αγοράσει online την δεδομένη χρονική στιγμή, από το σπίτι του, με την πιστωτική του κάρτα

και να το πολλαπλάσιε το επί 500.Το νόμισμα συναλλαγής πάντως, στις περισσότερες περιπτώσεις, είναι το BitCoin.

Είναι ασφαλές, είναι ανώνυμο;

Η απάντηση στο προηγούμενο ερώτημα είναι διφορούμενη, ο χρήστης δε θα πρέπει να περιμένει από κάποιον να τον προστατεύει για το περιεχόμενο που θα αντικρύσει, τότε δεν πρέπει να μπείτε καν στην διαδικασία αναζήτησης σε αυτό το δίκτυο. Στο Deep Web δεν υπάρχουν φίλτρα που θα προσδιορίσουν το περιεχόμενο ως «καλό» ή «κακό», ούτε υπάρχει κάποια διασφάλιση για τις συναλλαγές του εκεί.

Δεν είναι ανώνυμο αν ο χρήστης εξακολουθεί να κάνει χρήση των παλιών συνήθειων του κατά την διάρκεια του σερφαρίσματος. Για παράδειγμα, αν αποφασίσει για κάποιο λόγο να κάνει κάποια μεταβολή μπορεί να εκθέσει και δώσει τον λογαριασμό gmail του που δείχνει ξεκάθαρα το όνομά του, δεν μιλάμε για ανωνυμία. Σε ζητήματα που σχετίζονται με κακόβουλο λογισμικό, δεν διατρέχει τέτοιο κίνδυνο αν κάποιος έχει πρόσβαση στο Deep Web μέσω του Tails OS κυρίως. Ακόμα καθώς το πρόγραμμα τρέχει από την RAM του υπολογιστεί του χρήστη και δεν αφήνει ίχνη στον υπολογιστή που το χρησιμοποιεί, εκτός κι αν το ζητήσει ο ίδιος οπότε, τεχνικά, δεν προκύπτει κάποιο ζήτημα αναγνώρισης της ταυτότητάς του χρηστή.

(https://www.wired.com/insights/2013/08/deep-web-the-proverbial-safe-house-for-cybercriminals/)

Υπάρχει λόγος να έχω πρόσβαση στο Deep Web;

Ο σημαντικότερος λόγος που έχει ένας χρήστης να κάνει αναζητήσεις στο Deep Web είναι αν πιστεύει πως η ανωνυμία του πλήττεται και εξακολουθεί να θέλει να κάνει χρήση του διαδικτύου για να αναζητά πληροφορίες και να επικοινωνεί, χωρίς να τον καταγράφει κάποιος γι αυτό. Ακόμα στην περίπτωση αν ανήκει σε μία πληθυσμιακή ομάδα που βρίσκεται σε κίνδυνο ή παρακολούθηση και θέλει να επικοινωνήσει ανώνυμα.

Αρνητικά στην αναζήτηση μέσα από το Deep Web θα πρέπει να απαντήσει κάποιος αν πιστεύει πως το διαδίκτυο είναι τα socialmedia και τα selfies που μοιράζετε μέσα από αυτά, μαζί με τις κάθε 5 λεπτό αναφορών για το τι σκέφτεστε, τι κάνει και με ποιον το κάνει. Ακόμα, δεν υπάρχει κανένας λόγος να ασχοληθεί

κάποιος με το Deep Web αν η χρήση που κάνει στο διαδίκτυο δεν έχει κάποιες πιο «ευρείες» αναζητήσεις. Και όχι, δεν υπάρχει κανένας λόγος να εμπλακεί κάποιος με το Deep Web, αν δεν μπορείτε να αντιμετωπίσετε την εμπειρία ενός χαοτικού περιεχομένου με ότι προέκταση μπορεί να έχει αυτό. Από ενοχλητικές ή απειλητικές συζητήσεις, μέχρι παράνομο και επικίνδυνο υλικό (https://www.wired.com/insights/2013/08/deep-web-the-proverbial-safe-house-for-cybercriminals/).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι παραπάνω αναφορές στις μηχανές αναζήτησης έγινε στο πλαίσιο μιας προσπάθειας να προσδιοριστεί η σπουδαιότητα των εργαλείων αυτών του διαδικτύου και η χρησιμότητά τους ως εργαλεία πληροφόρησης των χρηστών της βιβλιοθήκης. Μετά από αυτές οι καθημερινές δραστηριότητες πληροφόρησης γίνονται πιο απλές και εύχρηστες μέσα από τις υποδείξεις προς τους χρήστες της κατάλληλης μηχανής αναζήτησης, πέρα από τις παραδοσιακές βάσεις δεδομένων.

Μια μηχανή αναζήτησης αποτελεί μια εφαρμογή που προσφέρει την δυνατότητα αναζήτηση κειμένων και αρχείων στο Διαδίκτυο. Αποτελείται από ένα πρόγραμμα υπολογιστή που ενυπάρχει σε έναν ή περισσότερους υπολογιστές στους οποίους κατασκευάζει μια βάση δεδομένων με τις πληροφορίες που συγκεντρώνει από το διαδίκτυο, και το διαδραστικό περιβάλλον που εμφανίζεται στον τελικό χρήστη ο οποίος κάνει χρήση της εφαρμογής από άλλον υπολογιστή συνδεδεμένο στο διαδίκτυο.

Παρά την διαφορετικότητα τους σε πολλά μέρη τους οι μηχανές αναζήτησης προσδιορίζονται από ορισμένες κοινές λειτουργίες. Μια από αυτές είναι ότι οι μηχανές αναζήτησης δεν αναζητούν σε πραγματικό χρόνο τον παγκόσμιο ιστό αλλά μία βάση δεδομένων που περιλαμβάνει ορισμένα αντίγραφα ιστοσελίδων. Στις ιστοσελίδες γίνεται μια επιλογή μεταξύ δισεκατομμυρίων σελίδων στο ιντερνέτ. Η επιλογή αυτή πραγματοποιείται μέσα από την χρήση ορισμένων προγραμμάτων που καλούνται «ρομπότ» ή «αράχνες». Τα προγράμματα αυτά διατρέχουν το Διαδίκτυο σε ένα πλήθος ιστοσελίδων με στόχο την συγκέντρωση πληροφοριών με βάση ορισμένα κριτήρια. Η λειτουργία αυτή είναι αναγκαίο να γίνεται συνεχώς διότι οι ιστοσελίδες μεταβάλλονται και η βάση της μηχανής θα είναι αναγκαίο να ανανεώνονται με καινούργιες πληροφορίες. Μια μηχανή αναζήτησης έχει την δυνατότητα να διαθέτει περισσότερα από ένα «ρομπότ». Με τον εντοπισμό από τα «ρομπότ» των ιστοσελίδων τις μεταφέρουν σε ένα άλλο πρόγραμμα από το οποίο θα λάβουν δείκτες. Με το πρόγραμμα αυτό γίνεται αναγνώριση του κείμενου, οι σύνδεσμοι, και το υπόλοιπο περιεχόμενο της ιστοσελίδας και αποθηκεύεται στα αρχεία της βάσης δεδομένων. Με την πραγματοποίηση της αποθήκευσης είναι δυνατόν να πραγματοποιηθεί αναζήτηση πάνω στη βάση της μηχανής αναζήτησης.

Τέλος στις γενικότερες λειτουργίες των μηχανών αναζήτησης αποτελεί το γεγονός ότι προσφέρουν την δυνατότητα στους χρήστες να «ψάχνουν» στη βάση δεδομένων τους μέσα από ένα περιβάλλον που παρέχει αρκετές δυνατότητες αναζήτησης. Η λειτουργία αυτή σχετίζεται με αυτό που αντιμετωπίζουν οι χρήστες μέσα από το περιβάλλον της μηχανής αναζήτησης.

Σήμερα λοιπόν η συντριπτική πλειοψηφία του "αναπτυγμένου" έχει την δυνατότητα για καθημερινή πρόσβαση σε πληροφορίες που βρίσκονται διαδίκτυο. Για τον εντοπισμό της πληροφορίας στο αχανές διαδίκτυο η συμβολή των μηχανών αναζήτησης αποτελεί σημαντικό εργαλείο με τη σημαντικότερη από αυτές να είναι η Google η οποία συγκεντρώνει το 98% περίπου των συνολικών αναζητήσεων και περίπου 80% μερίδιο αγοράς παγκοσμίως και 95% παγκοσμίως για πρόσβαση μέσω mobile συσκευών.

Από την άλλη πλευρά η Microsoft έχει την δική της μηχανή αναζήτησης και αποτελεί τη δεύτερη πιο γνωστή μετά τη Google και παράλληλα παρέχει υποστήριξη και στη μηχανή αναζήτησης της Yahoo!.

Μια ακόμα σημαντική μηχανή αναζήτησης είναι αυτή της DuckDuckGo που εισήρθε σχετικά πρόσφατα στο χώρο αλλά διώχνει να έχει εισέρθει με δυναμική και κερδίζει ολοένα και μεγαλύτερο έδαφος τα τελευταία χρόνια. Αυτό που κάνει τη DuckDuckGo να ξεχωρίζει είναι η πολιτική της ως σε σχέση με τα προσωπικά δεδομένα των χρηστών. Με βάση αυτή δεν αποθηκεύει καμία πληροφορία που να σχετίζεται με τις online δραστηριότητες των χρηστών και τις προσωπικές τους αναζητήσεις.

Υπάρχουν και άλλες σημαντικές μηχανές αναζήτησης, με ποιο στοχευόμενη βάση δεδομένων, αλλά και μηχανές αναζητήσεις με ποιο ειδικά ενδιαφέροντα. Αυτό που αποτελεί σημαντικό στοιχείο είναι ότι ο ανταγωνισμός μεταξύ των εταιριών που αναπτύσσουν τέτοια λογισμικά για μηχανές αναζητήσεις έχει συντελέσει στην αύξηση των δυνατοτήτων του συνόλου των μηχανών αυτών με αποτέλεσμα την παροχή σημαντικών εργαλείων στους χρηστές του διαδικτύου.

Με μια πιθανή αναζήτηση ενός χρήστη στο Google ή το Bing, λαμβάνει παρά πολλά αποτελέσματα, παρόλα αυτά δεν παύει να αποτελούν είναι ένα πολύ μικρό μέρος όλων εκείνων των πληροφοριών που πλημυρίζουν το διαδίκτυο με την

συντριπτική τους πλειοψηφία να είναι άγνωστη. Η ονομασία που έχει περιβάλει την άγνωστη πλευρά του διαδικτύου είναι αυτή του περίφημου «βαθύ» διαδικτύου, το Deep Web, είναι όλο εκείνο το δίκτυο από διασυνδεδεμένα συστήματα που αποτελεί το web και αυτό που δεν φαίνεται οι κοινοί χρήστες, το μέγεθος του αγγίζει σε μέγεθος τις 500 φορές μεγαλύτερο απ' ότι χρησιμοποιείται ή φαίνεται στις μηχανές αναζήτησης.

Μέσα από την περιήγησή σας στο Deep Web δεν έχει ως εμπειρία καμία σχέση με αυτό που βρίσκει ένας χρήστης στο συμβατικό διαδίκτυο που συνήθως κάνει χρήση. Και ο βασικός λόγος είναι γιατί δεν υπάρχει κατηγοριοποίηση στις πληροφορίες, οπότε, δεν λειτουργεί με τόσο απλοϊκό τρόπο η αναζήτηση, όπως στην Google, ή το Bing. Ακόμα δεν υπάρχουν «καταλήξεις» που γνωρίζατε στις ιστοσελίδες, όπως τα .com, .gr, .gov, .org. Τα περισσότερα domains στο Deep Web έχουν την κατάληξη .onion και πάρα πολλά από τα urls που δεν έχουν καμία λεκτική συνοχή.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Πούλος Μ.(2015).Σημασιολογική Επεξεργασία της Πληροφορίας, Περί Σημασιολογίας,https://repository.kallipos.gr/bitstream/11419/2854/5/00_master_document-KOY.pdf, σελ 42-44.
- Asnicar F. and Tasso C., (1997). ifWeb: A Protoype of User Model-Based Intelligent Agent for Documentation Filtering and Navigation in the World Wide Web. In Proceedings of the 6th International Conference on User Modeling.
- Barzilay R.,(1997). Lexical Chains for Summarization. Master's Thesis, BenGurion University.
- Brin S. and Page L., (1998)."The anatomy of a large-scale hypertextal Web search engine", *Computer Networks and ISDN Systems* Vol. 30(1-7).
-
- Netcraft LTD, "December 2017 Web Server Survey". Available: http://news.netcraft.com/archives/2008/12/24/december_2008_web_server_survey.html via the Internet. Last access: December 2017.
- Official Google Blog, "We knew the Web was big". Available:http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html via the Internet. Last access: December 2008.
- Spink A., Wolfram D., Jansen M.B.J. andSaracevic T.,(2001). "Searching the web: The public and their queries", *Journal of the American Society for Information Science and Technology* Vol. 52(3), pp.226-234.

Διαδικτυακές πήγες

- http://el.wikipedia.org/wiki/Google
- http://archive.is/20120712130315/http://docs.yahoo.com/info/misc/history.html#selection-8.5-97.961
- http://en.wikipedia.org/wiki/Bing_search
- https://www.crunchbase.com/organization/ask-com#/entity).
- http://www.wolfram.com/company/background.html?source=nav
- http://www.yippyinc.com/company-0
- https://duckduckgo.com/privacy)

- http://infospace.com/terms/privacy.html)
- http://www.goonline.gr/ebusiness/specials/article.html?article_id=22):
- http://www.robby.gr
- http://www.trinity.gr
- http://search.forthnet.gr
- http://anazitisis.gr
- http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html
- http://www.goonline. gr/ebusiness/specials
- article.html?article_id=232
- http://boston.lti.cs.cmu.edu/Data/clueweb09/
- http://www.webmasterslife.gr/search-engine-optimization/73.html
- http://complexnt.blogspot.gr/2012/04/hyperlink-induced-topic-search-hits.html
- http://www-digUb.stanford.edu/Jestbed/doc2/SDLIP/.
- oreilly.com/web2/archive/what-is-web-20.html
- www.broadband.gr/opencms/sites/Broadband/News/news071228c/
- https://learn20.wikispaces.com/Web+2.0%28%CE%BA%CE%B5%CE%AF%CE%BC%CE%B5%CE%BD%CE%BF%29
- skull.gr/blog/web-20
- www.netschoolbook.gr/epimorfosi/mashup.html
- https://thehackernews.com/2012/05/what-is-deep-web-first-trip-into-abyss.html
- http://3suaolltfj2xjksb.onion/hiddenwiki/index.php/Main_Page
- https://www.wired.com/insights/2013/08/deep-web-the-proverbial-safe-house-for-cybercriminals/
- https://www.wired.com/insights/2013/08/deep-web-the-proverbial-safe-house-for-cybercriminals/
- (http://www.dcs.bbk.ac.uk/~dell/teaching/cc/book/ditp/ditp_ch4.pdf).