ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΠΑΤΡΩΝ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Αριθμός 1318

ΤΕΧΝΙΚΕΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

ΣΠΟΥΔΑΣΤΗΣ:
ΘΕΟΔΩΡΟΥ ΑΝΤΡΕΑΣ

ΕΙΣΗΓΗΤΗΣ:
ΚΑΡΕΛΗΣ ΔΗΜΗΤΡΙΟΣ

ΠΑΤΡΑ ΙΟΥΝΙΟΣ 2013

ΠΡΟΛΟΓΟΣ

Η παρούσα πτυχιακή εργασία ασχολείται με θέματα που αφορούν τη λειτουργία των μηχανών αναζήτησης και ιδιαίτερα την εξαγωγή μεθόδων βελτιστοποίησης της κατάταξης των σελίδων. Στο πρώτο κεφάλαιο μελετώνται αναλυτικά οι τεχνολογίες και λειτουργίες των μηχανών αναζήτησης που σχετίζονται με την ανίχνευση και την ευρετηρίαση των σελίδων του Παγκόσμιου Ιστού, καθώς και την επεξεργασία των ερωτημάτων αναζήτησης. Στη συνέχεια, στο δεύτερο κεφάλαιο αναλύονται εκείνοι οι παράγοντες που αφορούν την εσωτερική οργάνωση και μορφοποίηση της ιστοσελίδας και του εξυπηρετητή φιλοξενίας αυτής και, βάσει συμπερασμάτων, παρατίθενται οι αντίστοιχες μέθοδοι βελτιστοποίησης των παραγόντων αυτών.

ΠΕΡΙΛΗΨΗ

Οι αναζητήσεις στις μηχανές αναζήτησης αγγίζουν τον αριθμό πολλών εκατομμυρίων καθημερινά, με τους χρήστες να απαιτούν την ταχύτατη εμφάνιση των αποτελεσμάτων που τους ενδιαφέρουν. Επίσης η κατασκευή μιας ιστοσελίδας αποτελεί πλέον μια απλή διαδικασία ακόμα και για αυτούς που δεν έχουν ιδιαίτερες γνώσεις επί του αντικειμένου, εκτοξεύοντας έτσι τον αριθμό των ιστοσελίδων στο Διαδίκτυο. Από την άλλη μεριά, η συνεχώς αυξανόμενη συμμετοχή και προβολή των επιχειρήσεων και των οργανισμών στο Διαδίκτυο έχει επιφέρει την σημαντική αύξηση του ανταγωνισμού για την κατάταξη των ιστοσελίδων τους στις περιορισμένες και πολύτιμες θέσεις των πρώτων σελίδων αποτελεσμάτων αναζήτησης για σχετικούς όρους, επιβάλλοντας ταυτόχρονα το δυναμικό χαρακτήρα μεταβολής της συμπεριφοράς των μηχανών αναζήτησης. Η παρούσα πτυχιακή εργασία ασχολείται με θέματα που αφορούν τη λειτουργία των μηχανών αναζήτησης και ιδιαίτερα την εξαγωγή μεθόδων βελτιστοποίησης της κατάταξης των σελίδων. Αρχικά μελετώνται αναλυτικά οι τεχνολογίες και λειτουργίες των μηχανών αναζήτησης που σχετίζονται με την ανίχνευση και την ευρετηρίαση των σελίδων του Παγκόσμιου Ιστού, καθώς και την επεξεργασία των ερωτημάτων αναζήτησης. Στη συνέχεια, μελετώνται αναλυτικά εκείνοι οι παράγοντες που αφορούν την εσωτερική οργάνωση και μορφοποίηση της ιστοσελίδας και του εξυπηρετητή φιλοξενίας αυτής και, βάσει συμπερασμάτων, καταστρώνονται οι αντίστοιχες μέθοδοι βελτιστοποίησης των παραγόντων αυτών.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Το μέγεθος του Internet είναι δεδομένο ότι είναι τεράστιο και αυξάνει με εκπληκτικούς ρυθμούς. Υπό αυτές τις συνθήκες, η εύρεση στοιχείων αποτελεί μία εξαιρετικά δύσκολη υπόθεση, που θα παρέμενε δύσκολη εάν δεν υπήρχαν εργαλεία όπως οι μηχανές αναζήτησης.

Ένα από τα σημαντικότερα χαρακτηριστικά του Internet είναι η ευκολία που παρέχει στην είσοδο οποιασδήποτε πληροφορίας, επιτρέποντας στους χρήστες του να εισάγουν στοιχεία για κάθε θέμα. Τα στοιχεία αυτά είναι συνήθως ελεύθερα διαθέσιμα σε όλους τους χρήστες, καθιστώντας έτσι το Internet στο σύνολό του μία μοναδική πηγή πληροφόρησης και εύρεσης στοιχείων, που παρόμοιά της δεν υπήρξε ποτέ μέχρι τώρα στην πορεία της ανθρωπότητας. Η ραγδαία αύξηση της χρήσης του World Wide Web, αλλά και των υπόλοιπων υπηρεσιών του δικτύου, έδωσε στους χρήστες τη δυνατότητα να αποκτήσουν εύκολη πρόσβαση στην πληροφορία, αλλά παράλληλα και τη δυνατότητα παροχής στο δίκτυο όλων όσων αυτοί θεωρούν κατάλληλα.

Ενώ όμως η πληθώρα πληροφοριών λογικά θα έπρεπε να είναι ευεργετική για τους χρήστες, οι οποίοι έχουν πλέον στη διάθεσή τους έναν τεράστιο όγκο στοιχείων, αυτή η ίδια πληθώρα προξενεί ένα σημαντικό πρόβλημα, που δεν είναι άλλο από το ότι οι χρήστες αδυνατούν τις περισσότερες φορές να εντοπίσουν τα σημεία εκείνα του δικτύου που περιέχουν τις πληροφορίες τις οποίες αυτοί χρειάζονται. Για παράδειγμα, έστω ότι κάποιος χρήστης αναζητεί πληροφορίες για ένα μουσικό συγκρότημα. Πιθανότατα, αρκετοί χρήστες από όλο το Internet θα έχουν συγκεντρωμένες πληροφορίες για το συγκεκριμένο συγκρότημα σε διάφορες σελίδες του Web ή ενδεχομένως να υπάρχουν σχετικές πληροφορίες από δισκογραφικές εταιρείες κ.λ.π. Το πρόβλημα που προκύπτει για τον ενδιαφερόμενο χρήστη είναι πώς θα εντοπίσει τις πληροφορίες που αυτός χρειάζεται, πώς δηλαδή θα μάθει τις σελίδες και τα sites που περιέχουν αυτό που αναζητά.

Μολονότι όλον και κάποιον τρόπο μπορεί να σκεφθεί ένας χρήστης για να το επιτύχει, κανένας τρόπος δεν μπορεί να συγκριθεί σε πληρότητα, ταχύτητα και αποτελεσματικότητα με την χρήση των περίφημων **μηχανών αναζήτησης** (search engines) του World Wide Web.

Οι μηχανές αναζήτησης είναι από τα λίγα εργαλεία του Internet που προσπαθούν να βάλουν τάξη και να προσφέρουν διέξοδο σε όσους αναζητούν μία πληροφορία στο Δίκτυο αλλά δεν γνωρίζουν πού ακριβώς θα την βρουν. Τυπικά, μία μηχανή αναζήτησης διαθέτει μία βάση δεδομένων με καταγεγραμμένες διευθύνσεις του Internet, στις οποίες ο χρήστης μπορεί να βρει συγκεκριμένα στοιχεία που τον ενδιαφέρουν. Ο χρήστης αναζητεί αυτό που θέλει με βάση κάποια συγκεκριμένα κριτήρια και η μηχανή αναζήτησης του παρουσιάζει τις διευθύνσεις εκείνες στις οποίες μπορεί αυτός να βρει σχετικές πληροφορίες.

Γενικά, μία μηχανή αναζήτησης μπορεί να περιέχει διευθύνσεις από όλες τις υπηρεσίες του Internet, όπως FTP, World Wide Web, Usenet, Telnet κ.λ.π. Οι περισσότερες όμως μηχανές αναζήτησης περιορίζονται στην “καταλογοποίηση” των πληροφοριών εκείνων που μπορούν να προβληθούν μόνο μέσω του World Wide Web, δηλαδή με βάση το πρωτόκολλο HTTP κατά κύριο λόγο, ενώ ορισμένες υποστηρίζουν επιπλέον FTP και Gopher διευθύνσεις του δικτύου. Πρέπει να καταστεί σαφές πάντως ότι σε κάθε περίπτωση η μηχανή αναζήτησης δεν έχει καταχωρημένο το περιεχόμενο αλλά μόνο τις διευθύνσεις και ό,τι άλλο αυτή χρειάζεται για να μπορέσει να εξυπηρετήσει τους χρήστες. Στην πράξη, δηλαδή, μία μηχανή αναζήτησης είναι ένα τεράστιο αρχείο με συνδέσμους (links) οι οποίοι οδηγούν σε διάφορους εξυπηρετητές, σελίδες Web, αρχεία κ.λ.π.

Στο Internet υπάρχουν αρκετές μηχανές αναζήτησης, οι οποίες τις περισσότερες φορές ξεκίνησαν από πειραματικά ερευνητικά προγράμματα (projects) και εξελίχθηκαν σε ολόκληρες εταιρείες, ενώ από πλευράς χρήσης εξυπηρετούν χιλιάδες χρήστες καθημερινά. Ενδεικτικά αναφέρονται εδώ οι πιο γνωστές από αυτές, όπως είναι η Yahoo, η Lycos, η InfoSeek, η Web Crawler κ.ά. Συνήθως, η παροχή των προσφερόμενων υπηρεσιών γίνεται δωρεάν, αν και ορισμένες μηχανές επιβάλλουν κάποιους περιορισμούς στη δωρεάν χρήση διαθέτοντας και πρόσβαση επί πληρωμή.

ΚΕΦΑΛΑΙΟ 1

ΤΙ ΕΙΝΑΙ ΚΑΙ ΠΩΣ ΛΕΙΤΟΥΡΓΕΙ ΜΙΑ ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ

1.1 Πως λειτουργεί μία μηχανή αναζήτησης

Το ερώτημα που λογικά προκύπτει είναι πώς εισάγονται οι διευθύνσεις σε κάθε μηχανή αναζήτησης, έτσι ώστε να δημιουργηθεί η βάση διευθύνσεων στην οποία κάνουν αναζητήσεις οι χρήστες. Η συνηθέστερη διαδικασία για την εισαγωγή των διευθύνσεων είναι οι ίδιοι οι κάτοχοι των σελίδων να ενημερώνουν τη μηχανή αναζήτησης για την ύπαρξη των σελίδων τους. Ακολούθως, η μηχανή αναζήτησης, αφού κάνει όλους τους απαραίτητους ελέγχους για τη διεύθυνση που δηλώθηκε, την καταχωρίζει στη βάση της. Φυσικά, κάθε μηχανή μπορεί να δέχεται διευθύνσεις μόνο του τύπου εκείνου που έχει καθορίσει ο κατασκευαστής της, λόγου χάρη HTTP, FTP, Gopher κ.λ.π. ενώ ο τρόπος διαχείρισης της διεύθυνσης-από τη στιγμή που αυτή θα εισαχθεί-διαφέρει από μηχανή σε μηχανή.

Εκτός από την εισαγωγή από τους ίδιους τους κατόχους των σελίδων ή λοιπών διευθύνσεων, ένας άλλος τρόπος ενημέρωσης της βάσης διευθύνσεων μίας μηχανής αναζήτησης είναι η έρευνα από την ίδια τη μηχανή στον Web ή σε άλλα μέρη του δικτύου Internet για εύρεση στοιχείων. Με τη διαδικασία αυτή, η μηχανή αναζήτησης συνδέεται με τους διάφορους υπολογιστές-εξυπηρετητές του δικτύου και καταγράφει τα δεδομένα τους, ανάλογα με τον σχεδιασμό της. Με τον τρόπο αυτό, δεν απαιτείται η συμμετοχή των χρηστών-κατόχων των σελίδων Web, η ενημέρωση γίνεται πιο άμεσα, ενώ το περιεχόμενο της βάσης διευθύνσεων είναι μεγαλύτερο και πληρέστερο απ'ότι θα ήταν εάν η βάση της μηχανής αναζήτησης ενημερωνόταν μόνο με πρωτοβουλία των χρηστών του δικτύου.

Από τη στιγμή που υπάρχουν οι διευθύνσεις στη βάση της μηχανής αναζήτησης, ο χρήστης μπορεί να αναζητήσει με βάση κάποιο θέμα τις διευθύνσεις που αναφέρονται σε αυτό. Ανάλογα με τη μηχανή, η αναζήτηση μπορεί να πραγματοποιηθεί είτε δίνοντας ο χρήστης κάποια έκφραση είτε μέσω κάποιας ιεραρχικής κατηγοριοποίησης των περιεχομένων. Στην πρώτη περίπτωση, ο χρήστης εισάγει μία έκφραση αναζήτησης (search expression ή string), είτε απλή είτε πιο σύνθετη με χρήση λογικών τελεστών, οπότε η μηχανή αναζητά στη βάση της σε ποιες ακριβώς διευθύνσεις υπάρχει η έκφραση αυτή. Ανάλογα με τη μηχανή, ο αλγόριθμος αναζήτησης που χρησιμοποιείται θα δώσει περισσότερο ή λιγότερο σχετικές διευθύνσεις. Στη δεύτερη περίπτωση, οι διευθύνσεις έχουν ήδη κατηγοριοποιηθεί από τη μηχανή σε γενικές ιεραρχικές κατηγορίες, οπότε ο χρήστης, χρησιμοποιώντας την ιεραρχία αυτή, οδηγείται στις διευθύνσεις που περιέχουν αυτό που αναζητά. Βέβαια, η περίπτωση αυτή απαιτεί να γνωρίζει ο χρήστης τι ακριβώς

ζητά. Τέλος, πρέπει να επισημανθεί ότι ο ένας τρόπος αναζήτησης δεν αναιρεί τον άλλο, αφού μπορούν κάλλιστα να υπάρχουν ταυτόχρονα και οι δύο τρόποι.

1.2 Οι δημοφιλέστερες μηχανές αναζήτησης

Πέρα από την εισαγωγή στις μηχανές αναζήτησης του World Wide Web και του τρόπου λειτουργίας τους, αναφέρονται ακολούθως οι πιο δημοφιλείς και εύχρηστες από τις μηχανές αυτές.

1.2.1 Lycos

Μία από τις γνωστότερες μηχανές αναζήτησης του Internet είναι ο Lycos. Η μηχαχή αυτή είναι μάλιστα τόσο πλούσια σε περιεχόμενο-διευθύνσεις που οι κατασκευαστές της την χαρακτηρίζουν ως κατάλογο του Internet, αφού σύμφωνα με στοιχεία τους περιλαμβάνει άνω του 90% των διευθύνσεων του Web! Για την εμπορική εκμετάλλευση της μηχανής αυτής δημιουργήθηκε στα τέλη Ιουνίου του 1995 η εταιρεία Lycos Inc., ενώ τεχνολογικά αποτελεί έργο του Dr. Michael Mauldin στο Πανεπιστήμιο Carnegie Mellon.

Το σημαντικότερο χαρακτηριστικό της μηχανής αυτής είναι ότι εκτός από τη δυνατότητα που παρέχει στους χρήστες να καταχωρούν οι ίδιοι τις διευθύνσεις των σελίδων τους, ο ίδιος ο Lycos αναζητεί καθημερινά μέσω ειδικών προγραμμάτων διευθύνσεις, τις οποίες και καταχωρεί στη βάση του. Τα προγράμματα αυτά, τα οποία ονομάζονται spiders, αναζητούν HTTP, FTP και Gopher sites (τις τρεις υπηρεσίες που καλύπτει ο Lycos) και είναι αυτά στα οποία οφείλεται το μεγάλο ποσοστό διευθύνσεων της μηχανής αυτής.
Φυσικά, η όλη διαδικασία είναι ιδιαίτερα πολύπλοκη, ενώ από τη στιγμή που συνδεθεί με κάποιο site ακολουθείται μία διαδικασία ελέγχου του εξυπηρέτη υπολογιστή. Είναι τέτοια η ποσότητα που συγκεντρώνεται στη βάση του Lycos από τη διαδικασία αυτή, ώστε ο Lycos είναι με τεράστια διαφορά η μεγαλύτερη μηχανή αναζήτησης και μάλιστα με βάση γενικώς αποδεκτά στοιχεία. Επίσης, η διαδικασία εύρεσης και ελέγχου των διευθύνσεων είναι καθημερινή, οπότε ο κατάλογος της βάσης διατηρείται συνεχώς ενημερωμένος με νέα στοιχεία και διευθύνσεις. Η μηχανή αναζήτησης Lycos παρέχει δυνατότητα δωρεάν πρόσβασης και εξυπηρέτησης των χρηστών μέσω του WWW στη διεύθυνση: http://www.lycos.com.

1.2.2 Yahoo

Η Yahoo είναι επίσης μία από τις πιο γνωστές μηχανές αναζήτησης. Αυτή χρησιμοποιεί έναν κατάλογο, θεματικά ταξινομημένο, επιτρέποντας έτσι στους χρήστες να αναζητήσουν διευθύνσεις ακολουθώντας έναν ιεραρχικό κατάλογο θεμάτων. Εκτός από την χρήση του καταλόγου, υπάρχει και η δυνατότητα για αναζήτηση λέξεων με χρήση λογικών τελεστών. Το περιεχόμενο του καταλόγου προέρχεται από τους χρήστες του Internet, οι οποίοι καταχωρούν τις διευθύνσεις των σελίδων τους σε αυτόν. Η μηχανή αυτή αναζήτησης παρέχει δωρεάν υπηρεσίες στους χρήστες μέσω του WWW στη διεύθυνση: http://www.yahoo.com.

1.2.3 Alta Vista

Μία από τις νεότερες και ιδιαίτερα αξιόλογες μηχανές αναζήτησης είναι η Alta Vista. Ξεκίνησε ως ερευνητικό πρόγραμμα (project) από τα εργαστήρια της εταιρείας Digital Research, ενώ η επίσημη λειτουργία της έγινε στις 15 Δεκεμβρίου του 1995. Ήδη μέσα στις τρεις πρώτες εβδομάδες της λειτουργίας της εξυπηρετούσε πάνω από 2 εκατομμύρια αναζητήσεις την ημέρα, ενώ εντυπωσιακός είναι και ο hardware εξοπλισμός που αυτή διαθέτει.

Ως μηχανή αναζήτησης, η Alta Vista ανήκει στην κατηγορία των μηχανών εκείνων που, εκτός από τις καταχωρίσεις σελίδων από τους ίδιους τους χρήστες, αναζητούν μόνες τις διευθύνσεις των σελίδων στον Web. Επιπλέον, υποστηρίζει και αναζητήσεις σε ομάδες νέων (newsgroups) του Usenet μέσα από τον τοπικό της server. Πέρα από την πλούσια βάση διευθύνσεων που διαθέτει, η οποία βρίσκεται στα επίπεδα του Lycos, παρέχει εξαιρετικές δυνατότητες αναζήτησης με την υποστήριξη ενός πλήρους συνόλου λογικών τελεστών. Ο χρήστης έχει έτσι τη δυνατότητα να κάνει απλές αναζητήσεις όπως σε όλες τις μηχανές αναζήτησης ή, εάν αυτός επιθυμεί κάτι πιο προχωρημένο, να χρησιμοποιήσει λογικούς τελεστές συντάσσοντας κάποιες πολύπλοκες “ερωτήσεις” (advanced queries) προς τη βάση διευθύνσεων της μηχανής. Η Alta Vista παρέχει δωρεάν υπηρεσίες στους χρήστες μέσω του WWW στη διεύθυνση : http://www.altavista.com.

1.2.4 Web Crawler

Είναι η μηχανή αναζήτησης που παρέχεται από τη γνωστή αμερικανική εταιρεία on-line υπηρεσιών America On Line. Διαθέτει μία σχετικά μικρή βάση διευθύνσεων, η οποία προέρχεται από καταχωρίσεις χρηστών και εν συνεχεία έλεγχο από την ίδια τη μηχανή. Λόγω του μικρού μεγέθους της βάσης, οι αναζητήσεις είναι σχετικά γρήγορες, οπότε η μηχανή αυτή αποτελεί την καλύτερη ίσως επιλογή των χρηστών όταν η ταχύτητα αναζήτησης είναι ένας κρίσιμος παράγοντας. Επίσης, δεν παρέχει ιδιαίτερες δυνατότητες ελέγχου της αναζήτησης αλλά μόνο τις στοιχειώδεις. Η Alta Vista προσφέρει δωρεάν υπηρεσίες στους χρήστες μέσω του WWW στη διεύθυνση :http://webcrawler.com.

1.2.5 Excite

Η Excite αποτελεί μία από τις νεότερες εταιρείες που δραστηριοποιούνται στο χώρο των μηχανών αναζήτησης. Παρέχει δωρεάν υπηρεσίες και προσφέρει αναζητήσεις σε σελίδες του Web και τις ομάδες νέων του Usenet. Η βάση διευθύνσεων της μηχανής είναι ικανοποιητική και περιλαμβάνει αρκετές σελίδες. Το περιβάλλον επικοινωνίας (interface) μεταξύ της μηχανής και του χρήστη είναι επίσης ικανοποιητικό ενώ οι δυνατότητες σύνταξης Queries βρίσκονται σε μέσο επίπεδο. Η Excite προσφέρει τις υπηρεσίες της στους χρήστες μέσω του WWW στη διεύθυνση: http://www.excite.com.

1.2.6 Google

Η **Google** είναι μια από τις μεγαλύτερες εταιρείες διαδικτυακών υπηρεσιών. Η λειτουργία του ξεκίνησε στης 27 Σεπτεμβρίου του 1998. Ο στόχος της είναι *να οργανώσει όλες τις πληροφορίες του κόσμου και να τις κάνει παγκόσμια διαθέσιμες*. Το Google ξεκίνησε σαν μια κολεγιακή εργασία από τον Λάρρυ Πέιτζ και τον Σεργκέι Μπριν το 1996 για μια μηχανή αναζήτησης. Σήμερα η μηχανή αναζήτησης google είναι μια από τις δημοφιλέστερες, και οι φράσεις «*κάνω **google***», «*γκουγκλάρω*», «*γκουγκλίζω*»,«***google it***» ή «*μπαίνω στον γκούγκλη*» είναι συνώνυμες με το «ψάχνω για πληροφορίες στο Διαδίκτυο». Αντίστοιχα, στην αγγλική γλώσσα το ρήμα "***to google***" έχει αποκτήσει πλέον ταυτόσημη έννοια με το ρήμα «αναζητώ», και, πρόσφατα, το ίδιο ρήμα προστέθηκε στο αγγλικό λεξικό Merriam-Webster με όλα τα παράγωγά του (***to google*** > ***googling*** > ***googled***) [1].

Η λέξη "Google" προήλθε από αναγραμματισμό της λέξης ***Googol***, η οποία εκφράζει μαθηματικό όρο (τον οποίο εισήγαγε ο Milton Sirotta) και σημαίνει το «1 ακολουθούμενο από 100 μηδενικά». Με τον όρο αυτόν η Google επιθυμεί να υποδηλώσει την αποστολή της εταιρίας να οργανώσει το τεράστιο πλήθος πληροφοριών του Ίντερνετ.

1.3 Κατηγορίες Μηχανών Αναζήτησης

Υπάρχουν μηχανές αναζήτησης που βασίζουν τη λειτουργία τους σε μηχανισμούς ανίχνευσης σελίδων (crawler – based μηχανές) αλλά και μηχανές που λειτουργούν με χειροκίνητους καταλόγους (human – powered directories). Η βασική διαφορά τους εντοπίζεται στο ότι οι μηχανές της δεύτερης κατηγορίας απαιτούν την ύπαρξη και συμμετοχή του ανθρώπινου παράγοντα για την επιλογή, καταχώρηση και κατάταξη των εγγράφων στους καταλόγους.

1.3.1 Crawler – based μηχανές

Αυτές οι μηχανές αναζήτησης περιλαμβάνουν λειτουργίες που τους επιτρέπουν να παρέχουν τα σχετικά αποτελέσματα όταν οι χρήστες χρησιμοποιούν το σύστημά τους για την εύρεση πληροφοριών. Αυτές οι λειτουργίες είναι οι εξής:

1. **Ανίχνευση του Παγκόσμιου Ιστού (Web Crawling)**

   Οι μηχανές αναζήτησης εκτελούν αυτοματοποιημένα προγράμματα, που ονομάζονται «bots» ή «αράχνες» (spiders), τα οποία χρησιμοποιούν τη δομή υπερσυνδέσμων του Ιστού για να ανιχνεύσουν σελίδες και έγγραφα.

2. **Ευρετηρίαση εγγράφων (Indexing)**

   Όταν ανιχνεύεται μία σελίδα, τα δεδομένα που περιέχει μπορούν να αποθηκευτούν σε μία τεράστια βάση δεδομένων από έγγραφα που όλα μαζί αποτελούν το ευρετήριο (index) μίας μηχανής αναζήτησης. Το ευρετήριο αυτό χρειάζεται να είναι αυστηρά οργανωμένο, έτσι ώστε οι αιτήσεις χρηστών να μπορούν να εξυπηρετηθούν σε μερικά κλάσματα του δευτερολέπτου.

3. **Επεξεργασία ερωτημάτων (Query Processing)**

   Όταν πραγματοποιείται μία αίτηση για πληροφορία, η μηχανή αναζήτησης ανακτά από το ευρετήριό της όλα τα έγγραφα που πιθανώς αντιστοιχούν στο ερώτημα. Η αντιστοιχία ορίζεται εάν οι όροι ή η φράση βρίσκεται στην σελίδα, με τρόπο που έχει καθοριστεί από τον χρήστη. Για παράδειγμα, μία αναζήτηση για *ΤΕΙ Πατρών- Τμήμα Ηλεκτρολογίας*, στην ελληνική εκδοχή της μηχανής Google, επιστρέφει περίπου 35.300 αποτελέσματα, ενώ η ίδια αναζήτηση με εισαγωγικά (“*ΤΕΙ Πατρών- Τμήμα Ηλεκτρολογίας* ”) επιστρέφει 2590 μόλις αποτελέσματα. Στο πρώτο σύστημα, ευρέως γνωστό ως λειτουργία «Findall» (εύρεση όλων), η μηχανή αναζήτησης της Google επέστρεψε όλα τα έγγραφα που περιελάμβαναν τους όρους «ΤΕΙ», «Πατρών», «Τμήμα», «Ηλεκτρολογίας» συμπεριλαμβανομένων όλων των πτώσεων και των καταλήξεων των παραπάνω λέξεων. Στη δεύτερη αναζήτηση, η μηχανή επέστρεψε μόνο τα έγγραφα που περιελάμβαναν την ακριβή φράση “*ΤΕΙ Πατρών- Τμήμα Ηλεκτρολογίας*

4. **Κατάταξη αποτελεσμάτων (Ranking)**

   Όταν η μηχανή καθορίσει ποια αποτελέσματα αντιστοιχούν σε ένα ερώτημα, ο αλγόριθμος της μηχανής εκτελεί υπολογισμούς σε κάθε ένα αποτέλεσμα για να καθορίσει τον βαθμό σχετικότητάς του με το δεδομένο ερώτημα. Οι μηχανές με τον τρόπο αυτό και με κριτήριο το βαθμό σχετικότητας, κατατάσσουν τα έγγραφα στις σελίδες αποτελεσμάτων, με φθίνουσα σειρά ταξινόμησης.

1.3.2 Human – powered κατάλογοι

Μία τέτοια μηχανή, όπως το Open Directory Project, βασίζει τη λειτουργία της στον ανθρώπινο παράγοντα για τις καταχωρήσεις της. Η διαδικασία της εγγραφής στον κατάλογο περιλαμβάνει την καταχώρηση μιας σύντομης περιγραφής για ολόκληρη την ιστοσελίδα, είτε από τον ενδιαφερόμενο κάτοχο του υπό καταχώρηση ιστοχώρου είτε από τους συντάκτες που αξιολογούν μια ιστοσελίδα, ενώ συνήθως πραγματοποιείται επί πληρωμή. Ένα αίτημα αναζητεί αντιστοιχίες μόνο στις περιγραφές που έχουν καταχωρηθεί. Τροποποιήσεις στην ιστοσελίδα που έχει ήδη καταχωρηθεί σε τέτοιες μηχανές δεν προκαλούν αντίστοιχες αλλαγές στην καταχώρηση καθ' αυτήν. Οι μέθοδοι βελτιστοποίηση, δεν επιδρούν στις εγγραφές του καταλόγου μιας human – powered μηχανής, ενώ εξαιρείται από αυτόν τον κανόνα η περίπτωση όπου μία ιδιαίτερα καλή ιστοσελίδα με εξαιρετικό περιεχόμενο ενδέχεται να αξιολογηθεί και να καταχωρηθεί από τους συντάκτες, χωρίς ο κάτοχος ή ο διαχειριστής της να παρουσιάσει κάποιο ενδιαφέρον και να προτείνει την καταχώρησή της.

1.4 Ανίχνευση του Παγκόσμιου Ιστού (Web Crawling)

Ο ανιχνευτής Διαδικτύου, παγκοσμίως γνωστός ως Web Crawler, είναι ένα υπολογιστικό πρόγραμμα το οποίο εξετάζει και περιηγείται στον Παγκόσμιο Ιστό με ένα μεθοδικό, αυτοματοποιημένο τρόπο. Η διαδικασία αυτή ονομάζεται «Ανίχνευση Ιστού» (web crawling ή spidering). Οι ανιχνευτές Ιστού δημιουργούν ένα αντίγραφο από όλες τις σελίδες που έχουν επισκεφθεί για τη μελλοντική επεξεργασία του από μία μηχανή αναζήτησης που θα ευρετηριάσει τις μεταφορτωμένες σελίδες και, ως εκ τούτου, θα παρέχει ταχείες αναζητήσεις. Οι ανιχνευτές μπορούν επίσης να χρησιμοποιηθούν για έλεγχο των συνδέσμων ή επικύρωση του HTML κώδικα.

Ο ανιχνευτής Ιστού δηλαδή αποτελεί τύπο διαδικτυακού ρομπότ, ή πράκτορα λογισμικού. Γενικά, ξεκινάει με μια λίστα από URLs, που ονομάζονται «σπόροι». Καθώς ο ανιχνευτής επισκέπτεται τις τοποθεσίες αυτές, αναγνωρίζει υπερσυνδέσμους στην σελίδα και τους προσθέτει στη λίστα των URLs που προορίζεται να επισκεφθεί, που ονομάζεται «σύνορο ανίχνευσης» (crawl frontier). Ο ανιχνευτής επισκέπτεται τις διευθύνσεις της λίστας αυτής αναδρομικά, σύμφωνα με ένα σύνολο πολιτικών, τις λεγόμενες πολιτικές ανίχνευσης

1.4.1 Πολιτικές ανίχνευσης

Το μεγάλο εύρος του Διαδικτύου, οι ταχύτατοι ρυθμοί με τους οποίους οι συνθήκες και οι ανάγκες σε αυτό μεταβάλλονται, καθώς και ο δυναμικός τρόπος παραγωγής ιστοσελίδων αποτελούν τους τρεις βασικούς παράγοντες που καθιστούν τη διαδικασία της ανίχνευσης ιδιαίτερα δύσκολη. Το μεγάλο εύρος σημαίνει ότι ο ανιχνευτής μπορεί να μεταφορτώσει μόνο ένα μικρό ποσοστό των σελίδων του Ιστού, σε δεδομένο χρόνο, με

αποτέλεσμα την αναγκαιότητα θέσπισης προτεραιοτήτων των μεταφορτώσεων. Ο ρυθμός μεταβολής σημαίνει ότι, κατά τη διάρκεια της μεταφόρτωσης των τελευταίων και πλέον πρόσφατων σελίδων ενός ιστοτόπου, είναι πολύ πιθανό νέες σελίδες να έχουν μόλις προστεθεί στον ιστότοπο, ή ορισμένες από τις μεταφορτωμένες ιστοσελίδες να έχουν ήδη ανανεωθεί ή διαγραφεί. Ο αριθμός των μέγιστων δυνατών ανιχνεύσιμων διευθύνσεων URL που μπορούν να παραχθούν από λογισμικό της πλευράς του διακομιστή οδηγεί τις περισσότερες φορές στην ανάκτηση διπλού περιεχομένου. Για παράδειγμα, σε μία απλή φωτογραφική γκαλερί εάν υπάρχουν τέσσερις τρόποι ταξινόμησης των εικόνων, τρεις επιλογές μεγέθους και δύο υποστηριζόμενοι τύποι αρχείου τότε στο ίδιο σύνολο περιεχομένου μπορεί να δίνεται πρόσβαση με 24 διαφορετικές τοποθεσίες URL, όλες εκ των οποίων μπορούν να συνδεθούν στη σελίδα. Αυτό ο μαθηματικός συνδυασμός δημιουργεί πρόβλημα στους ανιχνευτές, καθώς πρέπει να ταξινομήσουν ατελείωτους συνδυασμούς μερικώς τροποποιημένου περιεχομένου, με σκοπό την ανάκτηση μοναδικού περιεχομένου. Ένας ανιχνευτής πρέπει πολύ προσεκτικά να επιλέγει, σε κάθε βήμα, ποιες σελίδες να επισκεφθεί στο επόμενο βήμα. Η συμπεριφορά ενός ανιχνευτή Ιστού είναι το αποτέλεσμα ενός συνδυασμού πολιτικών:

- μία πολιτική επιλογής που δηλώνει ποιες σελίδες είναι προς μεταφόρτωση,
- μία πολιτική επανεπίσκεψης που δηλώνει πότε να πραγματοποιείται έλεγχος για αλλαγές στη σελίδα,
- μία πολιτική ευγένειας που δηλώνει πώς να αποφεύγεται η υπερφόρτωση ιστοσελίδων και
- μία πολιτική παραλληλοποίησης που δηλώνει πώς να συντονίζονται οι διανεμημένοι ανιχνευτές Ιστού.

1.4.1.1 Πολιτική επιλογής

Δεδομένου του σημερινού μεγέθους του Παγκόσμιου Ιστού, ακόμη και μεγάλες μηχανές αναζήτησης καλύπτουν μόλις ένα ποσοστό του δημοσίως διαθέσιμου κομματιού του. Μία έρευνα του 2005 έδειξε ότι μεγάλης κλίμακας μηχανές αναζήτησης ευρετηριάζουν λιγότερο από το 40% έως 70% του υπό ευρετηρίαση Ιστού. Μία προηγούμενη έρευνα, που διεξήχθη από τους Steve Lawrence και Lee Giles [1], έδειξε ότι καμία μηχανή αναζήτησης δεν ευρετηρίασε περισσότερο από το 16% του Ιστού, το 1999. Καθώς ένας ανιχνευτής πάντα μεταφορτώνει μόλις ένα μέρος των ιστοσελίδων, είναι ιδιαίτερα επιθυμητό το μεταφορτωμένο μέρος να περιλαμβάνει τις πιο σχετικές σελίδες και όχι απλώς ένα τυχαίο δείγμα του Ιστού. Αυτό προϋποθέτει την ύπαρξη ενός μέτρου σπουδαιότητας για την ιεράρχηση των ιστοσελίδων. Η σπουδαιότητα μίας σελίδας αποτελεί μία συνάρτηση της εσωτερικής ποιότητας, της δημοτικότητας σε όρους συνδέσμων ή επισκέψεων, ακόμη και της διεύθυνσης URL που την χαρακτηρίζουν.

1.4.1.2 Πολιτική επανεπίσκεψης

Όταν μία σελίδα *δημιουργείται*, δεν είναι ορατή ούτε διαθέσιμη στους χρήστες του Παγκόσμιου Ιστού μέχρι κάποια προϋπάρχουσα και γνωστή σελίδα δημιουργήσει έναν σύνδεσμο προς αυτήν. Άρα τουλάχιστον μία ανανέωση σελίδας, η οποία συμπεριλαμβάνει την προσθήκη συνδέσμου προς τη νέα ιστοσελίδα, πρέπει να πραγματοποιηθεί προτού η ιστοσελίδα αυτή να είναι ορατή. Ο ανιχνευτής ξεκινάει από ένα σύνολο εναρκτήριων διευθύνσεων URL, που συνήθως αποτελείται από μία λίστα από domains, οπότε η εγγραφή ενός domain μπορεί να εκφράσει τη διαδικασία δημιουργίας μίας URL.

Αντίστοιχα, όταν μία σελίδα *ανανεώνεται*, η ενημέρωση μπορεί να είναι κύρια ή δευτερεύουσα. Η ενημέρωση είναι δευτερεύουσα όταν αφορά αλλαγές σε επίπεδο παραγραφών ή προτάσεων, οπότε η σελίδα παραμένει σημασιολογικά σχεδόν η ίδια με προηγουμένως και οι αναφορές στο περιεχόμενό της εξακολουθούν να είναι ορθές. Αντίθετα, στην περίπτωση μίας κύριας ενημέρωσης, όλες οι αναφορές στο περιεχόμενο ακυρώνονται. Είναι σύνηθες να θεωρούνται όλες οι μεταβολές ως κύριες, καθώς η βασική δυσκολία διάκρισης έγκειται στην περιορισμένη δυνατότητα της γνώσης του κατά πόσο το περιεχόμενο μίας σελίδας παραμένει σημασιολογικά το ίδιο. Τέλος, μία σελίδα *διαγράφεται*, όταν αφαιρείται η ίδια από τον Ιστό ή όταν όλοι οι σύνδεσμοι προς αυτήν αφαιρούνται από τον Ιστό.

1.4.1.3 Πολιτική ευγένειας

Οι ανιχνευτές μπορούν να ανακτούν δεδομένα πολύ ταχύτερα και σε μεγαλύτερο βάθος από ότι οι άνθρωποι. Επομένως, παρόλο που ένας απλός ανιχνευτής διεκπεραιώνει πολλαπλά αιτήματα ανά δευτερόλεπτο και μεταφορτώνει αρχεία μεγάλου μεγέθους, ένας διακομιστής θα δυσκολευόταν να αντιμετωπίσει αιτήματα από πολλαπλούς ανιχνευτές. Η χρήση των ανιχνευτών Ιστού είναι χρήσιμη για έναν συγκεκριμένο αριθμό διεργασιών, αλλά επιβαρύνει την κοινότητα του Διαδικτύου με διάφορους τρόπους. Το κόστος χρήσης ανιχνευτών Ιστού περιλαμβάνει :

- δικτυακούς πόρους, καθώς οι ανιχνευτές απαιτούν σημαντικό εύρος σύνδεσης,
- υπερφόρτωση διακομιστών, ειδικά εάν η συχνότητα προσβάσεων σε δεδομένο διακομιστή είναι υψηλή,
- ανεπαρκώς γραμμένους ανιχνευτές, που μπορούν να καταστρέψουν διακομιστές ή δρομολογητές και που μεταφορτώνουν σελίδες που δε μπορούν να χειριστούν, και
- προσωπικούς ανιχνευτές που, εάν αναπτυχθούν και χρησιμοποιηθούν από πολλούς χρήστες, μπορούν να διαταράξουν δίκτυα και εξυπηρετητές.

Μία μερική λύση σε αυτά τα προβλήματα είναι το πρωτόκολλο εξαίρεσης «robots», ευρέως γνωστό και ως πρωτόκολλο robots.txt, που αποτελεί πρότυπο για διαχειριστές ιστοσελίδων για να υποδεικνύουν ποιές τοποθεσίες ενός ιστοχώρου ή διακομιστή δεν

πρέπει να είναι προσβάσιμες από έναν ανιχνευτή. Αυτό το πρότυπο δεν περιλαμβάνει μία πρόταση για το διάστημα των επισκέψεων στον ίδιο διακομιστή, παρότι αυτό το διάστημα αποτελεί τον πιο σημαντικό παράγοντα αποφυγής υπερφορτώσεων. Η πρώτη προτεινόμενη τιμή αυτού του διαστήματος μεταξύ των συνδέσεων ήταν 60 δευτερόλεπτα. Όμως, εάν οι σελίδες μεταφορτώνονται με αυτό το ρυθμό από έναν ιστότοπο με περισσότερες από 100,000 σελίδες, με «τέλεια» σύνδεση, άπειρο εύρος διασύνδεσης και μηδενικό λανθάνοντα χρόνο (latency), θα χρειάζονταν παραπάνω από δύο μήνες για τη μεταφόρτωση μόλις ολόκληρου του ιστότοπου και μόνο. Επίσης, μόλις ένα μέρος των πόρων αυτού του διακομιστή θα χρησιμοποιούνται, το οποίο δε θα ήταν αποδεκτό. Οι Cho και Garcia – Molina (2003) [2] χρησιμοποιούν διάστημα μεταξύ προσβάσεων 10 δευτερολέπτων. Ο ανιχνευτής MercatorWeb [3] ακολουθεί μία προσαρμοστική πολιτική ευγένειας: εάν χρειάζονταν t δευτερόλεπτα για τη μεταφόρτωση ενός εγγράφου από ένα δεδομένο διακομιστή, ο ανιχνευτής περιμένει 10t δευτερόλεπτα πριν τη μεταφόρτωση της επόμενης σελίδας.

1.4.1.4 Πολιτική παραλληλοποίησης

Ο παράλληλος ανιχνευτής διεκπεραιώνει πολλές διεργασίες παράλληλα. Στόχο αποτελεί η μεγιστοποίηση του ρυθμού μεταφόρτωσης, ελαχιστοποιώντας τις μεταφορτώσεις της ίδιας σελίδας. Για να γίνει αυτό, το σύστημα ανίχνευσης απαιτεί μία πολιτική ανάθεσης των νέων διευθύνσεων URL που ανακαλύπτονται, κατά τη διάρκεια της ανίχνευσης, καθώς οι ίδιες URL μπορούν να ευρεθούν από δύο διαφορετικές διεργασίες ανίχνευσης. Οι Cho και Garcia – Molina (2002) [4] μελέτησαν δύο τύπους πολιτικών παραλληλοποίησης:

- Την πολιτική δυναμικής ανάθεσης, με την οποία ένας κεντρικός διακομιστής αναθέτει νέες URL διευθύνσεις σε διαφορετικούς ανιχνευτές δυναμικά, επιτρέποντας, έτσι, τη δυναμική εξισορρόπηση της φόρτωσης κάθε ανιχνευτή.
- Την πολιτική στατικής ανάθεσης, σύμφωνα με την οποία υπάρχει ένας αρχικός κανόνας που ρυθμίζει με ποιον τρόπο γίνονται αναθέσεις νέων διευθύνσεων URL στους ανιχνευτές.

1.4.2 Βασικοί αλγόριθμοι ανίχνευσης

Οι βασικοί αλγόριθμοι ανίχνευσης που παρουσιάζονται αποτελούν παραλλαγές του best – first σχεδίου (αναζήτηση πρώτα στο καλύτερο). Η βασική διαφορά βρίσκεται στη μέθοδο εύρεσης που χρησιμοποιούν για την ανίχνευση σελίδων που δεν έχουν ήδη επισκεφθεί, προσαρμόζοντας και ρυθμίζοντας τις παραμέτρους του αλγορίθμου πριν ή κατά τη διάρκεια της ανίχνευσης.

1.4.2.1 «Αφελής» πρώτα στο καλύτερο ανίχνευση

Η ανίχνευση αυτή αντιλαμβάνεται μία ανακτημένη ιστοσελίδα ως ένα διάνυσμα της συχνότητας εμφάνισης, για κάθε λέξη. Ο ανιχνευτής, έπειτα, υπολογίζει την ομοιότητα της σελίδας με το ερώτημα αναζήτησης (query) ή την περιγραφή που δίνεται από το χρήστη και επισκέπτεται τις διευθύνσεις URL βάσει της ομοιότητας αυτής. Οι URLs, στη συνέχεια, προστίθενται σε ένα crawl frontier με σειρά προτεραιότητας, βάσει αυτών των ομοιοτήτων. Στην επόμενη επανάληψη, κάθε νήμα του ανιχνευτή επιλέγει την καλύτερη διεύθυνση URL για ανίχνευση και επιστρέφει με νέες μη επισκεφθείσες διευθύνσεις URL που, ομοίως, καταχωρούνται στην ουρά προτεραιότητας με την τιμή ομοιότητας της αρχικής σελίδας. Η ομοιότητα μεταξύ της σελίδας p και του ερωτήματος q υπολογίζεται με τη σχέση

$$sim(q, p) = \frac{v_q * v_p}{\|v_q\| * \|v_p\|}$$

όπου v_q και v_p είναι διανυσματικές αναπαραστάσεις, που βασίζονται στη συχνότητα όρων, του ερωτήματος και της σελίδας αντίστοιχα.

1.4.2.2 Αλγόριθμος SharkSearch

Ο αλγόριθμος SharkSearch χρησιμοποιεί ένα μέτρο ομοιότητας, παρόμοιο με αυτό της «αφελούς» πρώτα στο καλύτερο ανίχνευσης για τον υπολογισμό της σχετικότητας μίας μη επισκεφθείσας τοποθεσίας URL. Όμως, ο αλγόριθμος SharkSearch περιλαμβάνει μία πιο εξευγενισμένη έννοια του δυνητικού αποτελέσματος για τους συνδέσμους του crawl frontier. Σύμφωνα με τον SharkSearch , εάν ο ανιχνευτής φθάνει σε ασήμαντες σελίδες σε ένα υπό ανίχνευση μονοπάτι, διακόπτει την περαιτέρω ανίχνευση του συγκεκριμένου μονοπατιού. Για να δύναται να καταγράφει όλες τις πληροφορίες, κάθε διεύθυνση URL του frontier συνοδεύεται από ένα βάθος και ένα δυνητικό αποτέλεσμα. Το όριο βάθους (d) παρέχεται από το χρήστη, ενώ το δυνητικό αποτέλεσμα μίας μη επισκεφθείσας τοποθεσίας URL υπολογίζεται ως εξής:

$$score(url) = g * inherited(url) + (1-g) * neighborhood(url)$$

όπου γ<1 είναι μία παράμετρος, το neighborhood score είναι το αποτέλεσμα γειτνίασης και δηλώνει τις συναφείς αποδείξεις εντός της σελίδας που περιλαμβάνει τον υπερσύνδεσμο URL και το inherited score είναι το διαπερνόν αποτέλεσμα που υπολογίζεται από τα αποτελέσματα των προηγούμενων σελίδων. Συγκεκριμένα, το inherited score υπολογίζεται από την σχέση

$$inherited(url) = \begin{cases} d * sim(q,p) \text{ εάν } sim(q,p) > 0 \\ d * inherited(p) \text{ αλλιώς} \end{cases}$$

όπου δ<1 είναι μία παράμετρος, q το ερώτημα αναζήτησης και p η σελίδα από την οποία εξήχθη η διεύθυνση URL.

1.4.2.3 Αλγόριθμος επικεντρωμένου ανιχνευτή

Η βασική ιδέα ενός τέτοιου ανιχνευτή ήταν να ταξινομεί σελίδες που έχουν ανιχνευθεί σε ένα θεματικό κατάλογο. Αρχικά, ο ανιχνευτής χρειάζεται μία τέτοια κατηγοριοποίηση, όπως ο κατάλογος Yahoo ή το ODP (Open Directory Project της dmoz.org). Έπειτα, ο χρήστης παρέχει παραδείγματα URL διευθύνσεων σχετικού ενδιαφέροντος (όπως, περίπου, σε ένα ηλεκτρονικό αρχείο σελιδοδεικτών). Μέσα από μία διαδραστική διαδικασία, ο χρήστης μπορεί να διορθώσει την αυτόματη ταξινόμηση, να προσθέσει νέες κατηγορίες και να αξιολογήσει θετικά ορισμένες κατηγορίες ως «good». Ο ανιχνευτής χρησιμοποιεί αυτά τα παραδείγματα για να κατασκευάσει έναν ταξινομητή Bayes που μπορεί να υπολογίσει την πιθανότητα P(c | p) να ανήκει μία σελίδα p που έχει ανιχνευθεί σε μία κατηγορία c στον κατάλογο. Εξ ορισμού είναι , P(r | p)=1, όπου r είναι η αρχική κατηγορία της ταξινόμησης (root category). Ένας βαθμός συσχέτισης συνοδεύει κάθε σελίδα που υπολογίζεται από τη σχέση:

$$R(p) = \sum_{c \in good} P(c \mid \mathbf{p})$$

Όταν ο ανιχνευτής βρίσκεται σε μία «ελαφρώς» επικεντρωμένη λειτουργία, χρησιμοποιεί το βαθμό συσχέτισης της σελίδας για να βαθμολογήσει την μη επισκεφθείσα URL που προέρχεται από αυτήν. Οι βαθμολογημένοι αυτοί προορισμοί προστίθενται, στη

συνέχεια, στο crawl frontier. Έπειτα, με έναν τρόπο παρόμοιο με την «αφελή» πρώτα στο καλύτερο ανίχνευση, επιλέγει τους καλύτερους προορισμούς για να ανιχνεύσει επόμενους.

1.4.2.4 Αλγόριθμος InfoSpiders

Στον αλγόριθμο InfoSpiders των Menczer, Pant και Srinivasan (2004) [5], ένα προσαρμοστικό πλήθος ανιχνευτών αναζητά για σελίδες σχετικές με το θέμα αναζήτησης. Κάθε πράκτορας ουσιαστικά ακολουθεί τη διαδικασία ανίχνευσης, χρησιμοποιώντας μία προσαρμοστική λίστα ερωτημάτων για να αποφασίσει ποιους συνδέσμους θα ακολουθήσει. Ο αλγόριθμος παρέχει ένα αποκλειστικό crawl frontier για κάθε πράκτορα. Σε μία πολυεπίπεδη εφαρμογή του InfoSpiders, κάθε πράκτορας ανταποκρίνεται σε ένα νήμα εκτέλεσης. Ως εκ τούτου, κάθε νήμα έχει μία αδιαμφισβήτητη πρόσβαση στο δικό του frontier, δηλαδή κάθε νήμα κατέχει το δικό του crawl frontier.

Η προσαρμοστική αναπαράσταση κάθε πράκτορα αποτελείται από μία λίστα λέξεων – κλειδιών (με αφετηρία ένα ερώτημα ή μία περιγραφή) και ένα ουδέτερο δίχτυ για την αξιολόγηση νέων συνδέσμων. Κάθε είσοδος στο ουδέτερο δίχτυ λαμβάνει τη μέτρηση της συχνότητας με την οποία η λέξη – κλειδί εμφανίζεται γειτονικά κάθε συνδέσμου προς εξέταση, δίνοντας μεγαλύτερη βαρύτητα σε λέξεις – κλειδιά που βρίσκονται πολύ κοντά στο σύνδεσμο (και, φυσικά, τη μέγιστη βαρύτητα στο anchor text). Η μοναδική έξοδος του ουδέτερου διχτυού χρησιμοποιείται ως μία αριθμητική εκτίμηση ποιότητας για κάθε σύνδεσμο. Αυτές οι εκτιμήσεις, στη συνέχεια, συνδυάζονται με εκτιμήσεις που βασίζονται στην υπολογισθείσα ομοιότητα, που έχει ήδη εξεταστεί, μεταξύ της λέξης – κλειδιού του πράκτορα και της σελίδας που εμπεριέχει τους συνδέσμους. Με βάση τον τελικό βαθμό, ο πράκτορας χρησιμοποιεί έναν στοχαστικό επιλογέα για την επιλογή ενός εκ των συνδέσμων του frontier με πιθανότητα

$$P(l) = \frac{e^{bs(l)}}{\sum_{l' \in j} e^{bs(l')}}$$

όπου λ είναι μία URL του τοπικού frontier (φ) και σ(λ) είναι η τελική βαθμολογία συνδυασμού των εκτιμήσεων. Η παράμετρος β ρυθμίζει την υποκειμενικότητα του επιλογέα συνδέσμων. Αφού μία νέα σελίδα έχει επιλεχθεί, ο πράκτορας λαμβάνει «ενέργεια», εν αντιστοιχία με την ομοιότητα μεταξύ της λέξης – κλειδιού και της νέας σελίδας. Το ουδέτερο δίχτυ του πράκτορα μπορεί να εκπαιδευθεί με σκοπό τη βελτίωση

των εκτιμήσεων των συνδέσμων, προβλέποντας την ομοιότητα της νέας σελίδας, δεδομένων των εισόδων από τη σελίδα που περιελάμβανε τον σύνδεσμο που οδήγησε σε αυτή. Ένας αλγόριθμος οπισθοδιάδοσης χρησιμοποιείται για εκμάθηση. Μία τέτοια τεχνική εκμάθησης παρέχει στον αλγόριθμο InfoSpiders τη μοναδική ικανότητα να προσαρμόζει τη συμπεριφορά επίσκεψης συνδέσμων στην πορεία μίας ανίχνευσης συνδέοντας εκτιμήσεις συσχέτισης με την προτυποποίηση συχνοτήτων εμφάνισης της λέξης – κλειδιού κοντά στους συνδέσμους.

1.5 Ευρετηρίαση εγγράφων (indexing)

Η διαδικασία της ευρετηρίασης των μηχανών αναζήτησης συλλέγει, επεξεργάζεται και αποθηκεύει δεδομένα για να διευκολύνει την άμεση και ακριβή ανάκτηση πληροφοριών. Ο σχεδιασμός του ευρετηρίου (index) ενσωματώνει έννοιες από τη γλωσσολογία, την ψυχολογία, τα μαθηματικά, την πληροφορική, την φυσική και την επιστήμη των υπολογιστών. Ένα εναλλακτικό όνομα για τη διαδικασία, στα πλαίσια των μηχανών αναζήτησης, που σχεδιάστηκε για την εύρεση ιστοσελίδων στο Διαδίκτυο είναι η Ευρετηρίαση του Παγκόσμιου Ιστού (web indexing).

Δημοφιλείς μηχανές, όπως η Google και η Yahoo!, επικεντρώνονται στην πλήρους κειμένου ευρετηρίαση συνδεδεμένων (online) εγγράφων, γραμμένων σε φυσική γλώσσα. Τύποι πολυμέσων, όπως οπτικοακουστικά αρχεία, είναι επίσης ερευνήσιμα.

Οι Meta search engines επαναχρησιμοποιούν τις βάσεις άλλων μηχανών αναζήτησης ή υπηρεσιών και δεν διατηρούν τοπικό ευρετήριο, ενώ μηχανές αναζήτησης που βασίζονται στην κρυφή μνήμη αποθηκεύουν μόνιμα το ευρετήριο, μαζί με το περιεχόμενο των σελίδων.

1.5.1 Κατασκευή ευρετηρίου

Ο σκοπός της αποθήκευσης ενός ευρετηρίου είναι να βελτιστοποιείται η ταχύτητα και απόδοση στη διαδικασία εύρεσης σχετικών εγγράφων για ένα ερώτημα αναζήτησης. Χωρίς το ευρετήριο, η μηχανή αναζήτησης θα έπρεπε να σαρώνει κάθε έγγραφο του Ιστού, γεγονός που θα απαιτούσε μεγάλο χρονικό διάστημα και υπολογιστική δύναμη. Για παράδειγμα, ενώ ένα ευρετήριο x εγγράφων μπορεί να ερωτηθεί για σχετικά έγγραφα σε πολύ σύντομο χρονικό διάστημα, μία σύγχρονη διαδοχική σάρωση κάθε λέξης, μία προς μία, σε ένα σύνολο από $y < x$ έγγραφα, χωρίς ευρετήριο, μπορεί να πάρει ώρες. Η επιπρόσθετη χρήση υπολογιστικών πόρων που απαιτούνται για να αποθηκευτεί το ευρετήριο και ο απαιτούμενος χρόνος για κάθε ενημέρωση υπερτερούν σημαντικά του χρόνου που απαιτείται από την σύγχρονη ανάκτηση πληροφοριών.

1.5.1.1 Παράγοντες σχεδίασης του ευρετηρίου

Σημαντικοί παράγοντες που καθορίζουν το σχεδιασμό της αρχιτεκτονικής του ευρετηρίου μίας μηχανής αναζήτησης είναι οι εξής:

- **Συγχώνευση**

  Τέτοιοι παράγοντες αφορούν τον τρόπο με τον οποίο τα δεδομένα εισέρχονται στο υπάρχον ευρετήριο και οι λέξεις προστίθενται σε ευρετηριασμένα έγγραφα ή τη δυνατότητα ασύγχρονης λειτουργίας πολλαπλών διαδικασιών ευρετηρίασης (indexers).

- **Τεχνικές αποθήκευσης**

  Αφορούν τους τρόπους με τους οποίους αποθηκεύονται τα δεδομένα του ευρετηρίου, εάν, δηλαδή, τα δεδομένα πρέπει να συμπιέζονται ή να φιλτράρονται κατά την ευρετηρίαση.

- **Μέγεθος ευρετηρίου**

  Αφορά το μέγεθος, το πλήθος και τις δυνατότητες των υπολογιστικών πόρων αποθήκευσης που απαιτούνται για την υποστήριξη ενός ευρετηρίου.

- **Ταχύτητα αναζήτησης**

  Ο παράγοντας αυτός εξετάζει το πόσο γρήγορα μία λέξη μπορεί να ευρεθεί στο ανεστραμμένο ευρετήριο (πλήρως ταξινομημένο ευρετήριο).

- **Συντήρηση**

  Αφορά τους τρόπους, τη διάρκεια, τις τεχνικές και τον προγραμματισμό της συντήρησης του ευρετηρίου στο χρόνο.

- **Ανεκτικότητα σφαλμάτων**

  Ο παράγοντας αυτός αφορά τη σημασία που έχει η αξιοπιστία της υπηρεσίας, ενώ συμπεριλαμβάνει την αντιμετώπιση της φθοράς, την απομόνωση λανθασμένων δεδομένων, την αντιμετώπιση κακού υπολογιστικού υλικού (hardware) και την αντιγραφή.

1.5.1.2 Ευρετήριο παραπομπών

Η δομή αυτή αποθηκεύει παραπομπές ή υπερσυνδέσμους μεταξύ εγγράφων για την υποστήριξη αναλύσεων παραπομπών, ένα αντικείμενο της Βιβλιομετρίας.

1.5.1.3 Ευρετήριο Ngram

Χρησιμοποιείται ως δομή για την αποθήκευση ακολουθιών από μήκη δεδομένων για την υποστήριξη διαφορετικών τύπων ανάκτησης.

1.5.1.4 Πίνακας όρων – εγγράφων (ή εγγράφων – όρων)

Πρόκειται για ένα μαθηματικό δισδιάστατο πίνακα που περιγράφει τη συχνότητα των όρων που εμφανίζονται σε μία συλλογή εγγράφων. Στον εν λόγω πίνακα, οι γραμμές αντιστοιχούν στα έγγραφα και οι στήλες αντιστοιχίζονται στους όρους. Για παράδειγμα, έστω ότι διαθέτουμε τα παρακάτω δύο έγγραφα:

T_0=″μου αρέσουν πολύ τα πράσινα μήλα″

T_1=″υπάρχουν τα κόκκινα μήλα και τα πράσινα μήλα″

Ο πίνακας που προκύπτει είναι ο εξής:

Πίνακας 1 Παράδειγμα ευρετηρίου όρων-εγγράφων

	Μου	αρέσουν	πολύ	Τα	μήλα	υπάρχουν	κόκκινα	και	πράσινα
T_0	1	1	1	1	1	0	0	0	1
T_1	0	0	0	2	2	1	1	1	1

1.5.1.5 Ευθύ ευρετήριο

Το ευθύ ευρετήριο αποθηκεύει μία λίστα από λέξεις για κάθε έγγραφο. Μία απλουστευμένη μορφή ενός τέτοιου ευρετηρίου, βασισμένη στο παράδειγμα του ανεστραμμένου ευρετηρίου, είναι η εξής:

T_0 :{″είναι″,″αυτό″,″που″}

T_1 :{″είναι″,″αυτό″,″τι″}

T_2 :{″ένα″, ″ευρετήριο″, ″είναι″,″αυτό″}

Η λογική πίσω από την ανάπτυξη ενός τέτοιου ευρετηρίου είναι ότι καθώς τα έγγραφα αναλύονται, είναι καλύτερο να αποθηκεύονται απευθείας οι λέξεις ανά έγγραφο. Η σκιαγράφηση διευκολύνει την ασύγχρονη επεξεργασία συστημάτων, η οποία παρακάμπτει μερικώς τη συμφόρηση ενημερώσεων του ανεστραμμένου ευρετηρίου, ενώ, όντας ταξινομημένο, μπορεί να μετατραπεί σε ανεστραμμένο ευρετήριο.

1.5.1.6 Συμπίεση

Η παραγωγή ή συντήρηση του ευρετηρίου μηχανής αναζήτησης μεγάλης κλίμακας αναπαριστά μία σημαντική πρόκληση, σε όρους αποθήκευσης και επεξεργασίας. Πολλές μηχανές αναζήτησης χρησιμοποιούν μία μορφή συμπίεσης, λοιπόν, για να ελαττώσουν το μέγεθος των δεικτών στο δίσκο. Για παράδειγμα, έστω το ακόλουθο σενάριο για μία πλήρους κειμένου Διαδικτυακή μηχανή αναζήτησης:

- Το έτος 2000, υπήρχαν περίπου 2 δισεκατομμύρια διαφορετικές ιστοσελίδες.
- Έστω ότι υπάρχουν 250 λέξεις σε κάθε ιστοσελίδα.
- Απαιτούνται 8 bits (ή 1 byte) για να αποθηκευτεί ένας μόνο χαρακτήρας. Ορισμένες κωδικοποιήσεις, μάλιστα, απαιτούν 2 bytes ανά χαρακτήρα.
- Ο μέσος όρος των χαρακτήρων σε οποιαδήποτε λέξη μίας σελίδας μπορεί να υποτεθεί ότι είναι πέντε.
- Ο μέσος ηλεκτρονικός υπολογιστής διαθέτει περίπου από 100 έως 250 gigabytes ελεύθερου δίσκου.

Δεδομένων αυτών, ένα μη συμπιεσμένο, υποθετικά απλό ευρετήριο δυο δισεκατομμυρίων σελίδων θα έπρεπε να αποθηκεύσει 500 δισεκατομμύρια καταχωρήσεις. Με 1 byte ανά χαρακτήρα, ή 5 bytes ανά λέξη, θα απαιτούνταν περίπου 2500 gigabytes σκληρού δίσκου, περισσότερο, δηλαδή, από το μέσο σκληρό δίσκο 25 ηλεκτρονικών υπολογιστών. Αυτές οι χωρικές απαιτήσεις μπορεί να είναι ακόμη μεγαλύτερες για μία διανεμημένης αποθήκευσης αρχιτεκτονική με σχετική ανεκτικότητα σφάλματος. Ανάλογα με την τεχνική συμπίεσης που επιλέγεται, το μέγεθος του ευρετηρίου μπορεί να ελαττωθεί αρκετά. Το εναλλακτικό κόστος, σε χρόνο και επεξεργαστική ισχύ που απαιτούνται για τη διαδικασία της συμπίεσης και αποσυμπίεσης, υπερτερεί έναντι του μεγέθους ενός μη συμπιεσμένου ευρετηρίου. Σημειώνεται εδώ ότι οι σχεδιασμοί μεγάλης κλίμακας μηχανών αναζήτησης ενσωματώνουν το κόστος αποθήκευσης και ηλεκτρισμού που απαιτείται για αυτήν. Επομένως, η συμπίεση μετράται και σε κόστος.

1.5.2 Ανάλυση εγγράφων

Η διαδικασία της ανάλυσης εγγράφων διαχωρίζει τα περιεχόμενα ενός εγγράφου (λέξεις ή διάφορα στοιχεία πολυμέσων) για την καταχώρησή τους στους ευθείς και ανεστραμμένους δείκτες (ευρετήρια). Οι λέξεις που διαχωρίζονται ονομάζονται τεκμήρια

ή ενδείξεις (tokens) και, επομένως, στα πλαίσια της ευρετηρίασης των μηχανών αναζήτησης και της επεξεργασίας της φυσικής γλώσσας, η ανάλυση αναφέρεται ευρέως ως tokenization, ενώ αναφέρεται συχνά και ως ετικετοποίηση (tagging), ανάλυση περιεχομένου, ανάλυση κειμένου ή λεξική ανάλυση. Η επεξεργασία της φυσικής γλώσσας αποτελεί αντικείμενο συνεχούς έρευνας και τεχνολογικής ανάπτυξης. Ο διαχωρισμός των τεκμηρίων παρουσιάζει αρκετές προκλήσεις στη διαδικασία της εξαγωγής της απαραίτητης ή χρήσιμης πληροφορίας από τα έγγραφα για την ευρετηρίαση, ενώ περιλαμβάνει πολλαπλές τεχνολογίες, η εφαρμογή των οποίων συχνά αποτελεί επιχειρησιακό μυστικό.

1.5.2.1 Προκλήσεις στην επεξεργασία της φυσικής γλώσσας

- **Ασάφεια στα όρια μεταξύ των λέξεων**

  Μπορεί η διαδικασία, λαμβάνοντας υπόψη μόνο τους Άγγλους ή αγγλόφωνους χρήστες του Διαδικτύου, να φαίνεται απλοϊκή, αλλά δεν ισχύει αυτό, αν αναλογιστούμε τις δυσκολίες σχεδίασης ενός πολυγλωσσικού συστήματος ευρετηρίασης. Στην ψηφιακή μορφή, τα κείμενα άλλων γλωσσών, όπως τα κινέζικα, τα ιαπωνικά ή τα αραβικά φαντάζουν πολύ μεγαλύτερη πρόκληση, καθώς οι λέξεις σε αυτές τις γλώσσες δεν είναι σαφώς διαχωρισμένες με κενό χαρακτήρα. Ο στόχος, κατά τη διάρκεια του διαχωρισμού, είναι να ταυτοποιηθούν λέξεις για τις οποίες οι χρήστες θα καταχωρήσουν ερωτήματα, ως όρους αναζήτησης. Για το λόγο αυτό, επιστρατεύεται, συνήθως, η λογική της εξατομίκευσης με κριτήριο τη γλώσσα, ώστε να αναγνωρίζεται κανονικά το όριο των λέξεων, με αποτέλεσμα να σχεδιάζονται αναλυτές για κάθε γλώσσα ξεχωριστά (ή για ομάδες γλωσσών με παρόμοιες ενδείξεις οριοθέτησης λέξεων και παρόμοιο συντακτικό).

- **Γλωσσική ασάφεια**

  Με στόχο να υποστηριχθεί η λογική ταξινόμηση των αντιστοιχιζόμενων στο ερώτημα αναζήτησης αποτελεσμάτων, πολλές μηχανές αναζήτησης συλλέγουν επιπρόσθετες πληροφορίες για κάθε λέξη, όπως η γλωσσική ή λεκτική κατηγορία της (μέρος του λόγου). Τα έγγραφα δεν αναπαριστούν, πάντα, επακριβώς τη γλώσσα στην οποία είναι γραμμένα, γι’αυτό, κατά το διαχωρισμό (tokenization), ορισμένες μηχανές προσπαθούν να αναγνωρίσουν αυτόματα τη γλώσσα του εγγράφου.

- **Ποικίλοι τύποι αρχείων**

  Για τον ορθό διαχωρισμό των bytes ενός εγγράφου που αναπαριστούν χαρακτήρες, ο τύπος αρχείου πρέπει να υποστεί σωστό χειρισμό. Οι μηχανές αναζήτησης που υποστηρίζουν πολλαπλούς τύπους αρχείων οφείλουν να

μπορούν να προσπελαύνουν, να εκτελούν και να έχουν πρόσβαση στο έγγραφο και να έχουν τη δυνατότητα να διαχωρίσουν τους χαρακτήρες του εγγράφου.

§ **Ελαττωματική αποθήκευση**

Η ποιότητα των δεδομένων της φυσικής γλώσσας ενδέχεται να μην είναι πάντα άριστη. Ένας απροσδιόριστος αριθμός εγγράφων, ειδικά στο Διαδίκτυο, δεν υπακούουν σε μεγάλο βαθμό το πρωτόκολλο. Οι δυαδικοί χαρακτήρες ενδέχεται να κωδικοποιηθούν, κατά λάθος, σε διάφορα σημεία ενός εγγράφου. Χωρίς την αναγνώριση αυτών των χαρακτήρων και τον κατάλληλο χειρισμό, η ποιότητα και επίδοση του ευρετηρίου μπορεί να υποβαθμιστούν.

1.5.2.2 Διαχωρισμός λέξεων ή ενδείξεων

Σε αντίθεση με τον άνθρωπο, οι υπολογιστές δεν καταλαβαίνουν τη δομή ενός κειμένου γραμμένου σε φυσική γλώσσα και δεν μπορούν αυτόματα να αναγνωρίσουν λέξεις ή προτάσεις. Συγκεκριμένα, κάθε έγγραφο σημαίνει μόνο μία ακολουθία από bytes. Οι υπολογιστές δεν γνωρίζουν ότι ο χαρακτήρας του κενού διαχωρίζει λέξεις σε ένα έγγραφο. Αντίθετα, οι άνθρωποι πρέπει να προγραμματίσουν τον υπολογιστή για να αναγνωρίζει τι αποτελεί μία ξεχωριστή λέξη, μία μονάδα, που λέγεται ένδειξη. Κατά τη διάρκεια αυτού του προγραμματισμού, της διαδικασίας, δηλαδή, κατά την οποία διαχωρίζονται οι διάφορες μονάδες ενός εγγράφου, ο αναλυτής αναγνωρίζει ακολουθίες χαρακτήρων (συμβολοακολουθίες) που αναπαριστούν λέξεις και άλλα στοιχεία, όπως ο τονισμός, κάθε μία εκ των οποίων αναπαριστάται από αριθμητικούς κωδικούς. Ο αναλυτής μπορεί, επίσης, να αναγνωρίζει οντότητες, όπως διευθύνσεις ηλεκτρονικού ταχυδρομείου, αριθμούς τηλεφώνου και τοποθεσίες URL. Κατά την αναγνώριση κάθε ένδειξης, αρκετά χαρακτηριστικά μπορεί να αποθηκευτούν, όπως η πληροφορία που καθορίζει εάν ο κάθε χαρακτήρας ή λέξη είναι σε πεζά ή κεφαλαία, η γλώσσα ή η κωδικοποίηση, το μέρος του λόγου (ουσιαστικό, ρήμα, επίθετο κλπ.), η θέση, ο αριθμός της πρότασης, η θέση της πρότασης, το μήκος και ο αριθμός της γραμμής.

1.5.2.3 Αναγνώριση της γλώσσας

Εάν η μηχανή αναζήτησης υποστηρίζει πολλαπλές γλώσσες, ένα κοινό αρχικό βήμα, κατά τη διάρκεια του διαχωρισμού, είναι η αναγνώριση της γλώσσας κάθε εγγράφου. Πολλά από τα επόμενα βήματα εξαρτώνται από τη γλώσσα (όπως ο τονισμός και η ανάλυση και κατηγοριοποίηση του μέρους του λόγου). Αναγνώριση της γλώσσας ονομάζεται η διαδικασία εκείνη στην οποία ένα πρόγραμμα Η/Υ προσπαθεί να αναγνωρίσει ή να κατηγοριοποιήσει αυτόματα τη γλώσσα ενός εγγράφου. Η αυτοματοποιημένη αυτή διαδικασία αποτελεί αντικείμενο συνεχούς έρευνας, όσον αφορά την επεξεργασία της φυσικής γλώσσας.

1.5.2.4 Ανάλυση τύπου αρχείων

Η πρόκληση που έγκειται, σχετικά με την ανάλυση τύπου ενός αρχείου, γίνεται ολοένα και πιο απαιτητικό, όσο πιο περίπλοκος είναι ο εκάστοτε τύπος. Ορισμένοι τύποι αρχείων αναλύονται και ευρετηριάζονται με πολύ λίγες επιπρόσθετες πληροφορίες, ενώ άλλοι, πιο περίπλοκοι τύποι αρχείων όχι. Μερικά παραδείγματα των τελευταίων είναι τα εξής:

- HTML και XHTML έγγραφα
- Αρχεία κειμένου ASCII (έγγραφο κειμένου χωρίς ιδιαίτερη ευανάγνωστη από υπολογιστή μορφοποίηση)
- PDF αρχεία (Portable Document Format)
- PS (PostScript) έγγραφα
- LaTeX έγγραφα
- Usenet αρχεία
- XML έγγραφα (Extensible Markup Language) και παράγωγα (RSS)
- SGML αρχεία
- Αρχεία Microsoft Office (Excel, Word, PowerPoint)

Μία κοινή πρακτική που επιλύει το πρόβλημα ανάλυσης τέτοιων εγγράφων αποτελεί η έκδοση ενός δημοσίως διαθέσιμου προς χρήση εμπορικού εργαλείου ανάλυσης που προσφέρεται από την εταιρεία ή τον οργανισμό που ανέπτυξε, διατηρεί ή κατέχει τον τύπο αρχείου (όπως το PDF της Adobe).

1.5.3 Επεξεργασία ερωτημάτων

Υποθέτουμε ότι διαθέτουμε μία συλλογή από έγγραφα που περιέχονται σε n ιστοσελίδες, οι οποίες έχουν ήδη ανιχνευθεί και είναι διαθέσιμες στο ευρετήριο. Έστω ότι υπάρχουν m διαφορετικές λέξεις που εμφανίζονται οπουδήποτε στη συλλογή. Κάθε συμβολοακολουθία που περικλείεται από τα αντίστοιχα διαχωριστικά σύμβολα (κενό, κόμμα, τελεία κ.α.) αποτελεί μία έγκυρη λέξη (ή όρο) κατά την ευρετηρίαση σε μία μηχανή αναζήτησης. Το ευρετήριο για τη συλλογή των εγγράφων αποτελείται από ένα σύνολο ανεστραμμένων λιστών , όπου η λίστα εμπεριέχει μία εγγραφή για κάθε εμφάνιση της λέξης w. Κάθε τέτοια εγγραφή περιλαμβάνει την ταυτότητα του αρχείου στο οποίο η λέξη εμφανίζεται, τη θέση στην οποία βρίσκεται αυτή, καθώς και πληροφορίες για το πλαίσιο της λέξης (εάν βρίσκεται στο όνομα του εγγράφου, σε τίτλο, σε μεγάλη ή έντονη γραμματοσειρά, σε περιγραφή εικόνας ή σε anchor text). Συνήθως, οι εγγραφές αυτές ταξινομούνται με κριτήριο την ταυτότητα του εγγράφου, κι, ενδεχομένως, σε αύξουσα ή φθίνουσα σειρά των θέσεων των λέξεων εντός του εγγράφου, των χαρακτήρων των λέξεων, ή, στην καλύτερη περίπτωση, σε συνδυασμό των δύο, με αποτέλεσμα να διευκολύνεται η συμπίεση της λίστας καθώς και η εύκολη εύρεση της συχνότητας των λέξεων και των όρων που πλαισιώνουν μία συγκεκριμένη λέξη.

Αντίστοιχα ορίζεται το δεύτερο μέρος της σύγκρισης που λαμβάνει χώρα κατά τη διαδικασία της επεξεργασίας των ερωτημάτων. Ένα ερώτημα είναι ένα σύνολο όρων (λέξεων). Η σχέση μεταξύ των λέξεων που συναποτελούν το ερώτημα, η οποία καθορίζει την επεξεργασία και την εύρεση και παρουσίαση των αποτελεσμάτων αναζήτησης, καθορίζεται από τους τελεστές που ο χρήστης χρησιμοποιεί για να εκφράσει το ερώτημα. Ο πιο συνηθισμένος τρόπος για την κατάταξη σε Information Retrieval Systems βασίζεται στη σύγκριση των λέξεων (όρων) που περιλαμβάνονται στο έγγραφο και το ερώτημα. Πιο συγκεκριμένα, ένας αλγόριθμος κατάταξης αναθέτει ένα score (αποτέλεσμα) σε κάθε έγγραφο του ευρετηρίου, που βασίζεται στη συχνότητα με την οποία εμφανίζεται η συνολική φράση του ερωτήματος ή μέρος αυτής μέσα στη σελίδα, το μέγεθος του αρχείου, το πλαίσιο της εκάστοτε εμφάνισης (π.χ. μεγαλύτερο score λαμβάνει ένα έγγραφο εάν ο όρος της αναζήτησης περιλαμβάνεται εντός του τίτλου της σελίδας ή σε έντονη γραμματοσειρά κι αντίστοιχα μεγαλύτερο score λαμβάνει ένα άλλο έγγραφο που, συγκριτικά με το πρώτο, περιλαμβάνει τη λέξη στον τίτλο, ο οποίος, όμως, αποτελείται από λιγότερες λέξεις ή χαρακτήρες). Δηλαδή, συνάρτηση κατάταξης αποτελεί μία συνάρτηση F που, δεδομένου ενός ερωτήματος , αναθέτει σε κάθε έγγραφο D, ένα score . Το σύστημα, στη συνέχεια, επιστρέφει τα k έγγραφα με το μεγαλύτερο score, θέτοντας ως βάση πρόκρισης ένα αποτέλεσμα που θα εξασφαλίζει μία τυπική σχετικότητα του ερωτήματος με το έγγραφο. Επειδή, όμως, οι σύγχρονες μηχανές αναζήτησης αναθέτουν κάποιο θετικό score σε παράγοντες τόσο σχετικούς όσο και άσχετους με το ερώτημα, η συνάρτηση κατάταξης δε μπορεί να είναι αθροιστική, ή χρησιμοποιούνται περισσότερες από μία συναρτήσεις, ακριβώς για να εξασφαλίζεται αυτή η ζητούμενη συσχέτιση όλων των παρεχόμενων αποτελεσμάτων με το ερώτημα του χρήστη.

1.5.4 Τελεστές αναζήτησης

Απουσία τελεστών, οποιαδήποτε αναζήτηση σε μία σύγχρονη μηχανή θα επιστρέψει ιστοσελίδες που περιλαμβάνουν όλες τις λέξεις του ερωτήματος με την ακριβή σειρά τους, έπειτα τις σελίδες εκείνες που περιλαμβάνουν όλες τις λέξεις σε οποιαδήποτε σειρά και στη συνέχεια όλες τις σελίδες που είναι σχετικές με μεγάλο έως μικρό μέρος του ερωτήματος. Η κατάταξη των αποτελεσμάτων, φυσικά, εξαρτάται από πολλούς παράγοντες που θα αναπτυχθούν σε επόμενο κεφάλαιο της παρούσας Διπλωματικής εργασίας. Οι βασικότεροι τελεστές αναζήτησης που καθορίζουν τον τρόπο με τον οποίο μία μηχανή θα επεξεργαστεί ένα ερώτημα, συγκρίνοντάς το με το ευρετήριό της με διαφορετικό τρόπο κάθε φορά, είναι οι εξής:

Πίνακας 2 Οι βασικοί τελεστές αναζήτησης

Τελεστής	**Περιγραφή**	**Παράδειγμα**
+ ή AND	Οι σελίδες που επιστρέφει η μηχανή εμπεριέχουν όλους τους όρους που συνδέονται από τον τελεστή.	Information +retrieval
| ή OR	Οι σελίδες που επιστρέφει η μηχανή εμπεριέχουν όλους ή οποιονδήποτε μόνο	σταλακτίτες|σταλαγμίτες

	από τους όρους του ερωτήματος.	
""	Η φράση που περικλείεται από τα εισαγωγικά θα αναζητηθεί επακριβώς στο ευρετήριο.	"τμήμα ηλεκτρολογίας"
-	Η μηχανή αναζήτησης δεν επιστρέφει σελίδες που περιλαμβάνουν τη λέξη ή φράση δεξιά του τελεστή.	υπολογιστές -laptop
*	Ο αστερίσκος αναπαριστά μία οποιαδήποτε λέξη και η μηχανή θα τον αντικαταστήσει με όλες τις πιθανές λέξεις.	"TEI * "
~	Η μηχανή θα επιστρέψει τις σελίδες που περιλαμβάνουν τον όρο και όλα τα συνώνυμά του.	~διαδίκτυο
..	Ο τελεστής αυτός δηλώνει το εύρος μεταξύ δύο αριθμών, περιορίζοντας έτσι τα αποτελέσματα της αναζήτησης.	"ταινίες γαλλικού κινηματογράφου 2000..2010"

Πέραν, όμως, των βασικών τελεστών που καθορίζουν την επεξεργασία των όρων ενός ερωτήματος από τη μηχανή, υπάρχουν τελεστές που καθορίζουν το πλαίσιο, τον ιστοχώρο, ή τον τύπο του αρχείου στο οποίο θα πραγματοποιηθεί η αναζήτηση.

Πίνακας 3 Τελεστές Εναλλακτικής Αναζήτησης

Τελεστές Εναλλακτικής Αναζήτησης		
Τελεστής	**Περιγραφή**	**Παράδειγμα**
Site:	Η μηχανή αναζητά μόνο στον συγκεκριμένο ιστότοπο.	"Diploma Thesis" site:.teipat.gr
Cache:	Η μηχανή αναζητά αποθηκευμένα "στιγμιότυπα" μίας ιστοσελίδας στη μνήμη της.	cache:teipat.gr
Inurl:	Η μηχανή αναζητά τους όρους του ερωτήματος εντός της διεύθυνσης URL.	inurl: ele
Intitle:	Η μηχανή αναζητά τους όρους του ερωτήματος εντός τίτλων ιστοσελίδων	intitle:"Search Engine Optimisation"
Intext:	Η μηχανή αναζητά τους όρους του ερωτήματος εντός του περιεχόμενου κειμένου των εγγράφων.	intext: ele.teipat.gr
Inanchor	Η μηχανή αναζητά τους όρους του ερωτήματος εντός των anchor texts που περιλαμβάνουν τα έγγραφα.	inanchor:"τα νέα της Πάτρας"
Link:	Η αναζήτηση επιστρέφει τις σελίδες εκείνες που συνδέουν στον συγκεκριμένο ιστότοπο. Παρέχει ένα κατώτατο όριο των πραγματικών συνδέσμων προς τον ιστότοπο.	link:ele.teipat..gr
Related	Πραγματοποιεί αναζήτηση ιστοτόπων σχετικών με τον συγκεκριμένο ιστοχώρο	Related:wikipedia.org
Info:	Παρέχει όλες τις δυνατές πληροφορίες για έναν ιστότοπο, όπως καταγεγραμμένα στιγμιότυπα, συνδέσμους κ.λπ.	info:minedu.gov.gr
Filetype:	Η αναζήτηση πραγματοποιείται μόνο σε συγκεκριμένους τύπους αρχείων.	"crawling and indexing" filetype:pdf

ΚΕΦΑΛΑΙΟ 2

ΤΕΧΝΙΚΕΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΤΗΣ ΙΣΤΟΣΕΛΙΔΑΣ

2.1 Πρωτόκολλο αποκλεισμού ανιχνευτών (spiders)

Με το αρχείο robots.txt, οι κάτοχοι ιστοτόπων έχουν τη δυνατότητα να ορίσουν την προσβασιμότητα των ανιχνευτών στα έγγραφα του ιστοχώρου τους. Το αρχείο αυτό ονομάζεται «Πρωτόκολλο αποκλεισμού ανιχνευτών». Στην ουσία, ένας ανιχνευτής (ή robot, όπως ονομάζεται αλλιώς) επιθυμεί να επισκεφθεί μία διεύθυνση URL, για παράδειγμα τη διεύθυνση http://www.dete.gr. Προτού την επισκεφθεί, ελέγχει τη διεύθυνση http://www.dete.gr/robots.txt , στην οποία διαβάζει τα εξής:

```
User-agent: *
Disallow: /images/banners/
Allow: /
```

Με το παρόν αρχείο, οι διαχειριστές του ιστοχώρου dete.gr επιθυμούν να αποκλείσουν την είσοδο όλων των ανιχνευτών σε εικόνες και τίτλους

2.1.1 Σύνταξη

Υπάρχουν συγκεκριμένες εντολές που μας επιτρέπουν να κατευθύνουμε τους ανιχνευτές Ιστού στις σωστές διευθύνσεις του ιστοχώρου μας, όπως φαίνεται στο παρακάτω robots.txt:

```
#Τα σχόλια τοποθετούνται μετά από το σύμβολο "#" στην αρχή
 #μιας γραμμής ή ακριβώς δεξιά από μία εντολή.
 #Για τον αποκλεισμό όλων των ανιχνευτών από τον ιστοχώρο:
User-agent: *
Disallow: /
#Για να επιτραπεί η πρόσβαση όλων των ανιχνευτών παντού:
User-agent: *
Disallow:
#Εναλλακτικά, μπορούμε να δημιουργήσουμε ένα κενό αρχείο
#robots.txt
#Για τον αποκλεισμό ενός συγκεκριμένου ανιχνευτή:
User-agent: BotName
Disallow: /
#Για να επιτραπεί η πρόσβαση σε ορισμένο ανιχνευτή:
User-agent: BotName
Disallow:
#Για τον αποκλεισμό ανιχνευτή σε ορισμένα αρχεία μόνο και
#όχι ολόκληρο τον φάκελο:
User-agent: BotName
Disallow: /tmp/school/file1.php
Disallow: /tmp/school/file2.html
```

Τέλος, υπάρχουν ορισμένες εντολές που δεν αναγνωρίζονται από όλους τους ανιχνευτές, παρά μόνο από ορισμένους εκ των βασικών μηχανών αναζήτησης, όπως φαίνεται στο παρακάτω αρχείο robots.txt:

```
#Για να επιτραπεί η πρόσβαση ορισμένων ανιχνευτών σε
#κάποιο φάκελο ή αρχείο:
User-agent: GoogleBot
Allow: /folder2
Allow: /folder5/file1.html
#Η εντολή αυτή αναγνωρίζεται από την Google και την Bing.

#Αντίστοιχα, μπορούμε να επιτρέψουμε την πρόσβαση σε ένα
#μόνο αρχείο, αποκλείωντας την πρόσβαση στον υπόλοιπο
#φάκελο:
User-agent: MsnBot
Allow: /folder2/guestbook.html
Disallow: /folder2/
#Για να οριστεί το χρονικό διάστημα αναμονής, σε
#δευτερόλεπτα, μεταξύ δύο διαδοχικών αιτήσεων για επίσκεψη
#στον ίδιο διακομιστή, χρησιμοποιείται η εξής παράμετρος:
User-agent:
* Crawl-delay: sec
#όπου sec μία ακέραια τιμή δευτερολέπτων
#Η χρησιμότητα της εντολής αυτής είναι η αποφυγή
#κατασπατάλησης μεγάλου εύρους διασύνδεσης, εξαιτίας της
#δραστηριότητας των ανιχνευτών.
#Ορισμένοι ανιχνευτές, κυρίως των δημοφιλέστερων μηχανών
#αναζήτησης αναγνωρίζουν την εντολή υπόδειξης της
#θέσης ενός sitemap, δίνοντας τη δυνατότητα χρήσης
#πολλαπλών sitemaps:
Sitemap: http://www.dete.gr/sitemap.xml
Sitemap: http://www.dete.gr/updates/new_sitemap.xml

#Τέλος, ορισμένοι ανιχνευτές αναγνωρίζουν την παρακάτω
#εντολή, για την εξαίρεση συγκεκριμένων τύπων αρχείων:
User-agent: *
Disallow: /*pdf$
Disallow: /*xls$
```

2.2 Meta – Ετικέτες

Οι Meta – ετικέτες (Meta tags) τοποθετούνται στον τομέα <HEAD> μίας σελίδας και διαβάζονται από το φυλλομετρητή και τις μηχανές αναζήτησης. Αυτές οι ετικέτες αποκρύπτονται από τους απλούς χρήστες κι επισκέπτες μίας σελίδας, παρότι είναι διαθέσιμες στον πηγαίο κώδικα, ενώ χρησιμοποιούνται από όλες τις μεγάλες μηχανές αναζήτησης κατά την ευρετηρίαση της σελίδας. Οι σπουδαιότερες εξ αυτών είναι η ετικέτα της περιγραφής (Meta description tag) και η ετικέτα αποκλεισμού ανιχνευτών (Meta robots tag) με ανάλογη λειτουργία και χρησιμότητα με αυτήν του πρωτοκόλλου αποκλεισμού ανιχνευτών, robots.txt. Η συμβολή των ετικετών, όσον αφορά την τεχνική βελτιστοποίησης της ιστοσελίδας στα αποτελέσματα των μηχανών αναζήτησης αμφισβητείται, ήδη από τον Σεπτέμβριο του 2009, τόσο από επαγγελματίες SEOs όσο και κυρίως από τους μηχανικούς της Google (Google Webmaster Central Blog, 2009)[6]. Πάντως η εμπειρία έχει αποδείξει πως η ορθή χρήση και βελτιστοποίηση των Meta – ετικετών, καθιστά τις ιστοσελίδες πιο φιλικές στις μηχανές αναζήτησης, βοηθώντας την θέση που καταλαμβάνουν αυτές στα αποτελέσματα. Μία META ετικέτα περιλαμβάνει ένα γνώρισμα (HTTP-EQUIV ή NAME) που προσδιορίζει τον τύπο της ετικέτας και λαμβάνει αντίστοιχη τιμή, και ένα γνώρισμα (CONTENT) που λαμβάνει ως τιμή το περιεχόμενο που ανατίθεται στον τύπο της ετικέτας. Η σύνταξη των δύο περιπτώσεων φαίνεται παρακάτω:

```
<head>
        <meta name="…" content="…" />
        <meta http-equiv="…" content="…" />
</head>
```

2.2.1 Meta ετικέτα περιγραφής

Οι Meta – περιγραφές παρέχουν συνοπτικές πληροφορίες του περιεχομένου των ιστοσελίδων στους χρήστες. Χρησιμοποιούνται ευρέως από τις μηχανές αναζήτησης κατά την παρουσίαση των αποτελεσμάτων αναζήτησης με σκοπό την υπόδειξη ενός αποσπάσματος προεπισκόπησης μίας δεδομένης σελίδας. Οι ετικέτες αυτές, παρόλο που δεν είναι ιδιαίτερα σημαντικές για την κατάταξη των αποτελεσμάτων, είναι εξαιρετικά σημαντικές για την προσέλκυση επισκεπτών από τις σελίδες των αποτελεσμάτων των μηχανών.

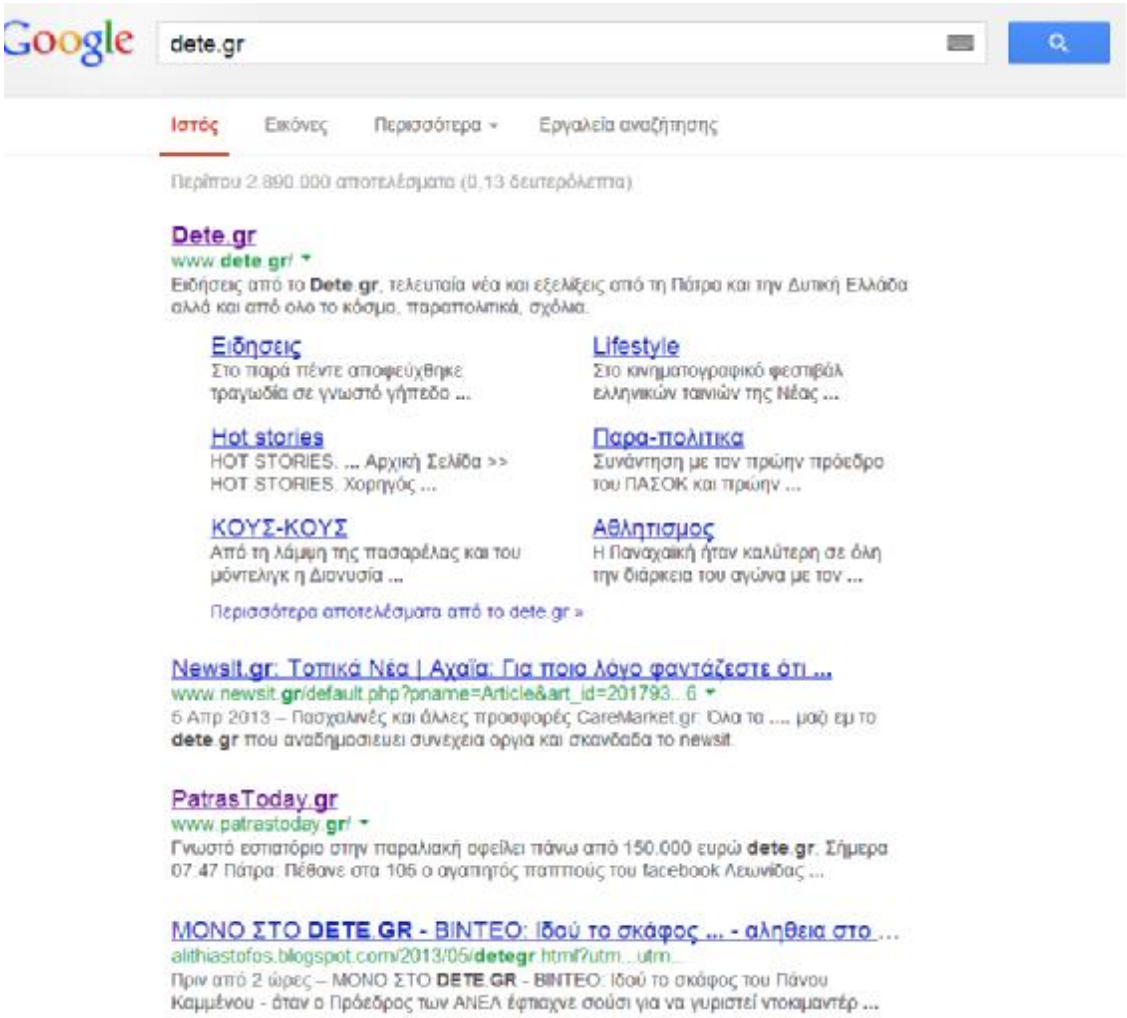
Πρόκειται για μικρές παραγράφους που δίνουν τη δυνατότητα στους διαχειριστές της σελίδας να διαφημίσουν περιεχόμενο στους χρήστες των μηχανών και να τους γνωστοποιήσουν τι σχέση έχει ακριβώς η σελίδα με τον όρο αναζήτησης. Με τον τρόπο αυτό, οι μηχανές αναζήτησης καταφέρνουν να τοποθετήσουν μία σχετική περιγραφή,

κάτω από τον τίτλο του κάθε αποτελέσματος που επιστρέφουν στον χρήστη για μία αναζήτηση, ακόμη κι αν δεν έχει οριστεί κάποια Meta ετικέτα περιγραφής.

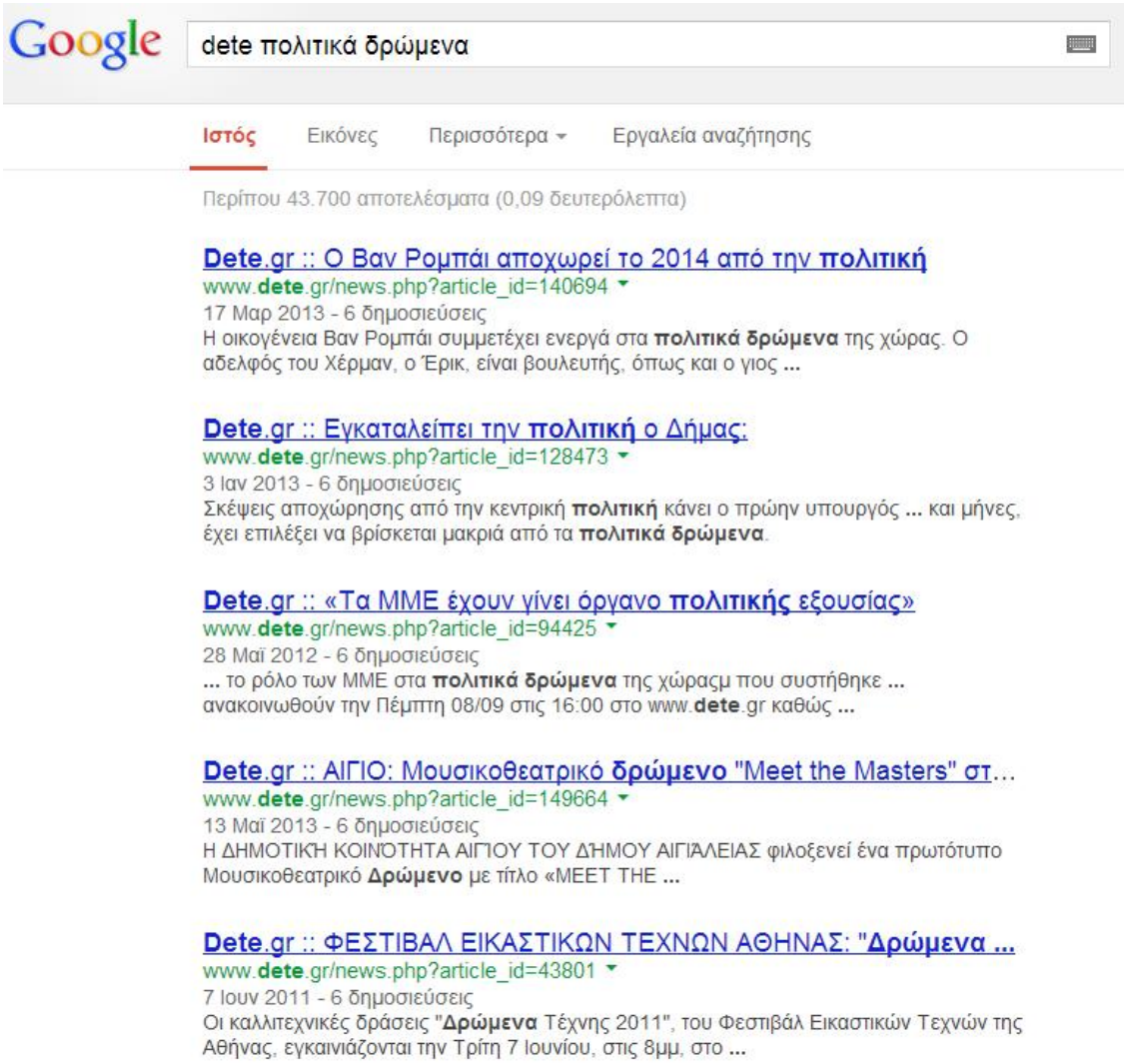
Παράλληλα, οι ΜΕΤΑ ετικέτες δίνουν τη δυνατότητα στον διαχειριστή να ελέγξει λίγο περισσότερο την περιγραφή που συνοδεύει τα αποτελέσματα αναζήτησης, ή και, εάν επιθυμεί, να μην συμπεριλάβει καμία περιγραφή. Για παράδειγμα, η ιστοσελίδα http://www.dete.gr με την ετικέτα

```
<meta name="description" content="Ειδήσεις από το Dete.gr, τελευταία νέα και εξελίξεις από τη Πάτρα και την Δυτική Ελλάδα αλλά και από όλο το κόσμο, παραπολιτικά, σχόλια"/>
```

για δύο διαφορετικές αναζητήσεις, επιστρέφει διαφορετική περιγραφή:



Εικόνα 1 Αποτελέσματα Αναζήτησης 1ου τρόπου



Εικόνα 2 Αποτελέσματα Αναζήτησης 2^{ου} τρόπου

Κατά την πρώτη, δηλαδή, αναζήτηση, επιστρέφει την περιγραφή που ο διαχειριστής έχει επιλέξει μέσω της ετικέτας meta description, ενώ έπειτα, στη δεύτερη αναζήτηση, επιστρέφει την ίδια σελίδα, με τον ίδιο τίτλο και διαφορετική περιγραφή. Η διαφορά αυτή έγκειται στο γεγονός ότι ο όρος της αναζήτησης δεν εμπεριέχεται στον τίτλο ή την ετικέτα περιγραφής, αλλά στο περιεχόμενο του εγγράφου.

Στην πρώτη περίπτωση, ο όρος αναζήτησης βρέθηκε στον τίτλο της σελίδας (page title), οπότε η μηχανή της Google επέλεξε να παρουσιάσει το αποτέλεσμα, παρέχοντας την επιλεγμένη περιγραφή του διαχειριστή του ιστοτόπου. Οι μηχανές αναζήτησης λοιπόν,

όσον αφορά την περιγραφή των ιστοσελίδων που επιστρέφονται ως αποτελέσματα, δίνουν προτεραιότητα στον τίτλο και τη meta ετικέτα περιγραφής.

2.2.1.1 Σύνταξη

Όπως όλες οι META ετικέτες, η περιγραφή γράφεται στον <HEAD> τομέα του HTML κώδικα ως εξής:

```
<head>
        <meta name="description" content="Το παρόν κείμενο αποτελεί παράδειγμα meta περιγραφής." />
</head>
```

2.2.2 Meta ετικέτα ανιχνευτών

Η ετικέτα ανιχνευτών (Meta robots tag) χρησιμοποιείται για να ελέγχεται η δραστηριότητα όλων των ανιχνευτών των μηχανών αναζήτησης σε επίπεδο σελίδας, και όχι σε επίπεδο διεύθυνσης ή διακομιστή όπως το robots.txt αρχείο. Υπάρχουν αρκετές λειτουργίες – τιμές που μπορούν να ανατεθούν στην ετικέτα αυτή, όπως φαίνεται παρακάτω.

2.2.2.1 Σύνταξη

Η ετικέτα meta robots, όπως όλες οι meta ετικέτες, συμπεριλαμβάνεται στον HEAD τομέα του HTML κώδικα και με παρόμοιο τρόπο, στο γνώρισμα robots, ανατίθενται οι διάφορες τιμές. Όπως προσδιορίζεται, πέραν των οδηγιών που απευθύνονται προς ανιχνευτές συγκεκριμένων μηχανών αναζήτησης, υπάρχουν ορισμένες τιμές που δεν αναγνωρίζονται από όλους τους ανιχνευτές.

```
<head>
<meta name="robots" content="VALUE, ...VALUE" />
</head>
```

Η τιμή VALUE μπορεί να πάρει τις εξής τιμές:

- Index/NoIndex

  Η τιμή αυτή επιτρέπει ή δεν επιτρέπει στις μηχανές αναζήτησης να ευρετηριάσουν τη συγκεκριμένη σελίδα. Με την ανάθεση της τιμής “noindex”, δηλαδή, η σελίδα θα εξαιρεθεί από τις μηχανές αναζήτησης. Ως προεπιλογή, όλες οι μηχανές αναζήτησης δίνουν αυτόματα την τιμή “index” στο συγκεκριμένο όρισμα, εκτός εάν έχει ορισθεί διαφορετική τιμή.

- Follow/NoFollow

  Η τιμή αυτή ενημερώνει τις μηχανές εάν οι σύνδεσμοι που περιλαμβάνονται στην σελίδα πρέπει να ακολουθηθούν και να ευρετηριασθούν. Με την επιλογή της τιμής “nofollow”, οι ανιχνευτές θα αγνοήσουν όλους τους συνδέσμους εντός της σελίδας, τόσο για σκοπούς ανακάλυψης και προσθήκης στο crawl frontier, όσο και για σκοπούς κατάταξης. Ως προεπιλογή, οι μηχανές αναζήτησης υποθέτουν ότι όλες οι σελίδες έχουν την τιμή “Follow”, ακολουθούν, δηλαδή, όλους τους συνδέσμους για να συνεχίσουν την ομαλή τους λειτουργία.

- Noarchive

  Η τιμή αυτή χρησιμοποιείται για να απαγορεύσει στις μηχανές αναζήτησης να αποθηκεύσουν κάποιο στιγμιότυπο – αντίγραφο (cached copy) της σελίδας. Ως προεπιλογή, οι μηχανές διατηρούν ορατά και διαθέσιμα προς τους χρήστες τους αντίγραφα όλων των σελίδων που επισκέπτονται και ευρετηριάζουν.

- NoODP

  Η τιμή αυτή αποτελεί εξειδικευμένη οδηγία προς ορισμένους ανιχνευτές (Google, Yahoo!, Bing), ενημερώνοντας τις μηχανές να μην αντικαταστήσουν την περιγραφή της σελίδας με αυτήν που εμφανίζεται στην καταχώρησή της στο Open Directory Project (κατάλογος DMOZ), αλλά να χρησιμοποιήσουν την τιμή της meta ετικέτας περιγραφής ή κάποιου πιο σχετικού αποσπάσματος από το περιεχόμενο της σελίδας. Ως προεπιλογή, οι μηχανές αντικαθιστούν τον τίτλο και την περιγραφή των αποτελεσμάτων με αυτά της αντίστοιχης καταχώρησης της σελίδας στο Open Directory Project, μόνο για την αρχική σελίδα.

§ NoYDir

Όπως και η τιμή NoODP, ενημερώνει αποκλειστικά τη μηχανή αναζήτησης της Yahoo! να μην επιλέξει για την εμφάνιση της σελίδας στα αποτελέσματα εκείνη την περιγραφή που έχει δοθεί στη σελίδα στον κατάλογο Yahoo! Directory. Οι υπόλοιπες μηχανές αναζήτησης την αγνοούν, καθώς δεν αναγνωρίζουν τον κατάλογο της Yahoo! ως τον επίσημο διαδικτυακό κατάλογο. Αντίστοιχα με το ODP, η αντικατάσταση, από προεπιλογή, της περιγραφής με αυτήν από τον κατάλογο της Yahoo! πραγματοποιείται μόνο για την αρχική σελίδα.

§ Unavailable_after:[ημερομηνία]

Η τιμή αυτή, με την ανάθεση και της επιθυμητής ημερομηνίας, αφαιρεί την σελίδα από τα αποτελέσματα των μηχανών αναζήτησης μετά το πέρας της ημερομηνίας αυτής. Η λειτουργία αυτή κρίνεται ιδιαίτερα χρήσιμη σε περιπτώσεις διαγωνισμών ή χρονικά ορισμένων προσφορών, διαφημιστικών καμπανιών και άλλων γεγονότων που ορίζονται και περιορίζονται χρονικά (Google Blog, 2007).

2.2.3 Άλλες χρήσιμες meta ετικέτες

Υπάρχουν πολλές ακόμη meta ετικέτες που μπορούν να χρησιμοποιηθούν για την τεχνική βελτιστοποίηση μίας ιστοσελίδας, άλλες γενικές και άλλες ιδιαίτερα εξειδικευμένες, αλλά όλες με πολύ μικρότερη σημασία και βαρύτητα από τις ετικέτες περιγραφής και ανιχνευτών. Οι σημαντικότερες εξ αυτών αναλύονται παρακάτω:

§ Ετικέτα μετάφρασης

```
<head>
<meta name="google" content="notranslate" />
 </head>
```

Όταν η μηχανή αναζήτησης της Google αντιλαμβάνεται ότι το περιεχόμενο μίας σελίδας δεν είναι γραμμένο στη γλώσσα που ο χρήστης δύναται ή ενδέχεται να επιθυμεί να διαβάσει (ανάλογα με τις προτιμήσεις της γλώσσας του λογαριασμού του στην Google, ήτη μητρική γλώσσα που ορίζεται από τον πάροχο σύνδεσης), συχνά προσφέρει έναν επιπλέον σύνδεσμο, κάτω από την περιγραφή του αποτελέσματος, για μία αυτόματη

μετάφραση της σελίδας. Σε γενικές γραμμές, με τον τρόπο αυτό δίνεται η δυνατότητα στον διαχειριστή της σελίδας να απευθύνει το μοναδικό περιεχόμενο της σελίδας του σε αρκετά μεγαλύτερο αριθμό χρηστών και δεν επηρεάζει καθόλου (αρνητικά) την επίδοση του συγκεκριμένου ιστοτόπου. Με την ανάθεση της τιμής “notranslate”, ο διαχειριστής απαγορεύει στη μηχανή της Google την παροχή δυνατότητας μετάφρασης.

§ Ετικέτα ανακατεύθυνσης

```
<head>

<meta http-equiv="refresh" content="…;url=…" /> </head>
```

Αυτή η meta ετικέτα στέλνει τον χρήστη σε ένα διεύθυνση URL, μετά από ένα ορισμένο χρονικό διάστημα, και συνήθως χρησιμοποιείται ως μία απλουστευμένη (συγκριτικά με την ανακατεύθυνση 301) μορφή ανακατεύθυνσης.

§ Ετικέτες σύνδεσης ιστοσελίδας με το Facebook

Η πιο δημοφιλής σελίδα κοινωνικής δικτύωσης (social networking) είναι το Facebook, επί του παρόντος, και χρησιμοποιείται από τις εταιρείες για λόγους προώθησης προϊόντων και υπηρεσιών, εξυπηρέτησης πελατών, ανατροφοδότησης από το καταναλωτικό κοινό, διαφήμισης γεγονότων και δραστηριοτήτων και, κυρίως, επικοινωνίας με το target group αυτών. Για λόγους διασύνδεσης ενός ιστοτόπου με την αντίστοιχη επίσημη σελίδα στο facebook και επικύρωσης της κατοχής αυτού, χρησιμοποιούνται οι παρακάτω ετικέτες:

```
<html xmlns="http://www.w3.org/1999/xhtml"
xmlns:fb="http://www.facebook.com/2008/fbml" xml:lang="en" lang="en"
> <head>

<meta property="fb:admins" content="…, …, …" />

<meta property="fb:page_id" content="…" />

</head>

</html>
```

Οι τιμές των ορισμάτων “fb:admins” και “fb:page_id” προσδιορίζονται από τον πίνακα ελέγχου διαχειριστή του facebook λογαριασμού, κατά τη διεκπεραίωση της διαδικασίας σύνδεσης ιστοσελίδας με την σελίδα facebook, με το πρώτο όρισμα να αφορά την ταυτότητα καθενός εκ των διαχειριστών και το δεύτερο την ταυτοποίηση της σελίδας. Όπως φαίνεται στο παραπάνω παράδειγμα, οφείλουμε να δηλώσουμε την υιοθέτηση όλων των απαραίτητων προτύπων, μαζί με αυτό για τη σύνταξη HTML/XHTML και το πρότυπο της facebook.

§ Ετικέτες πρωτοκόλλου Open Graph

Σε άμεση σχέση με τις προηγούμενες ετικέτες σύνδεσης με την σελίδα facebook, το Πρωτόκολλο Open Graph επιτρέπει την είσοδο των σελίδων ενός ιστοτόπου στο γράφημα κοινωνικής δικτύωσης του Facebook. Προς το παρόν, χρησιμοποιείται για σελίδες που αναπαριστούν πράγματα, ανθρώπους, δραστηριότητες, ταινίες, ομάδες, διασημότητες, ξενοδοχεία, εστιατόρια, οργανισμούς, αλλά και οποιαδήποτε διάσημη σελίδα μπορεί να γίνει “like” στο facebook profile των μελών. Με τη χρήση των meta ετικετών αυτών, η σελίδα λειτουργεί σαν μία σελίδα facebook, επιτρέποντας λειτουργίες που επιτρέπονται μέσα στον δημοφιλή αυτό ιστότοπο κοινωνικής δικτύωσης. Με τον τρόπο αυτό, η σελίδα εμφανίζεται στα ίδια σημεία με τις σελίδες facebook μέσα στον ιστοχώρο του Facebook, ενώ δίνεται η δυνατότητα στόχευσης διαφημίσεων σε άτομα που τους «αρέσει» το περιεχόμενο της σελίδας. Τέτοιες ετικέτες, είναι οι παρακάτω:

```
<html xmlns=http://www.w3.org/1999/xhtml
 xmlns:og=http://ogp.me/ns#
 xmlns:fb="http://www.facebook.com/2008/fbml">
<head>
<title>Κάποιος σχετικός τίτλος με το αντικείμενο</title>
<meta property="og:title" content="Τίτλος σελίδας"/>
<meta property="og:type" content="Τύπος σελίδας"/>
<meta property="og:url" content="Διεύθυνση URL της σελίδας"/>
<meta property="og:image" content="Διεύθυνση URL της εικόνας του
αντικειμένου που αναπαριστά η σελίδα"/>
<meta property="og:site_name" content="Όνομα ιστοχώρου"/>
<meta property="fb:admins" content="..."/>
<meta property="og:description" content="Μία περιγραφή του
αντικειμένου ή της σελίδας"/>
</head>
</html>
```

2.2.4 Ετικέτες σήμανσης περιεχομένου

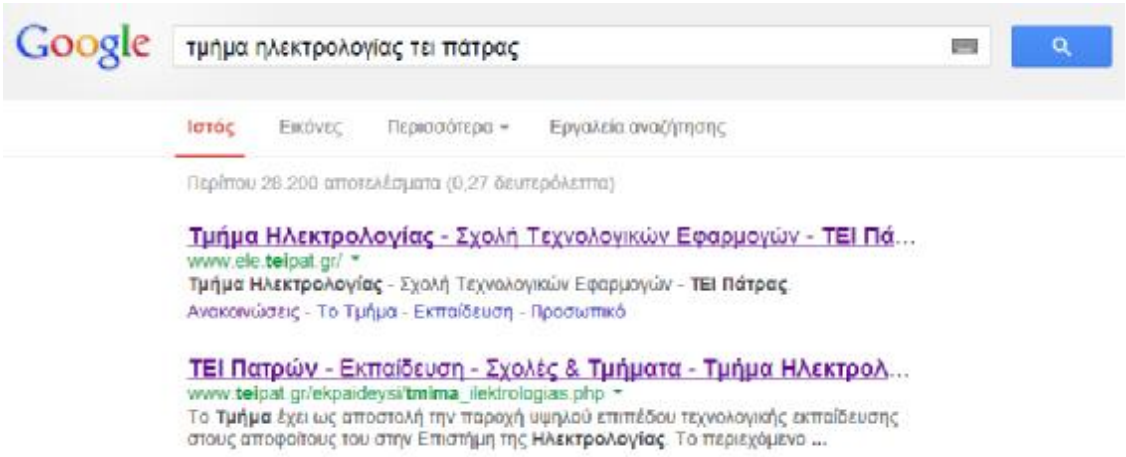
2.2.4.1 Τίτλος σελίδας

Η ετικέτα τίτλου (Page title tag) οφείλει να αποτελεί την ακριβή, συνοπτική περιγραφή του περιεχομένου μίας ιστοσελίδας. Μετά το ίδιο το περιεχόμενο του εγγράφου, αποτελεί τον σημαντικότερο παράγοντα βελτιστοποίησης της σελίδας και εμφανίζεται σε τρία πολύ σημαντικά επίπεδα:

α) Στην κορυφή του φυλλομετρητή του χρήστη, καθώς και ως όνομα της αντίστοιχης καρτέλας του φυλλομετρητή. Έρευνες έχουν δείξει, όμως, ότι ο χρήστης δίνει περιορισμένη προσοχή στα σημεία εκείνα, παρ' όλα αυτά διευκολύνει την εμπειρία του κατά τη διαχείριση πολλαπλών καρτελών.

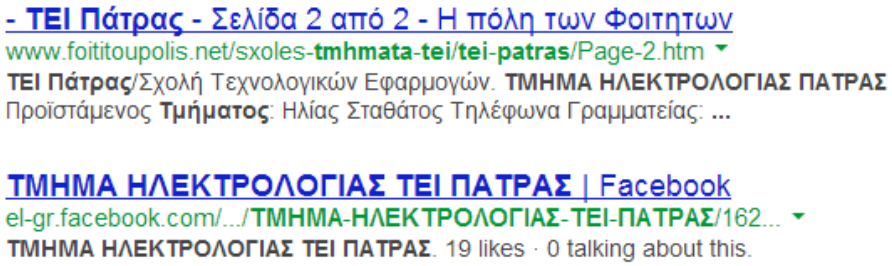
Εικόνα 3 Εμφάνιση του τίτλου σελίδας στο φυλλομετρητή

β) Στα αποτελέσματα των μηχανών αναζήτησης, δίνοντας μία πρώτη συνοπτική αλλά σαφή εικόνα του περιεχομένου προτού ο χρήστης εισέλθει στο έγγραφο αυτό. Μάλιστα, σε περίπτωση που κάποιος εκ των όρων της αναζήτησης συναντάται στο <title> tag, τότε ο όρος αυτός γράφεται με bold στα αποτελέσματα της αναζήτησης και, δεδομένης της μεγαλύτερης γραμματοσειράς του τίτλου, ελκύει άμεσα την προσοχή του χρήστη στη σελίδα αυτή, που είναι και ο απώτερος στόχος της βελτιστοποίησης.



Εικόνα 4 Εμφάνιση τίτλου σελίδας στα αποτελέσματα αναζήτησης

γ) Ως anchor text σε ορισμένους εξωτερικούς ιστοχώρους (και ιδιαίτερα τις σελίδες κοινωνικής δικτύωσης), όταν προτείνεται (ή «μοιράζεται») ένας σύνδεσμος ως ενδιαφέρων.



Εικόνα 5 Εμφάνιση τίτλου σελίδας στο anchor text ορισμένων συνδέσμων

2.2.4.2 Σύνταξη

Ο τίτλος της σελίδας γράφεται στον <head> τομέα του HTML αρχείου, μεταξύ των ετικετών <title> και </title>.

```
<head> <title> Τμήμα Ηλεκτρολογίας</title> </head>
```

2.3 Επικεφαλίδες

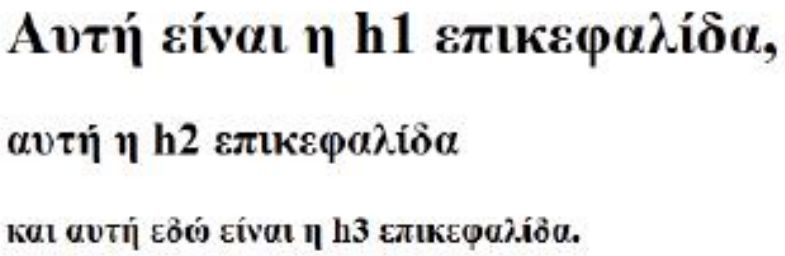
Πάρα πολλοί διαχειριστές και συντάκτες ιστοσελίδων στο Διαδίκτυο αγνοούν την ύπαρξη και την αξία των πρότυπων επικεφαλίδων (h1, h2, h3 headings) και είτε δεν χρησιμοποιούν γενικότερα επικεφαλίδες είτε συντάσσουν μικρά κείμενα στα οποία αποδίδουν διαφορετικά γνωρίσματα μεγέθους και χρώματος γραμματοσειράς, μέσω του αρχείου CSS (Cascading Style Sheet), για τη διαφοροποίηση αυτών από το υπόλοιπο κείμενο. Η χρήση επιφανών επικεφαλίδων και υποτίτλων που περιλαμβάνουν φράσεις – κλειδιά, μέσα σε ένα έγγραφο, κατά τη σύνταξη του περιεχομένου του, όμως, είναι ιδιαίτερα σημαντική για τις μηχανές αναζήτησης, αλλά βελτιώνει και τη χρηστικότητα και προσβασιμότητα των επισκεπτών. Οι αλγόριθμοι των μηχανών αναζήτησης εκλαμβάνουν το κείμενο που υπάρχει εντός των πρότυπων επικεφαλίδων ως πιο σημαντικό από το ίδιο το περιεχόμενο της σελίδας, καθώς θεωρούν ότι οι τίτλοι σχετίζονται θεματικά με αυτό. Σημειώνεται, εδώ, ότι ορισμένα συστήματα διαχείρισης περιεχομένου (CMSs) δεν υποστηρίζουν πλήρως τη χρήση επικεφαλίδων, καθώς μεταφράζουν τον υποτιτλισμό σε κείμενο με διαφοροποιημένη, πιο έντονη κι ευδιάκριτη γραμματοσειρά, ή αυτομάτως μεταφέρουν το περιεχόμενο της ετικέτας τίτλου της σελίδας στην ετικέτα επικεφαλίδας, κάτι που, προφανώς, δεν είναι επιθυμητό καθώς στερεί τον πλήρη έλεγχο όλων των πιθανών στοιχείων βελτιστοποίησης από τον διαχειριστή της ιστοσελίδας ενώ ενέχει τον κίνδυνο οι αλγόριθμοι να επιβάλλουν ποινή για κατάχρηση φράσεων – κλειδιών μέσα στη σελίδα (spam).

2.3.1 Σύνταξη

Παρακάτω, παρουσιάζεται η σύνταξη των τριών τύπων επικεφαλίδων, οι οποίες εντάσσονται στο κυρίως μέρος της ιστοσελίδας (τομέας <BODY></BODY>):

```
<html>
<head>
<title>Headings</title>
</head>
<body>
 <h1>Αυτή είναι η h1 επικεφαλίδα,</h1>
<h2>αυτή η h2 επικεφαλίδα</h2>
<h3>και αυτή εδώ είναι η h3 επικεφαλίδα.</h3>
 </body>
</html>
```

Το παραγόμενο αποτέλεσμα στο φυλλομετρητή είναι το εξής:



Εικόνα 6 Μορφή επικεφαλίδων στο φυλλομετρητή

Εικόνα 10 Οι διάφορες επικεφαλίδες στο φυλλομετρητή

Παρατηρούμε ότι το μέγεθος της γραμματοσειράς, σε κάθε μία από τις επικεφαλίδες, διαφέρει. Ανάλογα διαφοροποιείται και η βαρύτητα του περιεχομένου της κάθε ετικέτας στις μηχανές αναζήτησης, με μεγαλύτερη αξία να έχει αυτή της h1 επικεφαλίδας.

2.4 Μορφοποίηση Κειμένου

2.4.1 Σύνδεσμοι (links)

Οι υπερσύνδεσμοι μίας ιστοσελίδας αποτελούν την κινητήρια δύναμη της κατάταξης των ιστοσελίδων στα αποτελέσματα αναζήτησης και διαχωρίζονται σε εισερχόμενους (inbound) και εξερχόμενους (outbound). Εισερχόμενοι ονομάζονται οι σύνδεσμοι που βρίσκονται σε κάποια άλλη ιστοσελίδα προς την ιστοσελίδα που εξετάζουμε, ενώ εξερχόμενοι ονομάζονται εκείνοι που βρίσκονται πάνω στην εξεταζόμενη ιστοσελίδα προς κάποια άλλη ιστοσελίδα. Οι εξερχόμενοι μπορούν να διαχωριστούν, με τη σειρά τους, σε εσωτερικούς και εξωτερικούς συνδέσμους, ανάλογα με το αν η ιστοσελίδα στην οποία συνδέουν είναι εντός κι εκτός του ίδιου ιστοχώρου, αντίστοιχα.

Από την άποψη της τεχνικής βελτιστοποίησης, πολύ έντονο ενδιαφέρον έχουν όλοι οι εισερχόμενοι σύνδεσμοι, αλλά άμεση ή έμμεση σχέση με την επίδοση της ιστοσελίδας στις μηχανές αναζήτησης έχουν και όλοι οι εξερχόμενοι σύνδεσμοι, είτε αυτοί μεταφέρουν τον χρήστη σε κάποια άλλη σελίδα του ίδιου ιστοχώρου είτε τον κατευθύνουν σε ξένο διακομιστή. Οι μεγαλύτερες μηχανές αναζήτησης χρησιμοποιούν αρκετούς παράγοντες για τον προσδιορισμό της αξίας των συνδέσμων και μίας ιστοσελίδας βάση αυτών, οι βασικοί από τους οποίους είναι οι εξής:

- Η αξιοπιστία του ιστοχώρου που συνδέει προς την ιστοσελίδα
- Η δημοτικότητα του ιστοχώρου που συνδέει προς την ιστοσελίδα
- Η σχετικότητα του περιεχομένου μεταξύ των δύο ιστοσελίδων
- Το κείμενο που πλαισιώνει τον σύνδεσμο (anchor text, γειτονικές φράσεις, τίτλος)

- Ο αριθμός των συνδέσμων προς την ίδια σελίδα από τον αρχικό ιστοχώρο
- Ο αριθμός των διαφορετικών ιστοχώρων που συνδέουν προς την ιστοσελίδα
- Η διαφορετικότητα των anchor texts που περιγράφουν το σύνολο των συνδέσμων
- Η εμπορική σχέση ή σχέση ιδιοκτησίας μεταξύ των δύο ιστοχώρων

Δηλαδή, μεγάλη σημασία έχει τόσο η ποσότητα όσο και η ποιότητα των συνδέσμων από και προς μία ιστοσελίδα.

2.4.1.1 Σύνταξη

Μέσω του HTML αρχείου της ιστοσελίδας πραγματοποιείται ο έλεγχος μόνο των εξερχόμενων υπερσυνδέσμων. Η βασική σύνταξη που ακολουθείται για τη δημιουργία ενός συνδέσμου είναι η εξής:

```
<a href=”link_url”>Anchor Text</a>
```

Η πλέον αποτελεσματική σύνταξη, όμως, είναι αυτή που περιλαμβάνει και μία περιγραφή, με το γνώρισμα τίτλου εντός της ετικέτας <a href>, ως εξής:

```
<a href=”linkurl” title=”Περιγραφή”>Anchor Text</a>
```

Τέλος, όσον αφορά τους εξωτερικούς εξερχόμενους συνδέσμους και όπως θα αναλυθεί στο επόμενο κεφάλαιο εκτενέστερα, έχουμε τη δυνατότητα να αποκλείσουμε την επίσκεψη των ανιχνευτών μέσω αυτών και να επιτρέψουμε μόνο την ανακατεύθυνση των χρηστών, με την παρακάτω σύνταξη:

```
<a href="linkurl" title="Περιγραφή" rel="nofollow">Anchor
Text</a>
```

Η τιμή “nofollow”, στο γνώρισμα “rel=”, έχει ακριβώς την ίδια λειτουργία με αυτήν που ανατίθεται στη meta ετικέτα ανιχνευτών, δίνοντας την οδηγία στους ανιχνευτές των

μηχανών αναζήτησης να μην ακολουθήσουν τον σύνδεσμο. Με τον τρόπο αυτό, δεν μεταφέρουμε αξία από τη μία σελίδα στην άλλη και, όπως θα αναλυθεί στη συνέχεια, τη διατηρούμε για να την κατευθύνουμε εκεί όπου θέλουμε εμείς. Σημειώνεται εδώ ότι η συνολική αξία μίας ιστοσελίδας (σε PageRank) διαιρείται και διαμοιράζεται εξίσου σε όλους τους συνδέσμους που περιλαμβάνει. Αποτρέποντας, επομένως, τη διαρροή αξίας προς έναν εξωτερικό σύνδεσμο, δίνεται η δυνατότητα να προσφερθεί μεγαλύτερη αξία στους εσωτερικούς συνδέσμους που περιλαμβάνονται στην εξεταζόμενη ιστοσελίδα.

2.4.2 Εικόνες

Η ετικέτα <img> χρησιμοποιείται για να χαρακτηρίσει και να περιγράψει τις εικόνες στις μηχανές αναζήτησης, βελτιστοποιώντας την ευρετηρίαση των εικόνων.

2.4.2.1 Σύνταξη

Η βασική σύνταξη που ακολουθείται για την ενσωμάτωση εικόνων σε ένα HTML αρχείο είναι η εξής:

```
<img src="image_url">
```

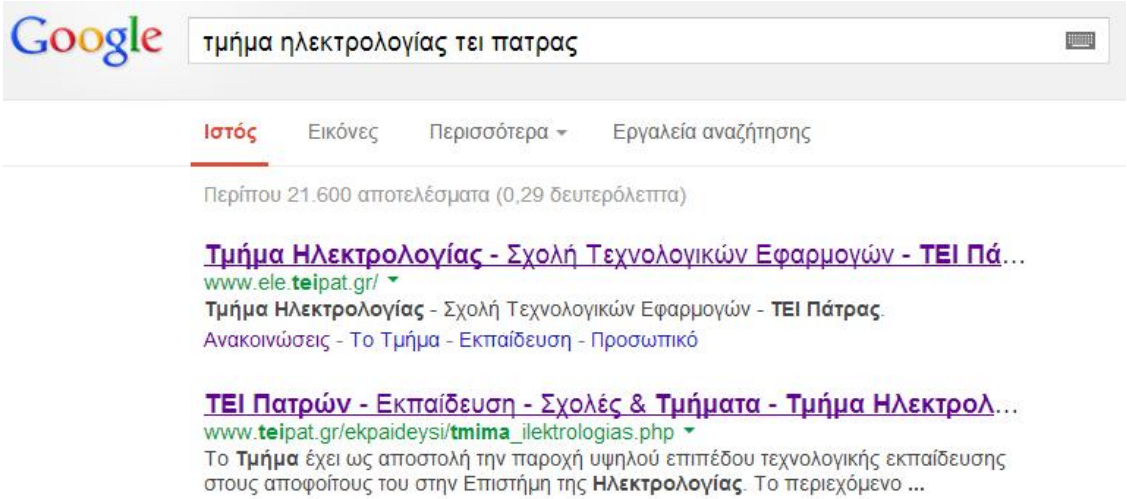
Οι βασικοί παράγοντες που επηρεάζουν θετικά την εικόνα της ιστοσελίδας στις μηχανές αναζήτησης είναι το όνομα του αρχείου της εικόνας και, κατ' επέκταση, η διεύθυνση URL αυτού, η περιγραφή της εικόνας στο γνώρισμα "title=" και το εναλλακτικό κείμενο που περιγράφει την εικόνα στο γνώρισμα "alt=". Η τιμή του γνωρίσματος "alt=" εμφανίζεται αντί της εικόνας, μέχρι να φορτωθεί πλήρως η εικόνα στον φυλλομετρητή. Η σύνταξη των δύο αυτών στοιχείων φαίνεται στο παρακάτω παράδειγμα:

```
<img src="image_url" title="Περιγραφή εικόνας"
alt="Εναλλακτικό κείμενο">
```

2.4.3 Δομή URL

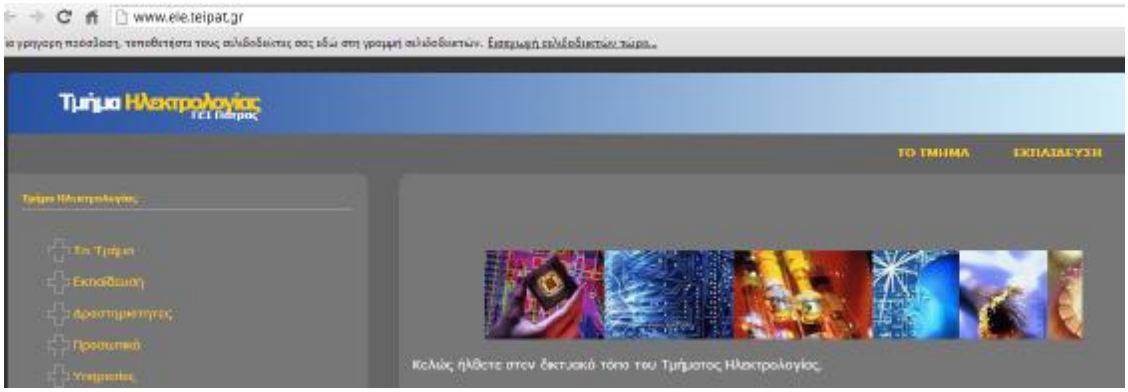
Όσον αφορά την αναζήτηση, είτε πρόκειται για μηχανές κι ανιχνευτές είτε για φυσικούς επισκέπτες, οι διαδικτυακές διευθύνσεις των εγγράφων, URLs, έχουν ιδιαίτερα μεγάλη αξία καθώς εμφανίζονται σε τρεις διαφορετικές σημαντικές τοποθεσίες:

α) Στα αποτελέσματα αναζήτησης, όπου παρουσιάζεται, με πράσινο χρώμα, κάτω από τον τίτλο και την περιγραφή του κάθε αποτελέσματος.



Εικόνα 7 Παράδειγμα εμφάνισης της δομής URL στα αποτελέσματα αναζήτησης

β) Στη μπάρα διευθύνσεων του φυλλομετρητή, όπου, ενώ δεν επηρεάζει τις μηχανές αναζήτησης, έχει άμεση επίδραση στην εμπειρία του χρήστη που επισκέπτεται τη σελίδα, κυρίως σε σχέση με την ευκολία πλοήγησης μέσα στον ιστοχώρο.



Εικόνα 8 Παράδειγμα εμφάνισης της δομής URL στο φυλλομετρητή

γ) Ως anchor text σε πάρα πολλές περιπτώσεις εξωτερικών συνδέσμων, στους οποίους παραλείφθηκε η ανάθεση κάποιας περιγραφής ή φράσης (κάτι που παρατηρείται έντονα σε blogs και forums, όπου οι συντάκτες δεν είναι επαγγελματίες διαχειριστές ιστοσελίδων και δεν ενδιαφέρονται να περιγράψουν ή να παρουσιάσουν τέλεια κάποια πληροφορία, παρά μόνο να τη μεταδώσουν)

Εικόνα 9 Παράδειγμα εμφάνισης της δομής URL σε συνδέσμους χωρίς anchor text

Οι διευθύνσεις URL, επομένως, ενός ιστοχώρου είναι ένα από τα πλέον σημαντικά, προγραμματιστικού επιπέδου στοιχεία, τόσο για τις μηχανές αναζήτησης όσο και για τους χρήστες. Οι μηχανές ανταμείβουν σύντομες, στατικές, σαφείς διευθύνσεις URL που εμπεριέχουν σημαντικές λέξεις ή φράσεις – κλειδιά και, διόλου τυχαία, οι χρήστες εκτιμούν τις ίδιες ακριβώς αξίες.

2.5 Χάρτες ιστοτόπων

Στην πιο απλή του μορφή, ο χάρτης ιστοτόπου είναι ένα αρχείο XML το οποίο καταγράφει τις διευθύνσεις URL για έναν ιστότοπο, μαζί με ορισμένα επιπρόσθετα meta δεδομένα για κάθε ένα έγγραφο (όπως την τελευταία ενημέρωσή του, τη συχνότητα αλλαγής του, την σπουδαιότητά του σε σχέση με τις υπόλοιπες URLs του ιστοτόπου), ώστε να μπορούν οι μηχανές αναζήτησης να ανιχνεύσουν ολόκληρο τον ιστότοπο πιο έξυπνα, γρήγορα και αποτελεσματικά. Το εργαλείο ανακοινώθηκε από την εταιρεία Google το 2005 κι επιτρέπει την καταγραφή διευθύνσεων ιστοσελίδων προς ανίχνευση. Έκτοτε, οι δυνατότητες ενημέρωσης των μηχανών αναζήτησης έχουν πολλαπλασιαστεί και διευρυνθεί, καθώς, πλέον, επιτρέπεται η καταχώρηση πολλαπλών χαρτών καθώς και η σύνταξη χαρτών των βίντεο, των εικόνων, των ειδήσεων, ιστοτόπων που προορίζονται για κινητά τηλέφωνα, ακόμη και του συνόλου των χαρτών που υπάρχουν σε ένα διακομιστή. Φαινομενικά, η σύνταξη ενός χάρτη ιστοτόπου, παρότι προαιρετική, διευκολύνει μόνο τη δουλειά των μηχανών αναζήτησης. Όμως, η δημιουργία ενός sitemap ενέχει κινδύνους για τον κάτοχο ή διαχειριστή ενός ιστοτόπου, ενώ προσφέρει πλεονεκτήματα τόσο για τις μηχανές αναζήτησης όσο και για τον ίδιο τον ιστότοπο.

Πλεονεκτήματα

1. Ο χάρτης μπορεί να καταγράφει όλες τις διευθύνσεις URL από τον ιστότοπο, συμπεριλαμβανομένων των σελίδων που, διαφορετικά, δεν είναι προσβάσιμες από τους ανιχνευτές και τις μηχανές αναζήτησης.
2. Οι μηχανές αναζήτησης αποκτούν περισσότερες πληροφορίες σχετικά με την ανανέωση των ιστοσελίδων, με αποτέλεσμα να βελτιστοποιείται η συχνότητα επίσκεψης των ανιχνευτών σε αυτές.
3. Με τη χρήση της προαιρετικής ετικέτας για την προτεραιότητα μίας σελίδας, στο χάρτη ιστοτόπου, υποδεικνύεται στους ανιχνευτές Ιστού πόσο σημαντική είναι μία σελίδα σχετικά με τις υπόλοιπες σελίδες στον ιστότοπο, δίνοντας σαφέστερη εικόνα για το πρωτεύον αντικείμενο του ιστοχώρου κι επιτρέποντας, έτσι, στις μηχανές αναζήτησης να θέσουν προτεραιότητες στην ανίχνευση του ιστοτόπου, βελτιώνοντας τα αποτελέσματά τους.

Μειονεκτήματα

1. Ένα σημαντικό πρόβλημα εποπτείας του ιστοτόπου προκύπτει από την ανίχνευση σελίδων που υποδεικνύονται από το χάρτη ιστοτόπου αλλά δε θα ήταν προσβάσιμες, χωρίς αυτόν. Αυτό σημαίνει ότι τυχόν προβλήματα ή ελαττώματα της αρχιτεκτονικής και δομής του ιστοχώρου δεν είναι ευδιάκριτα στο διαχειριστή αλλά αποκρύπτονται.
2. Όλες οι πληροφορίες που αναφέρονται στο χάρτη ιστοτόπου δεν είναι διαθέσιμες αποκλειστικά στις μηχανές αναζήτησης, αλλά και στους ανταγωνιστές αυτού του ιστοτόπου. Με άλλα λόγια, λίγο ή πολύ σημαντικές πληροφορίες (όπως η

προτεραιότητα των σελίδων) για τον ιστότοπο γίνονται άμεσα προσβάσιμες προς όλους τους χρήστες, άρα και τους ανταγωνιστές.

3. Όσον αφορά τους τεράστιους ιστοχώρους με πολλές χιλιάδες ιστοσελίδων που παράγονται δυναμικά (φόρουμ, διαδικτυακά καταστήματα, κλπ), η καταγραφή όλων των διευθύνσεων URL είναι μία επίπονη διαδικασία. Για το λόγο αυτό, χρησιμοποιούνται ορισμένα λογισμικά παραγωγής XML χαρτών ιστοτόπου, τα οποία σαρώνουν τον ιστότοπο με έναν τρόπο όμοιο με αυτό της ανίχνευσης από τις μηχανές αναζήτησης.

2.5.1 Γενικοί χάρτες XML

Πρόκειται για χάρτες ιστοτόπων που δηλώνουν στις μηχανές αναζήτησης ένα σύνολο από διευθύνσεις URL που απαρτίζουν έναν ιστότοπο. Η καταγραφή όλων των διευθύνσεων URL του ιστοτόπου δεν είναι υποχρεωτική και προτιμάται η καταγραφή των διευθύνσεων των σημαντικότερων εγγράφων, ή εκείνων που, σύμφωνα με την εμπειρία, δεν είναι προσβάσιμες από τις μηχανές αναζήτησης.

2.5.1.1 Σύνταξη

Η σύνταξη που χρησιμοποιείται αποτελείται από τις παρακάτω ετικέτες, μόνο οι τρεις πρώτες εκ των οποίων είναι απαραίτητες για την καταγραφή μίας διεύθυνσης URL:

Ετικέτα	**Περιγραφή**
<urlset>	> Περιέχει όλες τις πληροφορίες για το σύνολο των διευθύνσεων URL που συμπεριλαμβάνονται στο χάρτη
<url>	Περιέχει όλες τις πληροφορίες για μία διεύθυνση URL
<loc>	Διευκρινίζει τη διεύθυνση URL
. <lastmod>	Αφορά στην ημερομηνία που η URL τροποποιήθηκε για τελευταία φορά, στη μορφή [Χρονολογία-Μήνας-Ημέρα].
<changefreq>	Εκτιμά τη συχνότητα με την οποία προβλέπεται ότι τροποποιείται η διεύθυνση URL. Λαμβάνει τις τιμές always, hourly, daily, weekly, monthly, yearly, never.
. <priority>	Περιγράφει την προτεραιότητα μίας URL, σε σχέση με τις υπόλοιπες του ιστοτόπου. Λαμβάνει τιμές από 0.1 (καθόλου σημαντική) έως 1.0 (εξαιρετικά σημαντική. Η αρχική σελίδα θεωρείται ότι έχει μοναδιαία προτεραιότητα. Όμως, δεν επηρεάζει τη θέση των σελίδων στα αποτελέσματα αναζήτησης.

2.5.2 Χάρτες βίντεο

Οι χάρτες βίντεο επιτρέπουν την εισαγωγή κειμένων σχετικών με το οπτικοακουστικό υλικό που παρουσιάζεται σε μία σελίδα και, κατ' επέκταση, τη χρήση φράσεων ή λέξεων – κλειδιών. Το οπτικοακουστικό υλικό που δηλώνεται μέσω χαρτών προβάλεται στα αποτελέσματα βίντεο των μηχανών αναζήτησης Google, Bing και Ask, οι οποίες διαθέτουν εξειδικευμένο τομέα παρουσίασης αποτελεσμάτων οπτικοακουστικού υλικού, ενώ η Google παρεμβάλλει στην πρώτη σελίδα αποτελεσμάτων για κάποιον όρο ορισμένα σχετικά βίντεο.

2.5.2.1 Σύνταξη

Η σύνταξη που χρησιμοποιείται αποτελείται από τις παρακάτω βασικές ετικέτες, οι οκτώ πρώτες εκ των οποίων είναι απαραίτητες για την ορθή καταχώρηση του βίντεο:

Ετικέτα	**Περιγραφή**
<urlset>	Περιέχει όλες τις πληροφορίες για το σύνολο των διευθύνσεων URL που συμπεριλαμβάνονται στο χάρτη
<url>	Περιέχει πληροφορίες για μία διεύθυνση URL.
<loc>	Διευκρινίζει τη διεύθυνση URL. Ενδέχεται η τοποθεσία να περιλαμβάνει περισσότερα από ένα βίντεο, επομένως και μία ετικέτα <loc> μπορεί να περιλαμβάνει πολλαπλά βίντεο στο χάρτη.
<video:video>	Περιέχει όλες τις πληροφορίες για ένα βίντεο.
<video:thumbnail_loc>	Διευκρινίζει τη διεύθυνση URL της εικόνας που θα χρησιμοποιηθεί ως επισκόπηση
<video:title>	Διευκρινίζει τον τίτλο του βίντεο (έως 100 χαρακτήρες).
<video:description>	Αφορά μία σύντομη περιγραφή (μέχρι 2048 χαρ.) του βίντεο.
<video:content_loc>	Διευκρινίζει την τοποθεσία του οπτικοακουστικού υλικού. Ενδέχεται να αντικατασταθεί από την ετικέτα <video:player_loc>, σε περίπτωση που δεν πρόκειται για βίντεο αλλά για πρόγραμμα αναπαραγωγής υλικού Flash (αντικείμενο swf).
<video:duration>	Αναφέρει τη διάρκεια του υλικού σε δευτερόλεπτα.

<video:expiration_date>	Η ημερομηνία λήξης του βίντεο (π.χ. στην περίπτωση προσφοράς), στη μορφή [Έτος-Μήνας-Ημέρα] ή [Έτος-Μήνας ΗμέραΤ:Ώρες:Λεπτά:Δευτερόλεπτα+ΏραΖώνης]
<video:publication_date>	Η ημερομηνία έκδοσης του βίντεο, στη μορφή [Έτος-Μήνας-Ημέρα] ή [Έτος-Μήνας-ΗμέραΤ:Ώρες:Λεπτά:Δευτερόλεπτα+ ΏραΖώνης]
<video:tag>	Πρόκειται για ετικέτες – λέξεις που σχετίζονται με το περιεχόμενο του βίντεο. Έχουν περίπου την ίδια σχέση με το αντικείμενο που έχει η ετικέτα λέξεων – κλειδιών με μία ιστοσελίδα. Για κάθε μία λέξη – κλειδί, όμως, απαιτείται καινούρια ετικέτα <video:tag>, ενώ επιτρέπονται έως 32 συνολικά ετικέτες για κάθε βίντεο.

2.5.3 Χάρτες εικόνων

Όμοια με τους χάρτες βίντεο, οι χάρτες εικόνων καταχωρούν τις εικόνες ενός ιστοτόπου, οι οποίες, κατόπιν υποβολής του χάρτη, έχουν πολύ μεγαλύτερες πιθανότητες να εμφανισθούν στα αποτελέσματα αναζήτησης, είτε στο πεδίο των εικόνων (Google, Bing, Ask) είτε στην πρώτη σελίδα των αποτελεσμάτων στο Παγκόσμιο Ιστό (Google).

2.5.3.1 Σύνταξη

Η σύνταξη που χρησιμοποιείται αποτελείται από τις παρακάτω βασικές ετικέτες, οι οκτώ πρώτες εκ των οποίων είναι απαραίτητες για την ορθή καταχώρηση του βίντεο:

Ετικέτα	**Περιγραφή**
<urlset>	Περιέχει όλες τις πληροφορίες για το σύνολο των διευθύνσεων URL που συμπεριλαμβάνονται στο χάρτη.
<url>	Περιέχει πληροφορίες για μία διεύθυνση URL
<loc>	Διευκρινίζει τη διεύθυνση URL. Ενδέχεται η τοποθεσία να περιλαμβάνει περισσότερες από μία εικόνες, επομένως και μία ετικέτα <loc> μπορεί να περιλαμβάνει πολλαπλές εικόνες στο χάρτη
<image:image>	Περιέχει όλες τις πληροφορίες για μία εικόνα.

<image:loc>	Διευκρινίζει τη διεύθυνση URL της εικόνας.
<image:caption>	Αποτελεί ένα επεξηγηματικό κείμενο, μία περιγραφή της εικόνας (όπως το alt γνώρισμα της εικόνας στο HTML αρχείο).
<image:geo_location>	Η γεωγραφική τοποθεσία μίας εικόνας ή φωτογραφίας (π.χ. Athens, Greece).
<image:title>	Διευκρινίζει τον τίτλο της εικόνας.
<image:license>	Περιλαμβάνει τη διεύθυνση URL όπου διευκρινίζονται οι άδειες χρήσης και διάδοσης του περιεχομένου.

2.5.4 Χάρτες ιστοτόπων συμβατών με κινητά τηλέφωνα

Τα αποτελέσματα αναζήτησης στον Παγκόσμιο Ιστό πολλές φορές διαφέρουν από αυτά που εμφανίζονται στις αναζητήσεις από κινητό τηλέφωνο, καθώς οι μηχανές αναζήτησης διαθέτουν ξεχωριστούς ανιχνευτές για την ευρετηρίαση των ιστοτόπων που είναι συμβατοί με κινητά τηλέφωνα (mobile crawlers). Για το λόγο αυτό, δίνεται η δυνατότητα στους κατόχους ή διαχειριστές ιστοσελίδων να βοηθήσουν στην ευρετηρίαση του ιστοχώρου τους για προβολή στα κινητά τηλέφωνα. Ισχύουν ακριβώς οι ίδιοι κανόνες σύνταξης με αυτούς των βασικών XML χαρτών ιστοτόπων, καθώς επίσης και η ίδια δομή, με τη διαφορά ότι τοποθετείται η παρακάτω ετικέτα, για κάθε τοποθεσία <loc>:

```
</mobile:mobile/>
```

2.5.5 Πολλαπλοί χάρτες

Οι διαχειριστές μίας σελίδας που έχουν δημιουργήσει πολλαπλούς χάρτες για τον ιστότοπό τους, μπορούν να δημιουργήσουν και να δηλώσουν ένα χάρτη που θα περιλαμβάνει όλους τους υπόλοιπους και θα ανακατευθύνει, κάθε φορά, τον ανιχνευτή, με τη σειρά, σε όλους τους χάρτες. Ο αρχικός χάρτης στον οποίο καταχωρούνται οι διευθύνσεις άλλων χαρτών ιστοτόπων ονομάζεται και ευρετήριο χαρτών. Με τον τρόπο αυτό, οι διαχειριστές χρησιμοποιούν πολλαπλούς χάρτες, στην περίπτωση που ο ιστότοπός τους περιέχει χιλιάδες σελίδες, είτε διαφορετικούς χάρτες για τη δήλωση των σελίδων, των εικόνων, των βίντεο, κ.ο.κ., ενώ βελτιστοποιείται η χρήση και η συχνότητα επίσκεψης των ιστοτόπων από τις μηχανές αναζήτησης,

2.5.5.1 Σύνταξη

Η XML μορφοποίηση ενός ευρετηρίου χαρτών είναι όμοια με τη σύνταξη ενός απλού χάρτη ιστοτόπου και χρησιμοποιεί τις παρακάτω ετικέτες, μόνο οι τρεις εκ των οποίων είναι απαραίτητες για την ορθή σύνταξη:

Ετικέτα	**Περιγραφή**
<sitemapindex>	Περιέχει όλες τις πληροφορίες για το σύνολο των χαρτών ιστοτόπου.
<sitemap>	Περιέχει πληροφορίες για ένα χάρτη ιστοτόπου
<loc>	Διευκρινίζει τη διεύθυνση URL του εκάστοτε χάρτη
<lastmod>	Προαιρετική ετικέτα που αναφέρεται στην ημερομηνία της τελευταίας επεξεργασίας του εκάστοτε χάρτη.

2.5.6 Δήλωση των χαρτών

Μετά τη δημιουργία ενός χάρτη ιστοτόπου ακολουθεί η υποβολή του στις μηχανές αναζήτησης, για να μπορέσουν οι ανιχνευτές αυτών να το προσπελάσουν και να το αξιοποιήσουν.

Μία ενιαία διαδικασία που οφείλει να πραγματοποιείται από το διαχειριστή είναι η δήλωση των χαρτών στο αρχείο robots.txt.

2.6 Στρατηγική domain

2.6.1 Επιλογή ονόματος και τύπου domain

Υπάρχουν πολλά πιθανά ονόματα domain (τομέα) που μία επιχείρηση ή ένας οργανισμός μπορεί να επιλέξει να καταχωρήσει στο Διαδίκτυο, ειδικά εάν σχετίζεται με πολλές ονομασίες ή αντικείμενα. Πέραν του ονόματος, όμως, του ιστοτόπου, οι τύποι του domain που προσφέρονται για καταχώρηση είναι πολλοί και ποικίλλουν, βάσει του τύπου του ιστοτόπου (όπως .com, .biz, .info, .org, .edu.) αλλά και της χώρας προέλευσης (όπως .gr, .it, .co.uk, .fr).

Για παράδειγμα, έστω ένας ελληνικός οργανισμός «TEIPAT». Ο οργανισμός μπορεί να κατοχυρώσει μόνο το domain name «teipat.org», όμως πολλοί ενδιαφερόμενοι και υποψήφιοι επισκέπτες που θα τον αναζητήσουν διαφορετικά στη μπάρα διευθύνσεων URL δε θα τον βρουν εύκολα. Έτσι, συνίσταται η καταχώρηση διαφορετικών ονομάτων αλλά και τύπων domain, όπως για παράδειγμα τα εξής ονόματα:

- teipat.gr
- teipat.org
- teipat.eu
- teipat.info
- tei-pat.gr
- tei-pat.org
- tei-pat.eu
- tei-pat.info

Τα διαφορετικά αυτά ονόματα domain μπορούν να χρησιμοποιηθούν είτε ως ξεχωριστές σελίδες, με την εφαρμογή της ετικέτας κανονικοποίησης, είτε ως μία σελίδα στην οποία ανακατευθύνονται όλα τα ονόματα, με τη χρήση κάποιας μεθόδου ανακατεύθυνσης. Τα δύο αυτά εργαλεία περιγράφονται αναλυτικά παρακάτω.

Η περίπτωση στην οποία επιλεχθεί η κανονικοποίηση όλων των domains και η παράλληλη χρήση τους πλεονεκτεί έναντι της ανακατεύθυνσης ως προς το πλήθος των σελίδων που θα εμφανισθούν στα οργανικά αποτελέσματα των μηχανών αναζήτησης, δίνοντας τη δυνατότητα μεγαλύτερης εκπροσώπησης σε αυτά, πετυχαίνοντας μεγαλύτερο αριθμό επισκεπτών, συνολικά, στην σελίδα του οργανισμού (ή της επιχείρησης) και μετακινώντας τους ανταγωνιστές χαμηλότερα (ή και σε επόμενες σελίδες των αποτελεσμάτων).

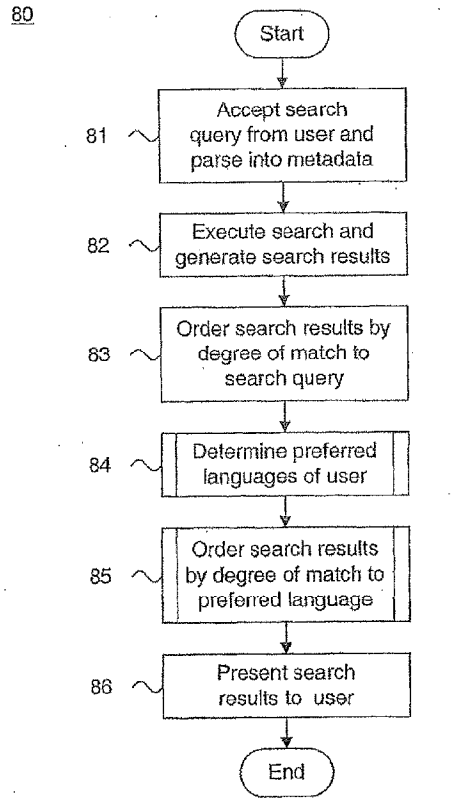
Αντίθετα, όμως, η ύπαρξη πολλών αυτόνομων ονομάτων domain για την προβολή του ίδιου περιεχομένου ενδέχεται να διαιρέσει το βαθμό κατάταξης της ιστοσελίδας. Αυτό προκύπτει από το γεγονός ότι οι σύνδεσμοι από τρίτους ιστοτόπους προς τον οργανισμό θα ποικίλλουν, καθώς θα συνδέουν σε διαφορετικές διευθύνσεις URL, και θα έχει ως αποτέλεσμα τη χαμηλότερη θέση στα αποτελέσματα αναζήτησης. Το πρόβλημα αυτό επιλύεται πλήρως με την χρήση μίας συγκεκριμένης μεθόδου ανακατεύθυνσης σε μία κεντρική διεύθυνση URL, της ανακατεύθυνσης τύπου 301, αντί της ετικέτας κανονικοποίησης.

2.6.2 Γεωγραφική τοποθέτηση

Ως geolocation ορίζεται η διαδικασία κατά την οποία ένας ιστότοπος αντιστοιχίζεται σε μία γεωγραφική περιοχή (χώρα, συνήθως), με σκοπό την ενίσχυση του έργου των μηχανών αναζήτησης να παρέχουν αποτελέσματα που δεν σχετίζονται μόνο με τον όρο αναζήτησης, αλλά και την γεωγραφική τοποθεσία του χρήστη. Το ελληνικό εκπαιδευτικό ίδρυμα teipat, για παράδειγμα, δεν έχει ιδιαίτερο ενδιαφέρον να εμφανίζεται στη Χιλή, ή

το Εκουαδόρ, οπότε η μηχανή αναζήτησης οφείλει να «ωθήσει» τον ελληνικό ιστότοπο στους Έλληνες χρήστες. Αυτό, προφανώς, σημαίνει ότι η γεωγραφική θέση ενός ιστοχώρου ή, έστω, οποιαδήποτε άμεση σύνδεση αυτού με κάποια συγκεκριμένη περιοχή αποτελεί παράμετρο στους αλγορίθμους κατάταξης των ιστοσελίδων, παρότι η μηχανή αναζήτησης της Google τοποθετεί τη διαδικασία γεωγραφικού φιλτραρίσματος των αποτελεσμάτων ως ξεχωριστή διαδικασία που έπεται της κατάταξης των αποτελεσμάτων, όπως φαίνεται στο παρακάτω διάγραμμα ροής (Gupta et al., 2003) [7]

Figure 7.



Εικόνα 10 Διάγραμμα ροής γεωγραφικού φιλτραρίσματος αποτελεσμάτων της Google

Στο ίδιο έγγραφο, η Google εξηγεί ότι οι πιο σημαντικοί παράγοντες που αξιολογούνται για την παροχή σχετικών αποτελεσμάτων για ένα χρήστη, σε δεδομένη γεωγραφική περιοχή, είναι οι εξής:

1. Ο κωδικός της χώρας προέλευσης με τον οποίο έχει καταχωρηθεί το όνομα domain (όπως, για παράδειγμα, «teipat.gr» για Ελλάδα, ή «politics.into.co.uk» για Ηνωμένο Βασίλειο).
2. Η φυσική διεύθυνση της υπηρεσίας κατοχύρωσης ονομάτων domain, μέσω της οποίας πραγματοποιήθηκε η καταχώρηση του ονόματος.
3. Η διεύθυνση IP του διακομιστή που φιλοξενεί τον ιστότοπο.
4. Το ευρύτερο φραστικό πλαίσιο της εκάστοτε σελίδας των συνδέσμων που παραπέμπουν στον εξεταζόμενο ιστότοπο (το anchor text του συνδέσμου, η χώρα προέλευσης της σελίδας που συνδέει προς τον ιστότοπο)

2.6.3 Ανακατεύθυνση

Ως ανακατεύθυνση ορίζεται η διαδικασία προώθησης μίας URL σε μία διαφορετική. Οι περιπτώσεις στις οποίες η διαδικασία αυτή εμφανίζεται χρήσιμη είναι κυρίως οι εξής:

- Η μετακίνηση του ιστοτόπου σε νέο domain.
- Η πρόσβαση των χρηστών στον ιστότοπο μέσω διαφορετικών διευθύνσεων URL, όπως αναφέρθηκε και στην προηγούμενη παράγραφο, με αποτέλεσμα οι ποικίλες διευθύνσεις να διαιρούν την κατάταξη του ιστοτόπου και να διατρέχουν τον κίνδυνο οι μηχανές αναζήτησης να αποδώσουν κάποια ποινή για διπλότυπο περιεχόμενο. Για παράδειγμα, οι διευθύνσεις http://www.teipat.org/index.php, http://www.teipat.org και http://teipat.org.
- Η συγχώνευση δύο ιστοτόπων σε έναν, περίπτωση κατά την οποία ανενεργές, πλέον, διευθύνσεις URL δεν αξιοποιούνται.

Προκύπτει, επομένως, στις παραπάνω περιπτώσεις, η αναγκαιότητα ανακατεύθυνσης μίας διεύθυνσης URL σε μία άλλη, για λόγους βελτιστοποίησης, τόσο σε όρους επισκεψιμότητας χρηστών όσο και σε όρους μηχανών αναζήτησης. Υπάρχουν τρεις τύποι ανακατεύθυνσης:

- **Η 301 ανακατεύθυνση**, η οποία είναι μία μόνιμη ανακατεύθυνση που διαβιβάζει το 90-99% των συνδέσμων στην τελική διεύθυνση URL, σε αυτήν που ανακατευθύνει η πρώτη, δηλαδή. Ο αριθμός 301 αντιστοιχεί στον κωδικό κατάστασης HTTP για το συγκεκριμένο τύπο ανακατεύθυνσης. Πρόκειται για την προτεινόμενη μέθοδο ανακατεύθυνσης, καθώς αποδίδει τη μεγαλύτερη αξία, από πλευράς βελτιστοποίησης μηχανών αναχζήτησης, από κάθε άλλη μορφή ανακατεύθυνσης ιστοτόπων ή ιστοσελίδων. Μεταφράζεται από τις μηχανές αναζήτησης ως «μόνιμη μετακίνηση ιστοτόπου» και για το λόγο αυτό αυτές φροντίζουν να διασφαλίσουν τις επιδόσεις του ιστοτόπου στα αποτελέσματα αναζήτησης και για το νέο τομέα.
- **Η 302 / 307 ανακατεύθυνση**. Η 302 ανακατεύθυνση αφορά τις προσωρινές μετακινήσεις ιστοσελίδων και δεν μεταφέρει την αξία των συνδέσμων και την κατάταξη της σελίδας στο νέο τομέα (ή τη νέα σελίδα), για να μην προκληθεί σύγχυση κατά την επαναφορά του ιστοτόπου στον αρχικό τομέα. Η 307 ανακατεύθυνση αφορά την ίδια ακριβώς περίπτωση. Η βασική διαφορά των δύο είναι ότι η ανακατεύθυνση τύπου 302 χρησιμοποιείται για ιστοτόπους που

φιλοξενούνται σε διακομιστές συμβατούς με την έκδοση HTTP 1.0, ενώ η ανακατεύθυνση κωδικού κατάστασης 307 αφορά εξυπηρετητές συμβατούς με την έκδοση HTTP 1.1

- **Η Meta ανανέωση (Meta refresh).** Πρόκειται για μία ανακατεύθυνση που δεν φέρει κάποιον κωδικό κατάστασης HTTP, καθώς δεν πραγματοποιείται σε επίπεδο διακομιστή, αλλά σε επίπεδο σελίδας. Παρότι, σε αντίθεση με την 302 ανακατεύθυνση, διαβιβάζει μέρος της αξίας των συνδέσμων προς τον ιστότοπο στο νέο τομέα, αποτελεί την πλέον αργή ανακατεύθυνση.

Επομένως, η χρήση τα Meta Ανανέωσης συνίσταται μόνο σε περίπτωση που ο διαχειριστής εκούσια επιθυμεί τη μη μεταβίβαση οποιασδήποτε αξίας από την παλαιότερη στη νέα σελίδα. Παράλληλα, η 302 ανακατεύθυνση χρησιμοποιείται μόνο στην περίπτωση που η μετατόπιση του ιστοτόπου είναι προσωρινή και όχι μόνιμη, όταν, για παράδειγμα, υπάρχει κάποιο τεχνικό πρόβλημα στην αρχική σελίδα που αναμένεται να διορθωθεί ή το περιεχόμενο μεταφέρεται προσωρινά σε μία νέα ιστοσελίδα προσφορών και προώθησης προϊόντος ή υπηρεσίας. Επομένως, η πλέον προτεινόμενη πρακτική για τη βελτιστοποίηση της ιστοσελίδας είναι η χρήση της ανακατεύθυνσης 301, η οποία υποδεικνύει, τόσο στους φυλλομετρητές όσο και τους ανιχνευτές των μηχανών αναζήτησης ότι η σελίδα έχει μεταφερθεί μόνιμα. Οι μηχανές αναζήτησης, τότε, αντιλαμβάνονται ότι η σελίδα έχει αλλάξει τοποθεσία αλλά και πως το περιεχόμενο, το ίδιο ή αναβαθμισμένο, μπορεί να βρεθεί στη νέα διεύθυνση URL, και μεταφέρουν τη βαρύτητα των εισερχόμενων συνδέσμων προς τη συγκεκριμένη σελίδα, από την προηγούμενη διεύθυνση στην καινούρια. Φυσικά, απαιτείται χρόνος προκειμένου οι μηχανές αναζήτησης ανακαλύψουν την 301 ανακατεύθυνση, την αναγνωρίσουν και «πιστώσουν» στη νέα διεύθυνση URL όλες τις κατατάξεις για τους αντίστοιχους όρους αναζήτησης.

2.7 Βελτιστοποίηση Flash περιεχομένου

Οι μηχανές αναζήτησης έχουν μικρή ή καθόλου δυνατότητα ανάγνωσης και ευρετηρίασης περιεχομένου Flash, σε σύγκριση με τις δυνατότητές τους στην ανάγνωση HTML σελίδων. Αυτό συμβαίνει διότι τα αρχεία Flash δεν παρουσιάζουν περιεχόμενο άμεσα στο χρήστη, αλλά λειτουργούν ως ένα ξεχωριστό πρόγραμμα ή script. Άλλωστε, η ίδια η φύση ενός Flash στοιχείου (οπτικοακουστικό υλικό, παιχνίδι, διαδραστική εφαρμογή) είναι αντίθετη στη χρήση κειμένου και λέξεων – κλειδιών που, σχεδόν αποκλειστικά, καθοδηγούν τις μηχανές αναζήτησης. Σήμερα, η Google έχει καταφέρει να αναγνώσει Flash ιστοσελίδες, να εντοπίζει και να ευρετηριάζει ορισμένα αντικείμενα κειμένου, καθώς και να ακολουθεί ορισμένους συνδέσμους URL που παρατίθενται εντός του στοιχείου Flash, με την προϋπόθεση η κωδικοποίηση αυτού να γίνεται με τα πρότυπα που η μηχανή αναζήτησης έχει θέσει.

Υπάρχουν αρκετές τεχνικές που χρησιμοποιούνται για τη βελτιστοποίηση, σε όσο το δυνατόν υψηλότερο επίπεδο, μίας σελίδας Flash, όμως η πιο αποδοτική που επιλύει εξ ολοκλήρου το πρόβλημα της βελτιστοποίησης, δίνοντας τις απεριόριστες δυνατότητες

που δίνει και μία σελίδα HTML, είναι η χρήση του «swfobject» [8]. Το «swfobject» είναι πρακτικά ένα κομμάτι κώδικα JavaScript που προηγείται της φόρτωσης του Flash περιεχομένου. Παρέχει στους χρήστες ορισμένα πολύ σημαντικά πλεονεκτήματα, όπως ο εντοπισμός τεχνικής υποστήριξης για την Flash, ο έλεγχος συμβατότητας έκδοσης της τεχνολογίας μεταξύ ιστοσελίδας και φυλλομετρητή, ο έλεγχος ενημερώσεων για την έκδοση Flash του φυλλομετρητή και την πολύτιμη υποστήριξη εμφάνισης εναλλακτικού περιεχομένου στους χρήστες που δεν έχουν οποιαδήποτε ή την απαιτούμενη έκδοση Flash εγκατεστημένη στο φυλλομετρητή. Από άποψη τεχνικής βελτιστοποίησης, το «swfobject» παρέχει, με τον ίδιο ακριβώς τρόπο, τη δυνατότητα εναλλακτικής παρουσίασης του περιεχομένου του Flash στοιχείου (ή σελίδας) όχι μόνο στους επισκέπτες που το χρειάζονται αλλά και στις ίδιες τις μηχανές αναζήτησης, σε γλώσσα HTML. Παράλληλα, εξασφαλίζει την ευρετηρίαση του Flash περιεχομένου από τις μηχανές αναζήτησης.

Όσον αφορά το πρώτο πλεονέκτημα, με τη χρήση του «swfobject», δίνεται η δυνατότητα στο διαχειριστή της ιστοσελίδας να παρέχει HTML κείμενο «πίσω» από το στοιχείο Flash. Έτσι, παρέχεται περιεχόμενο κειμένου, πλούσιο σε λέξεις και φράσεις – κλειδιά που μπορεί, με τον παραδοσιακό τρόπο, να ευρετηριασθεί από τις μηχανές αναζήτησης. Ο μοναδικός και λογικός περιορισμός που θέτουν οι μηχανές αναζήτησης, ως προς τη χρήση αυτής της μεθόδου, είναι το εναλλακτικό περιεχόμενο σε HTML να είναι πανομοιότυπο με το πρωταρχικό σε Flash. Η απαίτηση αυτή γίνεται για δύο λόγους: Αφενός για να μην επιχειρούνται παράνομες τεχνικές (black – hat SEO), με την καταχρηστική άσκοπη επανάληψη λέξεων, φράσεων και περιεχομένου ή τον εμπλουτισμό του εγγράφου με ασύνδετες ή και άσχετες με το περιεχόμενο λέξεις, αφετέρου για να διασφαλιστεί η ορθή παρουσίαση του ακριβούς περιεχομένου σε όλους τους χρήστες εξίσου, ανεξάρτητα από το εάν αυτό θα προβληθεί σε Flash ή HTML.

Το δεύτερο πλεονέκτημα ξεπερνά ένα ακόμη γνωστό πρόβλημα της τεχνολογίας Flash, κατά το οποίο ο διαχειριστής δεν μπορεί να γνωρίζει εάν η μηχανή αναζήτησης που δύναται να αναγνωρίζει Flash περιεχόμενο θα καταφέρει να εντοπίσει τα Flash αρχεία. Εάν αυτά είναι «κρυμμένα» από τις μηχανές, πίσω από JavaScript φορτωτές, δεν θα ευρεθούν, εμποδίζοντας την ήδη περιορισμένη δυνατότητα ανίχνευσης κι ευρετηρίασης που υπάρχει σήμερα. Επειδή, όμως, το «swfobject» αποτελεί πρότυπο της βιομηχανίας του Διαδικτύου, οι μηχανές αναζήτησης μπορούν να μεταφράσουν το swfobject και να εντοπίσουν τα αρχεία αυτά.

Πλεονεκτήματα

- Όπως αναφέρθηκε ήδη, η μέθοδος «swfobject» παρέχει έναν αποδοτικό τρόπο να καθιστά ορατό ένα εναλλακτικό περιεχόμενο μίας Flash ιστοσελίδας στις μηχανές αναζήτησης, καθιστώντας σίγουρη την ανίχνευση και ευρετηρίαση ακόμη και των Flash στοιχείων.
- Αποτελεί πρότυπο της βιομηχανίας, με αποτέλεσμα να υποστηρίζεται από όλους τους φυλλομετρητές και τις μηχανές αναζήτησης.
- Το εναλλακτικό περιεχόμενο μορφοποιείται σε γλώσσα HTML, παρέχοντας έτσι κάθε δυνατότητα στο διαχειριστή να το βελτιστοποιήσει με τις τεχνικές που

έχουν ήδη αναπτυχθεί προηγουμένως, ενώ παρέχει τη δυνατότητα σε όλους τους χρήστες να προβάλουν τη σελίδα, ανεξάρτητα από την κατάσταση του φυλλομετρητή και της έκδοσης Flash αυτού.

Μειονεκτήματα

- Είναι προφανές ότι η βελτιστοποίηση μίας σελίδας Flash χρειάζεται πολύ περισσότερη δουλειά απ' ό,τι θα χρειαζόταν η αντίστοιχη διαδικασία σε μία HTML σελίδα. Αυτό συμβαίνει διότι απαιτείται η επανακατασκευή της σελίδας σε HTML και στη συνέχεια η βελτιστοποίηση του HTML περιεχομένου.
- Η αντιμετώπιση του εναλλακτικού (κρυφού) περιεχομένου επαφίεται, κάθε φορά, στην πολιτική της εκάστοτε μηχανής αναζήτησης. Ενδέχεται, για παράδειγμα, να δίνουν μικρότερη σημασία στο περιεχόμενο μίας σελίδας που επιστρατεύει τη δέσμη «swfobject», κατατάσσοντάς το δευτερεύον, ειδικά εάν η ιστοσελίδα παρουσιάσει κατάχρηση της μεθόδου ή και άσχετο περιεχόμενο.

2.8 Θέματα χρόνου και συχνότητας

2.8.1 Το φαινόμενο «sandbox»

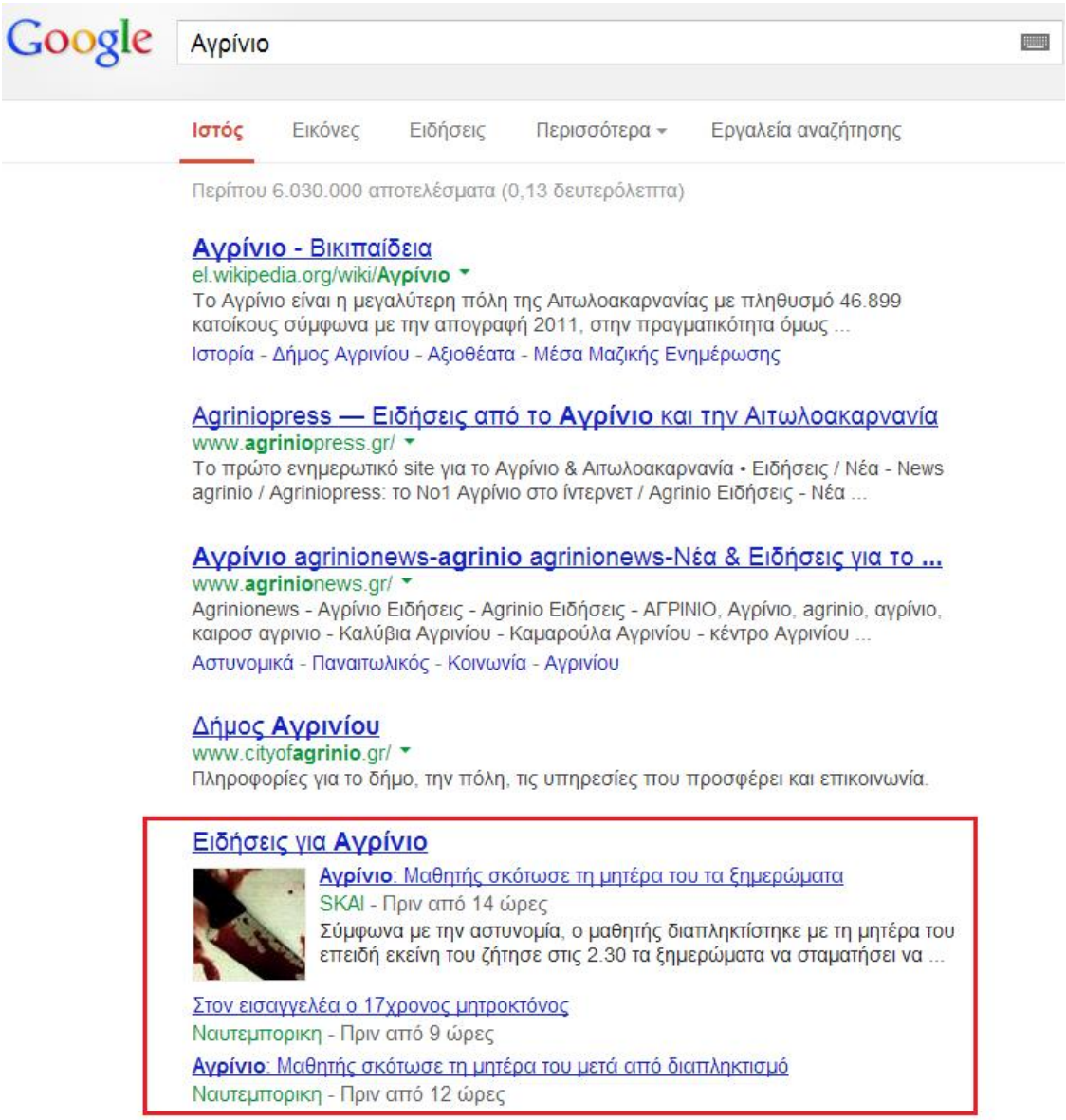
Το φαινόμενο «sandbox» αφορά αποκλειστικά στη μηχανή αναζήτησης της Google και είναι γνωστό ως το «Google Sandbox Effect». Πρόκειται για ένα φαινόμενο που επιβραδύνει σημαντικά την αντικειμενική εκπροσώπηση ενός νεοκαταχωρηθέντος ιστοχώρου στα αποτελέσματα αναζήτησης της μηχανής της Google. Η ίδια η εταιρεία αρνείται, μέχρι και σήμερα, την ύπαρξη ενός τέτοιου φαινομένου ενώ διαφωνεί και με τη χρήση του όρου αυτού. Σύμφωνα με την έρευνα που πραγματοποίησε, στις αρχές του 2009, ο Rand Fishkin της SEOmoz [9], σε συνεργασία με τον ιστότοπο Grader.com, παρατήρησε ότι για όρους αναζήτησης που ταυτίζονταν 100% με τίτλους σελίδων του ιστοτόπου και δεδομένης της εφαρμογής τεχνικών βελτιστοποίησης της ιστοσελίδας από τους μηχανικούς της, οι αντίστοιχες σελίδες εμφανίζονταν στις θέσεις #50 έως και #300 των αποτελεσμάτων αναζήτησης, την ίδια στιγμή που, για τους ίδιους όρους, οι σελίδες αυτές καταλάμβαναν την πρώτη θέση των αποτελεσμάτων των μηχανών Yahoo και MSN/Live. Από το τελευταίο, γίνεται εμφανές ότι το φαινόμενο αυτό αφορά μόνο την Google και όχι τις υπόλοιπες μεγάλες μηχανές αναζήτησης. Μάλιστα, έχουν παρατηρηθεί ορισμένα κοινά γνωρίσματα των ιστοσελίδων που έρχονται αντιμέτωπες με το φαινόμενο αλλά και κοινά συμπτώματα αυτού του φαινομένου:

- Αφορά αποκλειστικά νέους τομείς (ηλικίας, συνήθως, μικρότερης του ενός έτους).

- Οι σελίδες αδυνατούν να καταταχθούν στα πρώτα αποτελέσματα της Google, ακόμη και για μοναδικούς όρους αναζήτησης ή φράσεις που ταυτίζονται με τίτλους σελίδων του ιστοτόπου.
- Οι τεχνικές βελτιστοποίησης που εφαρμόζονται αποδίδουν εξίσου με οποιαδήποτε άλλη ιστοσελίδα, όσον αφορά τις θέσεις στα αποτελέσματα που επιτυγχάνονται στις υπόλοιπες μηχανές αναζήτησης.
- Σημειώνεται μία προσωρινή περίοδος στην οποία οι ιστότοποι αυτοί καταλαμβάνουν ιδιαίτερα ανταγωνιστικές θέσεις, ίσως περισσότερο ανταγωνιστικές από αυτές που θα καταλάμβαναν διαφορετικά, και στη συνέχεια πέφτουν απότομα 30 – 500 θέσεις στα αποτελέσματα μέχρις ότου παύσει η ισχύς του φαινομένου.

2.8.2 Συχνότητα ανανέωσης περιεχομένου

Έχει παρατηρηθεί ότι όσο πιο συχνά ανανεώνεται το περιεχόμενο ή όσο πιο φρέσκο είναι, τόσο πιο μεγάλη βαρύτητα έχει η σελίδα που το φιλοξενεί στους αλγορίθμους κατάταξης των ιστοσελίδων. Ειδικότερα, η Google περιλαμβάνει έναν εξειδικευμένο ανιχνευτή, το FreshBot, ο οποίος αναζητά διαρκώς νέο περιεχόμενο. Η βαρύτητα που αποκτούν οι νέες καταχωρήσεις στο Διαδίκτυο τις καθιστά πιο ανταγωνιστικές στα αποτελέσματα αναζήτησης για ένα διάστημα, μετά το οποίο παύει να επιδρά στους αλγορίθμους κατάταξης και οι σελίδες αυτές χάνουν θέσεις μέχρι μία πιο «μόνιμη» κατάταξη. Ως απόρροια της παραπάνω διαδικασίας που ακολουθείται από τις μηχανές αναζήτησης, παρατηρείται ότι οι ειδήσεις και οι καταχωρήσεις σε blogs ή forums, όπως φαίνεται και στο παρακάτω παράδειγμα, εμφανίζονται προσωρινά αρκετά υψηλά στα αποτελέσματα:



Εικόνα 11 Παράδειγμα ευνοϊκής κατάταξης φρέσκων σελίδων

Επομένως, είναι απαραίτητο να εμπλουτίζεται διαρκώς ένας ιστότοπος με νέες καταχωρήσεις, προϊόντα, ανακοινώσεις, νέα δεδομένα γενικότερα. Ακόμη και σε περιπτώσεις που αυτό είναι ανέφικτο, είναι καλό να ανανεώνεται ή να επεξεργάζεται ελαφρώς το ήδη υπάρχον περιεχόμενο, καθώς η λειτουργία των ανιχνευτών περιλαμβάνει τη σύγκριση της ιστοσελίδας με την προσφάτως καταχωρημένη εικόνα αυτής. Μία άλλη τακτική είναι η προσθήκη πρόσθετων δυναμικής ενημέρωσης και μηχανισμών αυτόματης παραγωγής δυναμικού περιεχομένου, όπως είναι τα πρόσθετα μετεωρολογικής ενημέρωσης, πρόσθετα εμφάνισης πρόσφατων καταχωρήσεων στους λογαριασμούς twitter συγκεκριμένων προσώπων κ.α., για να μη μένει στάσιμη η εικόνα της ιστοσελίδας.

2.8.3 Μακροβιότητα ιστοτόπου

Η μακροβιότητα του ιστοτόπου, του ονόματος τομέα και του διακομιστή αποτελεί εξίσου σημαντική παράμετρο των αλγορίθμων κατάταξης των αποτελεσμάτων αναζήτησης. Φαίνεται πως αποδίδεται περισσότερη «εμπιστοσύνη» σε καθιερωμένους ιστοτόπους και ονόματα τομέα, κάτι που σχετίζεται εμμέσως και με το φαινόμενο «sandbox» της μηχανής αναζήτησης της Google. Ένας τρόπος αντιμετώπισης του προβλήματος ανεπαρκούς εμπιστοσύνης που αντιμετωπίζουν οι σχετικά νέοι ιστότοποι είναι η μακροχρόνια διατήρηση του ίδιου ονόματος τομέα, η χρήση μόνιμων ανακατευθύνσεων (κωδικού κατάστασης 301) από ανενεργές σελίδες με μεγάλο ιστορικό ποιοτικών εισερχόμενων συνδέσμων, προς νέες ενεργές σελίδες ή τομείς, καθώς και η αγορά καθιερωμένων ονομάτων domain για την κατασκευή ενός νέου ιστοχώρου που πλεονεκτούν έναντι των αχρησιμοποίητων.

2.8.4 Συχνότητα δημιουργίας εσωτερικών και εισερχόμενων συνδέσμων

Οι μηχανές αναζήτησης ελέγχουν το ρυθμό με τον οποίο οι ιστότοποι επεκτείνονται εσωτερικά αλλά και με τον οποίο αποκτούν συνδέσμους από ξένες σελίδες προς αυτούς. Ακριβώς όπως το νέο περιεχόμενο ευνοείται, το ίδιο ευνοούνται και οι ιστοχώροι που αναπτύσσονται, σε όρους διευθύνσεων URL, με μεγάλη ταχύτητα. Παράλληλα, ο ρυθμός αύξησης των εισερχόμενων συνδέσμων επιδρά στον αλγόριθμο κατάταξης των αποτελεσμάτων αναζήτησης κάθε μηχανής. Παράλληλα, φυσικά, με το ρυθμό, μεγάλη σημασία έχει η ποιότητα των συνδέσμων αυτών, ο έλεγχος, όμως, αυτής της ποιότητας ενεργοποιείται από τις μηχανές με κριτήριο το ρυθμό. Έτσι, ένας ενδεχομένως αρκετά υψηλός ρυθμός δημιουργίας εισερχόμενων συνδέσμων (backlinks) αποτελεί υπόδειξη χειροκίνητων τεχνικών δημιουργίας συνδέσμων ή κατάχρησης κάποιας φάρμας συνδέσμων (link farm).

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] http://lvk.cs.msu.su/~bruzz/articles/web_retrieval/94.pdf

[2] http://oak.cs.ucla.edu/~cho/papers/cho-evol.pdf

[3] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.9242&rep=rep1&type=pdf

[4] http://oak.cs.ucla.edu/~cho/papers/cho-parallel.pdf

[5] http://informatics.indiana.edu/fil/Papers/TOIT.pdf

[6] http://googlewebmastercentral.blogspot.gr/2009/09/google-does-not-use-keywords-meta-tag.html

[7] http://www.google.com/patents/US20130060768?dq=System+and+method+for+providing+preferred+country+biasing+of+search+results&hl=el&sa=X&ei=hWifUf3QF4WBhAfD8ICABw&ved=0CE4Q6wEwAw

[8] https://code.google.com/p/swfobject/

[9] http://www.seomoz.org/blog/googles-sandbox-still-exists-exemplified-by-gradercom,

§ "SEO - Search Engine Optimization Bible" - Jerri L. Ledford

§ "SEO Made Easy - Everything You Need to Know About SEO and Nothing More" - Evan Bailyn

§ "Search Engine Optimization (SEO) Secrets" – Danny Dover