ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Βελτιστοποίηση Ιστοσελίδων για Μηχανές Αναζήτησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΕΞΑΝΔΡΟΥ ΝΙΚΑ

Επιβλέπων : Γρηγόριος Μέντζας
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Βελτιστοποίηση Ιστοσελίδων για Μηχανές Αναζήτησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΕΞΑΝΔΡΟΥ ΝΙΚΑ

Επιβλέπων : Γρηγόριος Μέντζας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Νοεμβρίου 2011.

.....	.....	.....
Γρηγόριος Μέντζας Καθηγητής Ε.Μ.Π.	Δημήτριος Ασκούνης Καθηγητής Ε.Μ.Π.	Ιωάννης Ψαρράς Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2011

.....

ΑΛΕΞΑΝΔΡΟΣ Κ. ΝΙΚΑΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Πρόλογος

Θα ήθελα θερμότατα να ευχαριστήσω τον Καθηγητή Γρηγόριο Μέντζα για την πολύτιμη συμβολή του, την επιστημονική αλλά και ηθική υποστήριξη που μου παρείχε μέσω της στενής επίβλεψής του, σε όλα τα στάδια της εκπόνησης της παρούσας διπλωματικής εργασίας, και της συνεργασίας που είχαμε το τελευταίο διάστημα αλλά και στα πλαίσια του μαθήματος «Διοίκηση της Ψηφιακής Επιχείρησης». Κυρίως, όμως, για την ευκαιρία που μου έδωσε να ασχοληθώ ερευνητικά με το συγκεκριμένο θέμα της βελτιστοποίησης της κατάταξης των ιστοσελίδων. Ως άτομο με καινοτόμο και διερευνητικό πνεύμα, πολύπλευρες και βαθύτατες γνώσεις σε ένα ευρύ επιστημονικό φάσμα, αλλά και ανθρώπινο χαρακτήρα, αποτελεί παράδειγμα προς μίμηση και για εμένα προσωπικά πηγή έμπνευσης.

Θα ήθελα, επίσης, να ευχαριστήσω εκ βαθέων τους Καθηγητές Ι. Ψαρρά και Δ. Ασκούνη, καθώς θεωρώ ιδιαίτερη τιμή μου τη συμμετοχή τους στην επιτροπή εξέτασης της διπλωματικής εργασίας.

Παράλληλα, για την καθοριστικής σημασίας συμβολή τους, τις εποικοδομητικές τους προτάσεις, τις εύστοχες συμβουλές και την πιστή τους συμπαράσταση, ιδιαίτερες ευχαριστίες οφείλουν να αποδοθούν στους Κώστα Χρηστίδη και Μπάμπη Μαγκούτα, ερευνητές της μονάδας Διοίκησης Πληροφοριακών Συστημάτων, καθώς και σε όλα τα υπόλοιπα μέλη της μονάδας για τη στήριξη κι εμψύχωση που αφειδώς μου προσέφεραν, καθ' όλη τη διάρκεια της εκπόνησης της εργασίας.

Τέλος, θέλω ειλικρινά να ευχαριστήσω την αγαπημένη μου οικογένεια και το στενό φιλικό μου περιβάλλον που μου παρείχαν ηθικά και ψυχικά εφόδια και με βοήθησαν να ανταπεξέλθω και να ανταποκριθώ στις διάφορες προκλήσεις και δυσκολίες.

Περίληψη

Η ολοένα αυξανόμενη συμμετοχή των επιχειρήσεων και των οργανισμών στο Διαδίκτυο την τελευταία δεκαετία 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, καθώς κι επιβάλλει το δυναμικό χαρακτήρα μεταβολής της συμπεριφοράς των μηχανών αναζήτησης. Η παρούσα διπλωματική εργασία ασχολείται με θέματα που αφορούν τη λειτουργία των μηχανών 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. Επίσης, γίνεται αναλυτική περιγραφή των προσπαθειών να προσεγγισθούν πειραματικά οι παράμετροι που επιδρούν στους αλγορίθμους των μηχανών αναζήτησης για την κατάταξη των ιστοσελίδων. 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. Ακολουθεί η μαθηματική και θεωρητική ανάλυση της διασύνδεσης των ιστοσελίδων στο διαδικτυακό γράφο και του βαθμού PageRank και μελετώνται οι τρόποι με τους οποίους αυτός επηρεάζει την κατάταξη των αποτελεσμάτων αναζήτησης, ενώ, στη συνέχεια και με βάση την ανάλυση αυτή, διατυπώνονται ορισμένα αξιώματα που τον χαρακτηρίζουν και αναπτύσσονται οι βασικότερες τεχνικές κατασκευής συνδέσμων που προκύπτουν από αυτά. Τέλος, διαπιστώνεται η σχέση της διαδικασίας βελτιστοποίησης με τον Σημασιολογικό Ιστό κι επιχειρείται μία θεωρητική αναπροσαρμογή της στα δεδομένα του Web 3.0.

Λέξεις Κλειδιά: τεχνικές βελτιστοποίησης, ιστοσελίδες, κατάταξη, οργανικά αποτελέσματα, μηχανές αναζήτησης, προώθηση

Abstract

The increasing activity of businesses and organizations in the Web over the past decade has brought a major rise of competition in the search engine result pages, for search terms related to them, as well as enforced the ongoing and dynamic changes of attitude of the major search engines. This diploma thesis deals with the various search engine functions, in order to approach webpage ranking optimization techniques. At first, the search engine technologies and functions related to web crawling, indexing and query proccessing are studied. An effort to experimentally approach the parameters that affect the search engine ranking algorithms is also made. Later on, all those factors concerning the internal and on-page structure and formatting of a website, as well as the server hosting it, are studied in detail and, based on the conclusions, the corresponding optimization methods are approached. Subsequently, a mathematical and theoretical approach to webpage interlinking inside the web graph as well as PageRank and the ways it affects page ranking is made and, as a result, certain PageRank axioms are put forward to help develop link building strategies. Finally, the connection between search engine optimization and Semantic Web is concluded and the theoretical readjustment of the former to the aspects of Web 3.0 is discussed.

Keywords: Search Engine Optimization, SEO, ranking, webpages, websites, organic results, internet marketing

Πίνακας περιεχομένων

Ευρετήριο Πινάκων

Ευρετήριο Εικόνων

1 *Εισαγωγή*

1.1 Εισαγωγή στη βελτιστοποίηση των ιστοσελίδων

Η συντριπτική πλειοψηφία των επιχειρήσεων και οργανισμών, σήμερα, δεν περιορίζονται στη φυσική τους παρουσία και δραστηριοποίηση, αλλά επεκτείνονται και στο Διαδίκτυο, προωθώντας τα φυσικά τους καταστήματα, ή πωλώντας απευθείας τα προϊόντα και τις υπηρεσίες που παράγουν. Ορισμένες φορές, μάλιστα, δραστηριοποιούνται αποκλειστικά ηλεκτρονικά.

Έτσι, μία τυπική τοπική επιχείρηση αναμένεται να επενδύσει στο Διαδίκτυο και να κατασκευάσει έναν λειτουργικό, υψηλής αισθητικής ιστότοπο με την state of the art τεχνολογία, τον οποίο θα θέσει σε λειτουργία με πολλές προσδοκίες και προοπτικές. Παράλληλα, όμως, η έρευνα αγοράς αλλά και η ευρύτερη αναζήτηση της πληροφορίας γίνεται αποκλειστικά μέσα από τις μηχανές αναζήτησης, στις σελίδες των αποτελεσμάτων αναζήτησης των οποίων ο μέσος χρήστης θα δοκιμάσει να εμπιστευθεί έναν ορισμένο αριθμό πρώτων ιστοσελίδων που επιστρέφονται μέχρις ότου να βρει αυτό που ζητάει ή να εξαντληθεί η υπομονή του. Αυτό έχει ως αποτέλεσμα η μέση αυτή επιχείρηση να αποτυγχάνει εν τέλει στους σκοπούς της και ο πολλά υποσχόμενος ιστοχώρος να μην ανταποκρίνεται στις προσδοκίες των ιδιοκτητών – διαχειριστών.

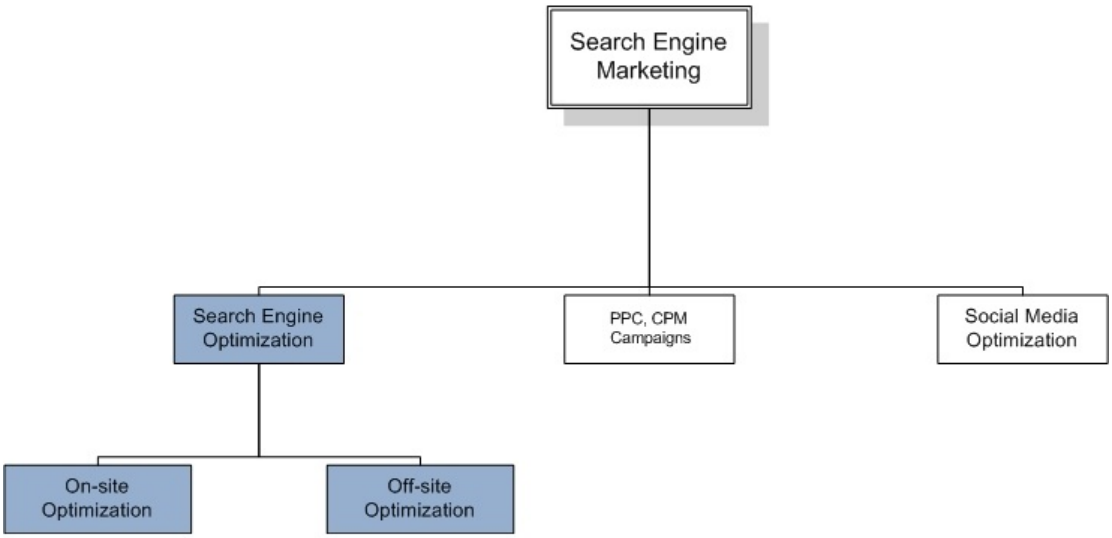
Με τον τρόπο αυτό, δημιουργείται το παράδοξο φαινόμενο, σύμφωνα με το οποίο ολοένα και περισσότερες επιχειρήσεις και οργανισμοί δραστηριοποιούνται στο Διαδίκτυο και ανταγωνίζονται για μία θέση ανάμεσα στις εξαιρετικά και διαχρονικά περιορισμένες θέσεις των αποτελεσμάτων των διαφόρων σχετικών με τις επιχειρήσεις αυτές αναζητήσεων. Οι επιχειρήσεις και οι οργανισμοί, λοιπόν, παύουν να περιορίζονται αποκλειστικά και μόνο στην κατασκευή ενός ιστοτόπου και δημιουργούνται οι ανάγκες αλλά και οι προϋποθέσεις για την προώθηση αυτού στις όσο το δυνατόν υψηλότερες θέσεις των αποτελεσμάτων αναζήτησης.

Οι θέσεις των αποτελεσμάτων των μηχανών αναζήτησης συχνά καθορίζουν την ηλεκτρονική ή και γενικότερη βιωσιμότητα μίας επιχείρησης ή ενός οργανισμού, καθώς αυτή εξαρτάται από την ορατότητα (visibility) των ιστοχώρων αυτών από τους διαδικτυακούς χρήστες.

Τη λύση στο πρόβλημα αυτό έρχεται να δώσει μία νέα, σχετικά, και διαρκώς αναπτυσσόμενη στο χώρο και τη βιομηχανία του Διαδικτύου δραστηριότητα, αυτή της βελτιστοποίησης της κατάταξης των ιστοσελίδων στα οργανικά αποτελέσματα των μηχανών αναζήτησης (**Search Engine Optimization – SEO**), ή αλλιώς βελτιστοποίησης για τις μηχανές αναζήτησης. Με τον όρο **οργανικά αποτελέσματα** αναφερόμαστε στα αποτελέσματα της αναζήτησης που προκύπτουν αποκλειστικά από τη σχέση που έχουν με τον όρο αναζήτησης, σε αντίθεση με εκείνα που εμφανίζονται ως αποτέλεσμα κάποιας επί πληρωμή διαφημιστικής καταχώρησης, τα οποία ονομάζονται μη οργανικά Τα τελευταία συνήθως εμφανίζονται πάνω (ή κάτω) και δεξιά από τα οργανικά αποτελέσματα του δικτύου αναζήτησης (Search Network) των μηχανών αναζήτησης, με ελαφρώς διαφορετικό χρώμα, αλλά και στο δίκτυο περιεχομένου (Content Network) των μηχανών που περιλαμβάνει ένα πολύ μεγάλο μέρος του Διαδικτύου και συναποτελείται από όλες τις ιστοσελίδες που παραθέτουν διαφημιστικό χώρο στην εκάστοτε μηχανή.

Η βελτιστοποίηση για τις μηχανές αναζήτησης αποτελεί μία μόνο από τις τρεις πτυχές ενός ευρύτερου πλαισίου, αυτού του **Search Engine Marketing ή SEM** (μάρκετινγκ για τις μηχανές αναζήτησης). Πρόκειται για μία μορφή Διαδικτυακού Μάρκετινγκ που σκοπό έχει την προώθηση των ιστοτόπων για τη βελτίωση της ορατότητάς τους (visibility) και την αύξηση της επισκεψιμότητας και της οικονομικής αποδοτικότητας αυτών. Το Search Engine Marketing, πέραν της βελτιστοποίησης στις μηχανές αναζήτησης (SEO), περιλαμβάνει τις επί πληρωμή διαφημιστικές καταχωρήσεις (**PPC** – Pay Per Click, για την κοστολόγηση κάθε κλικ των χρηστών σε αυτές, και **CPM** – Cost per Mile, για την κοστολόγηση ανά χίλιες προβολές της διαφήμισης), καθώς και τη βελτιστοποίηση στα μέσα κοινωνικής δικτύωσης (**Social Media Optimization - SMO**), που σκοπό έχει την προώθηση των ιστοτόπων στα Social Media & Networks, όπως τα Facebook, Twitter και YouTube.

Έτσι, όσον αφορά τα οργανικά αποτελέσματα των μηχανών αναζήτησης, ιδιαίτερο ενδιαφέρον αποκτά η διερεύνηση των παραγόντων που οι αλγόριθμοι των μηχανών λαμβάνουν υπόψη για την κατάταξη των αποτελεσμάτων αναζήτησης και η περιγραφή των τεχνικών βελτιστοποίησης της κατάταξης των ιστοσελίδων σε αυτά. Η βελτιστοποίηση των στοιχείων και παραμέτρων της σελίδας που πραγματοποιείται σε επίπεδο ιστοσελίδας (κώδικας, περιεχόμενο, δομή) και διακομιστή ονομάζεται βελτιστοποίηση εντός της ιστοσελίδας (**on-page ή on-site optimization**), ενώ η βελτιστοποίηση που αφορά σε εξωτερικούς παράγοντες (σύνδεσμοι, PageRank) και δεν πραγματοποιείται εσωτερικά ονομάζεται βελτιστοποίηση εκτός της ιστοσελίδας (**off-page ή off-site optimization**).



Εικόνα 1 Η σχέση της βελτιστοποίησης με το Search Engine Marketing

1.2 Αντικείμενο διπλωματικής

Σκοπός της διπλωματικής εργασίας είναι η μελέτη των λειτουργιών των μηχανών αναζήτησης, ο προσδιορισμός και η ανάλυση όλων των εσωτερικών κι εξωτερικών, ως προς την ιστοσελίδα, παραγόντων που επιδρούν στην κατάταξη των αποτελεσμάτων αναζήτησης, η ανάπτυξη τεχνικών βελτιστοποίησης του κώδικα, της δομής και του περιεχομένου μίας ιστοσελίδας καθώς και του τρόπου διασύνδεσης αυτής στο Διαδίκτυο, στο βαθμό που συνδέονται με τους παράγοντες αυτούς.

Έτσι, στο 2^{ο} κεφάλαιο αναλύονται οι διαφορετικές λειτουργίες που διεκπεραιώνει μία τυπική μηχανή αναζήτησης και, για κάθε μία από αυτές, ερευνάται η εξέλιξη των τεχνολογιών που χρησιμοποιήθηκαν. Παρουσιάζονται, λοιπόν, οι διάφοροι αλγόριθμοι με τους οποίους ανιχνεύονται τα διαδικτυακά έγγραφα από τα επιμέρους προγράμματα πλοήγησης των μηχανών αναζήτησης και οι πολιτικές με τις οποίες το κάνουν αυτό. Στη συνέχεια, μελετώνται οι παράμετροι σχεδίασης και οι πιθανές δομές ενός ευρετηρίου, αναλύονται οι τρόποι με τους οποίους αναλύονται τα έγγραφα και οι δυσκολίες και προκλήσεις που αντιμετωπίζουν οι μηχανές αναζήτησης κατά την ευρετηρίαση των σελίδων. Τέλος, περιγράφεται ο τρόπος με τον οποίο οι μηχανές αναζήτησης επεξεργάζονται τα ερωτήματα, βάσει της διαδοχής των λέξεων και των διαφορετικών τελεστών αναζήτησης.

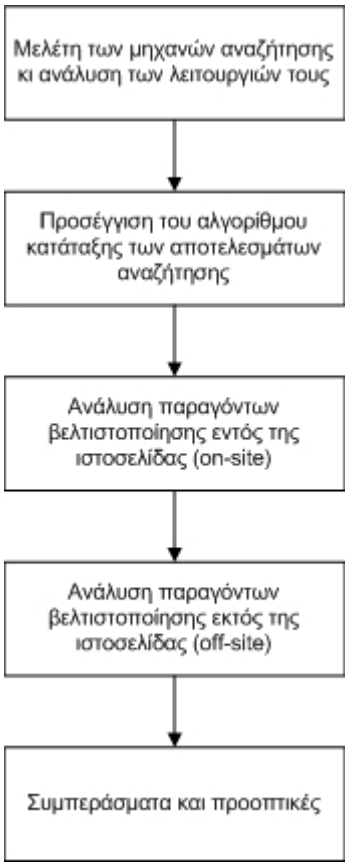
Στο 3ο κεφάλαιο, προσεγγίζεται ο αλγόριθμος κατάταξης των αποτελεσμάτων αναζήτησης της μηχανής Google, όπως περιγράφηκε από τους ερευνητές και ιδρυτές της Google, Lawrence Page και Sergey Brin, επιχειρήθηκε να μοντελοποιηθεί σε αντίστοιχες έρευνες στο παρελθόν, με σκοπό τον προσδιορισμό των εσωτερικών κι εξωτερικών, σχετικών με τον όρο αναζήτησης και μη παραγόντων που επιδρούν στους αλγόριθμους κατάταξης που οι σύγχρονες μηχανές αναζήτησης χρησιμοποιούν. Στη συνέχεια, παρουσιάζονται τα αποτελέσματα των ερευνών αυτών και η σημασία τους στη βιομηχανία της βελτιστοποίησης για τις μηχανές αναζήτησης.

Έπειτα, στο 4ο κεφάλαιο, μελετώνται κι αναλύονται οι παράμετροι επίδρασης στην κατάταξη των αποτελεσμάτων αναζήτησης, η βελτιστοποίηση των οποίων πραγματοποιείται αποκλειστικά σε επίπεδο ιστοσελίδας και διακομιστή φιλοξενίας αυτής (on-page optimization). Έτσι, γίνονται αντικείμενο μελέτης παράγοντες σε επίπεδο ιστοσελίδας, όπως η συχνότητα των λέξεων – κλειδιών στο κείμενο, οι διάφορες meta ετικέτες, παράγοντες μορφοποίησης του περιεχομένου (κείμενο, φωτογραφίες, σύνδεσμοι, αντικείμενα Flash), οι διευθύνσεις URL και η εσωτερική δομή του ιστοτόπου, καθώς και παράγοντες σε επίπεδο διακομιστή, όπως το πρωτόκολλο αποκλεισμού ανιχνευτών, οι χάρτες ιστοτόπων, τα χαρακτηριστικά του τομέα (domain), η κανονικοποίηση και η ανακατεύθυνση, αλλά και ο παράγοντας του χρόνου. Παράλληλα, κατά τη μελέτη κάθε παράγοντα, παρουσιάζονται προτεινόμενες πρακτικές αξιοποίησης της γνώσης που εξάγεται από την ανάλυση αυτή, οι οποίες κι εφαρμόζονται σε παραδείγματα (test cases), που βασίζονται σε πραγματικές ιστοσελίδες ή σελίδες που δημιουργήθηκαν στα πλαίσια της εργασίας.

Στο 5ο κεφάλαιο, πραγματοποιείται μία εκτενής ανάλυση του αλγορίθμου και του βαθμού PageRank, όπως αναπτύχθηκε από τους ιδρυτές της μηχανής Google, καθώς και του τρόπου με τον οποίο χαρακτηρίζει κάθε ιστοσελίδα του Διαδικτύου και επηρεάζει τους αλγορίθμους κατάταξης των εγγράφων από τις μηχανές αναζήτησης. Παράλληλα, περιγράφονται οι αλγόριθμοι που έχουν, στο παρελθόν, μελετηθεί για την εξατομίκευση του PageRank και θέτουν τις βάσεις για μελλοντική έρευνα και εξέλιξη, στα πλαίσια της ανάπτυξης του Ιστού. Έπειτα, συμπεραίνονται ορισμένες βασικές αρχές που χαρακτηρίζουν το βαθμό PageRank και, με γνώμονα αυτές, αναλύεται η διαδικασία της κατασκευής συνδέσμων, προσεγγίζοντας τους διαφορετικούς τρόπους με τους οποίους αυτή πραγματοποιείται, και παρουσιάζονται, σε αντιστοιχία με το κεφάλαιο που πραγματεύεται τους εσωτερικούς παράγοντες βελτιστοποίησης, τεχνικές και προτάσεις αξιοποίησης της έρευνας για τη βελτιστοποίηση των εξωτερικών παραγόντων κατάταξης των ιστοσελίδων (off-site optimization).

Τέλος, στο 6ο κεφάλαιο, συνοψίζονται τα συμπεράσματα της εργασίας και αναλύεται πώς αυτά αντιμετωπίζουν το αρχικό πρόβλημα που περιγράφηκε παραπάνω, ενώ αναπτύσσονται

ορισμένες ιδέες για την εξέλιξη και μελλοντική αξιοποίηση των συμπερασμάτων αυτών, στα πλαίσια της προσαρμογής των μηχανών αναζήτησης και των ιστοσελίδων στον Σημασιολογικό Ιστό.



Εικόνα 2 Η δομή της διπλωματικής εργασίας

1.3 Προσδιορισμός των στόχων της βελτιστοποίησης

Πριν την εφαρμογή των διαφόρων τεχνικών βελτιστοποίησης της κατάταξης μίας ιστοσελίδας στα αποτελέσματα αναζήτησης, πρέπει να προσδιορίζονται οι στόχοι της διαδικασίας.

Σε πρώτο στάδιο, οφείλουμε να προσδιορίσουμε τους στόχους του Search Engine Marketing, γενικότερα. Τέτοιοι μπορεί να είναι οι εξής:

- **Αύξηση της επισκεψιμότητας**

  Η βελτιστοποίηση ενός ιστοτόπου για τις μηχανές αναζήτησης οδηγεί περισσότερους χρήστες σε αυτόν, δημιουργεί δηλαδή απευθείας κίνηση (direct traffic) στον ιστότοπο. Εάν σε αυτό αποσκοπεί η δραστηριότητα της βελτιστοποίησης, προφανώς επιδιώκεται η βελτίωση της κατάταξης του ιστοτόπου σε όσο το δυνατόν περισσότερους όρους

αναζήτησης και όχι μόνο σε επιλεγμένες, εύκολες για βελτιστοποίηση και περιορισμένες λέξεις – κλειδιά, ενώ το περιεχόμενο και η παρουσίαση αυτού πρέπει να είναι έτσι ώστε να ενθαρρύνουν τη διάδοσή τους.

- **Επίτευξη οικονομικών στόχων**

Στην περίπτωση αυτή, η επιχείρηση αποσκοπεί στην προσέλκυση πελατών και όχι επισκεπτών. Καθώς, πολλές φορές, τα επίπεδα επισκεψιμότητας είναι αντιστρόφως ανάλογα με τα επίπεδα πωλήσεων, είναι προτιμότερη η επένδυση σε αυστηρά καθορισμένους, ίσως και σπάνιους (μακροσκελούς ουράς αναζήτησης, όπως αυτή αναλύεται στο παράρτημα Α) όρους αναζήτησης, οι οποίοι συνοδεύονται από υψηλό δείκτη ROI (Return On Investment). Είναι προφανές ότι ο στόχος αυτός επιδιώκεται από επιχειρήσεις που δραστηριοποιούνται σε μεγάλο βαθμό στο Διαδίκτυο και προσφέρουν δυνατότητες απευθείας ηλεκτρονικών αγορών. Έτσι, ένα κατάστημα ηλεκτρονικών ειδών οφείλει να επιδιώκει να κατατάσσεται υψηλά στα αποτελέσματα των μηχανών για όρους αναζήτησης που σχετίζονται με τα μοντέλα των προϊόντων που εμπορεύεται, παρά τον όρο «ηλεκτρονικά είδη».

- **Επίτευξη μη οικονομικών στόχων**

Υπάρχουν περιπτώσεις στις οποίες μία εταιρεία δε δραστηριοποιείται στον τομέα του e-shop, επομένως δεν επιδιώκει τις άμεσες πωλήσεις μέσω του ιστοτόπου της. Το ίδιο, προφανώς, ισχύει για το σύνολο σχεδόν των οργανισμών (π.χ. φιλανθρωπικά ιδρύματα, φιλοζωικές οργανώσεις), που δεν σχετίζονται με την αγορά και τους κανόνες της. Έτσι, ο στόχος, στις περιπτώσεις αυτές, εκφράζεται με κάποιον άλλον τρόπο, όπως, για παράδειγμα, είναι η εγγραφή σε κάποιο newsletter, η άμεση επικοινωνία (μέσω ηλεκτρονικής φόρμας, τηλεφωνική συνομιλίας, ηλεκτρονικής αλληλογραφίας ή επίσκεψης στο φυσικό κατάστημα), η εγγραφή στη σελίδα ενός κοινωνικού δικτύου του φορέα, η συμμετοχή σε διαγωνισμούς, ακόμη και κάποια δωρεά φιλανθρωπικού χαρακτήρα.

- **Η αναγνωρισιμότητα (branding)**

Σε αυτή την περίπτωση, η επιχείρηση (ή ο οργανισμός) επιδιώκει να προβληθεί στην αγορά, στους χρήστες του Διαδικτύου, ανεξάρτητα από τον τομέα δραστηριοποίησής της, ενώ αποσκοπεί στην εμφάνιση του ιστοτόπου στα αποτελέσματα αναζήτησης για λέξεις και φράσεις τετριμμένες και πολύ γενικές με τις οποίες συνδέεται άμεσα. Στο παραπάνω παράδειγμα με το κατάστημα ηλεκτρονικών ειδών, η επιχείρηση επιδιώκει περισσότερο την επίτευξη υψηλής ορατότητας στα αποτελέσματα των μηχανών αναζήτησης για τους όρους «ηλεκτρονικά είδη», «ανακύκλωση ηλεκτρικών συσκευών», «οικιακές συσκευές», παρά όρους σχετικούς με τα μοντέλα των προϊόντων.

- **Ιδεολογική επιρροή**

Πολλές φορές επιδιώκεται η προώθηση ιδεών ή η επιρροή όσο το δυνατόν περισσότερων χρηστών προς μία κατεύθυνση. Πρόκειται για έναν στόχο αρκετά παρόμοιο με αυτόν του branding. Στην περίπτωση αυτή, η βελτιστοποίηση πραγματοποιείται κυρίως εξωτερικά και όχι σε επίπεδο ιστοσελίδας ή διακομιστή, με την αξιοποίηση της δύναμης των συνδέσμων (οι οποίοι, άλλωστε, μεταφράζονται ως «ψήφοι»). Παράδειγμα αυτής της κατηγορίας υπήρξε η δυσφήμηση που συνέχισε να επιτυγχάνεται, μέχρι πρόσφατα, με την τακτική του Google-Bombing, που χρησιμοποιήθηκε κατά κόρον με στόχους αρκετούς αμερικανούς ρεπουμπλικάνους βουλευτές (Zeller, 2006).

- **Διαχείριση εταιρικής φήμης**

Σε περίπτωση αρνητικών κριτικών και άρθρων σε ιστοσελίδες στις οποίες η επιχείρηση ή ο οργανισμός που θίγεται δεν έχει δικαιοδοσία, όπως συνήθως συμβαίνει, είναι επιθυμητό να «θαφτούν» αυτές οι σελίδες στα αποτελέσματα αναζήτησης, δηλαδή επιδιώκεται η βελτιστοποίηση της κατάταξης πολλών σελίδων από διαφορετικούς διακομιστές που θα ξεπεράσουν τις αρνητικές ιστοσελίδες που εμφανίζονται στα αποτελέσματα αναζήτησης. Πρόκειται για την πιο απαιτητική εκδοχή της βελτιστοποίησης, καθώς επιδιώκεται η επανάληψη της διαδικασίας για πολλές σελίδες πάνω στον ίδιο όρο αναζήτησης, σε αντίθεση με όλες τις περιπτώσεις που αναφέρθηκαν προηγουμένως. Παράλληλα, η βελτιστοποίηση των ιστοσελίδων είναι μία διαδικασία παραμετροποίησης των διαφόρων παραγόντων που οι μηχανές αναζήτησης λαμβάνουν υπόψη κατά την εκτέλεση των αλγορίθμων κατάταξης και όχι απαραίτητα βελτίωσης της κατάταξης των ιστοσελίδων. Πολλές τεχνικές βελτιστοποίησης δεν αποβλέπουν αποκλειστικά στην άνοδο μίας ιστοσελίδας στα αποτελέσματα αναζήτησης, αλλά μπορούν να ωφελήσουν στο εκ διαμέτρου αντίθετο αποτέλεσμα. Αυτό επιδιώκεται όταν μία επιχείρηση ή οργανισμός παρέχει στον ιστοχώρο της δεδομένα που δε θέλει να γνωστοποιούνται και να είναι προσβάσιμα από τις μηχανές αναζήτησης (ευαίσθητα προσωπικά στοιχεία, κακές κριτικές, βάσεις δεδομένων, σελίδες σχολίων, φόρμες αγοράς, κατάλογοι τιμολόγησης).

Οι παραπάνω περιπτώσεις συνοψίζουν την πλειοψηφία των διαφορετικών στόχων που μία επιχείρηση ή ένας οργανισμός ενδέχεται να θέτει και να επιδιώκει να επιτύχει μέσα από την ηλεκτρονική δραστηριοποίηση. Η επιλογή του κατάλληλου στόχου σχετίζεται με τη φύση της επιχείρησης, πρωτίστως, και τις ανάγκες της, δευτερευόντως.

Σε δεύτερο στάδιο και βάσει των στόχων της βελτιστοποίησης, πρέπει να επιλεχθούν προσεκτικά οι λέξεις – κλειδιά γύρω από τις οποίες θα επικεντρωθεί η δραστηριότητα. Η στρατηγική που ακολουθείται για την επιλογή των κατάλληλων όρων βελτιστοποίησης αναπτύσσεται στο παράρτημα Α.

1.4 Η ηθικότητα της βελτιστοποίησης

Ένα ζήτημα που προκύπτει από τη φιλοσοφία του SEO είναι η ηθικότητά του. Η ίδια του η γενική ιδέα αφορά την όσο το δυνατόν πληρέστερη εκμάθηση των αλγορίθμων κατάταξης των μηχανών αναζήτησης και τη χειραγώγηση των παραγόντων που επιδρούν σε αυτούς.

Προφανώς, όμως, προσεγγίζοντας το θέμα από την σκοπιά των μηχανών αναζήτησης, δεν πρόκειται περί χειραγώγησης των παραγόντων αυτών αλλά παραμετροποίησης των στοιχείων που σχετίζονται με αυτούς, έτσι ώστε να διευκολύνεται το έργο των μηχανών. Έτσι, η διαδικασία της βελτιστοποίησης ξεκινά με τον προσδιορισμό του τι επιζητούν οι χρήστες του διαδικτύου, κατά τη διεξαγωγή μίας αναζήτησης, και τη δημιουργία του κατάλληλου περιεχομένου ενώ, στη συνέχεια, οι σελίδες που απαρτίζουν έναν ιστότοπο καθίστανται προσβάσιμες στις μηχανές αναζήτησης, αναδεικνύοντας τα στοιχεία εκείνα που, κατ' εκτίμηση, περιγράφουν και χαρακτηρίζουν το περιεχόμενο του ιστοτόπου, έτσι ώστε οι μηχανές να συσχετίσουν τον ιστότοπο αυτό με ενδεχόμενους όρους αναζήτησης.

Άλλωστε, τα αποτελέσματα αναζήτησης καθαυτά αποτελούν μία μορφή διαφήμισης, καθώς ο χρήστης μπορεί να τα αγνοήσει ή να τα ακολουθήσει, προσφέροντάς του μία ευκαιρία να ανακαλύψει κάτι σχετικό με την αναζήτησή του και χρήσιμο. Με τον τρόπο αυτό και από αυτή την οπτική γωνία, ο στόχος της διαφήμισης είναι να βελτιστοποιήσει τη διαφήμιση αυτή αλλά και την πιθανότητα να προβληθεί αυτή στους χρήστες.

Γίνεται εμφανές ότι το SEO, όπως κάθε άλλη μορφή επιρροής μπορεί να έχει ένας άνθρωπος πάνω σε κάποιον άλλο μπορεί να χρησιμοποιηθεί είτε για καλό σκοπό, είτε για κακό, και οι τεχνικές που μπορούν να εφαρμοσθούν, στα πλαίσια της προσαρμογής μίας ιστοσελίδας, προσφέρουν τη δυνατότητα και τα περιθώρια και για τα δύο.

Έτσι, οι μηχανές αναζήτησης με τη συμπεριφορά τους (επιβράβευση ή απόδοση κάποιας ποινής, μέσω της κατάταξης των ιστοσελίδων στα αποτελέσματα) δείχνουν την κατεύθυνση προς την οποία διαχωρίζονται οι διάφορες τεχνικές βελτιστοποίησης των ιστοσελίδων, σε ηθικές (**white-hat**) και ανήθικες (**black-hat**).

Στις δεύτερες, εντάσσεται το παραγέμισμα με λέξεις ή φράσεις – κλειδιά (keyword stuffing / spamming) των meta ετικετών, η απόκρυψη κώδικα και περιεχομένου πίσω από το εμφανές περιεχόμενο μίας ιστοσελίδας, οι φάρμες συνδέσμων, το διπλότυπο περιεχόμενο και η εφαρμογή των διαφόρων τεχνικών για τη σύνδεση ενός ιστοτόπου με όρους αναζήτησης που δε σχετίζονται, σε κανένα επίπεδο, με αυτόν.

Στα πλαίσια και για τους σκοπούς της παρούσας διπλωματικής, προσεγγίζονται, αναλύονται αλλά και συνίσταται να εφαρμόζονται αποκλειστικά οι white-hat τεχνικές βελτιστοποίησης των ιστοσελίδων.

2 *Ανίχνευση, Ευρετηρίαση & Επεξεργασία Ερωτημάτων*

2.1 Κατηγορίες μηχανών αναζήτησης

Ο όρος μηχανή αναζήτησης χρησιμοποιείται για να περιγράψει τόσο τις μηχανές που βασίζουν τη λειτουργία τους σε μηχανισμούς ανίχνευσης σελίδων (crawler – based μηχανές) όσο και τις μηχανές που λειτουργούν με χειροκίνητους καταλόγους (human – powered directories). Αυτοί οι δύο τύποι μηχανών αναζήτησης συγκεντρώνουν τις καταχωρήσεις τους με πολύ διαφορετικό τρόπο. Η διαφορά των δύο έγκειται στο γεγονός ότι οι μηχανές της δεύτερης κατηγορίας που αναφέρθηκε απαιτούν την ύπαρξη και συμμετοχή του ανθρώπινου παράγοντα για την επιλογή, καταχώρηση και κατάταξη των εγγράφων στους καταλόγους.

2.1.1 Crawler – based μηχανές

Αυτές οι μηχανές αναζήτησης περιλαμβάνουν μια λίστα κρίσιμων λειτουργιών 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 τις σελίδες και τα έγγραφα που συναποτελούν τον Παγκόσμιο Ιστό.

2. **Ευρετηρίαση εγγράφων (Indexing)**

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

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

Όταν πραγματοποιείται μία αίτηση για πληροφορία εκ των εκατοντάδων εκατομμυρίων που πραγματοποιούνται σε μία μέρα, 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 *φροντιστήρια γερμανικών*, στην παγκόσμια εκδοχή της μηχανής Google, επιστρέφει περίπου 3.900.000 αποτελέσματα, ενώ η ίδια αναζήτηση με εισαγωγικά ("φροντιστήρια γερμανικών") επιστρέφει 33.200 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 «φροντιστήρια» και «γερμανικών», συμπεριλαμβανομένων όλων των πτώσεων του ουσιαστικού *φροντιστήριο* και του επιθέτου *γερμανικός*. Στη δεύτερη αναζήτηση, η μηχανή επέστρεψε μόνο τα έγγραφα που περιελάμβαναν την ακριβή φράση «φροντιστήρια γερμανικών». Άλλοι προχωρημένοι τελεστές (η Google περιλαμβάνει περισσότερους από 40 τελεστές για τη συγκεκριμενοποίηση και διευκόλυνση των αναζητήσεων, παρότι, στην ιστοσελίδα του κέντρου βοηθείας της ίδιας, καταγράφει και παρουσιάζονται μόλις 6 βασικοί τελεστές αναζήτησης) μπορούν να μεταβάλλουν τον τρόπο με τον οποίο μία μηχανή αναζήτησης κρίνει την αντιστοιχία αποτελεσμάτων και ερωτημάτων.

4. **Κατάταξη αποτελεσμάτων (Ranking)**

Όταν η μηχανή καθορίσει ποια αποτελέσματα αντιστοιχούν σε ένα ερώτημα, ο αλγόριθμος της μηχανής (μια μαθηματική εξίσωση που συνήθως χρησιμοποιείται για την ταξινόμηση) εκτελεί υπολογισμούς σε κάθε ένα αποτέλεσμα για να καθορίσει τον βαθμό σχετικότητας 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.

Παρότι οι λειτουργίες μίας μηχανής αναζήτησης που βασίζεται στην αυτόματη ανίχνευση σελίδων δεν είναι ιδιαίτερα εκτενείς ή χρονοβόρες, συστήματα όπως αυτό της Google, της Yahoo!, της AskJeeves (Ask.com) ή της MSN (bing) βρίσκονται μεταξύ των πιο σύνθετων, εντατικών υπολογιστικών συστημάτων στον κόσμο, καθώς διαχειρίζονται εκατομμύρια υπολογισμών το δευτερόλεπτο και συγκεντρώνουν κι εξυπηρετούν αιτήματα για πληροφορίες ταυτόχρονα σε ένα τεράστιο αριθμό χρηστών.

2.1.2 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 στον κατάλογο 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 δε φέρουν αντίστοιχες αλλαγές στην καταχώρηση καθ' αυτήν. Οι μέθοδοι βελτιστοποίησης, που είναι το αντικείμενο της παρούσας διπλωματικής, 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.

Για τους σκοπούς του παρόντος κεφαλαίου, θα μελετήσουμε την αυτοματοποιημένη καταχώρηση των ιστοσελίδων στις crawler – based μηχανές αναζήτησης.

2.2 Ιστορική αναδρομή

Κατά την πρώιμη ανάπτυξη του Ιστού, υπήρχε μία λίστα από διακομιστές (webservers), την οποία συνέταξε ο Tim Berners-Lee και που φιλοξενήθηκε στο διακομιστή του Ευρωπαϊκού Οργανισμού Πυρηνικών Ερευνών (CERN – Centre Européen pour la Recherche Nucléaire). Καθώς περισσότεροι διακομιστές συνδέονταν, η λίστα αυτή δεν ήταν δυνατό να ανταποκριθεί στις εξελίξεις, ενώ στην ιστοσελίδα του αμερικάνικου Διεθνούς Κέντρου Υπερυπολογιστικών Εφαρμογών (NCSA - National Center for Supercomputing Applications), το οποίο διέθετε κι

έναν από τους πρωταρχικούς διακομιστές της προαναφερθείσας λίστας, η λειτουργία νέων διακομιστών ανακοινώνονταν διαρκώς, υπό τον τίτλο «What's new!».

Το πρώτο εργαλείο που χρησιμοποιήθηκε στο Διαδίκτυο ήταν το Archie (εκ του archive, που σημαίνει αρχείο), το οποίο δημιουργήθηκε το 1990 από τους Alan Emtage, Bill Heelan και Peter Deutsch, φοιτητές της Επιστήμης Υπολογιστών στο Πανεπιστήμιο McGill του Μοντρεάλ, στον Καναδά. Το πρόγραμμα μεταφόρτωνε όλες τις καταχωρήσεις καταλόγου όλων των αρχείων που βρίσκονταν σε δημόσιες και ανώνυμες ιστοσελίδες με Πρωτόκολλο Μεταφοράς Δεδομένων (FTP – File Transfer Protocol), δημιουργώντας μία ερευνήσιμη βάση δεδομένων με τα ονόματα των αρχείων. Όμως, το Archie δεν ευρετηρίαζε το περιεχόμενο των σελίδων αυτών, εφόσον η ποσότητα των δεδομένων ήταν περιορισμένη με τρόπο τέτοιο ώστε να είναι ευανάγνωστα.

Το 1991, ο Mark McCahill, στο Πανεπιστήμιο της Μινεσότα, δημιούργησε το Gopher, η ανάπτυξη του οποίου οδήγησε στη δημιουργία δύο νέων προγραμμάτων αναζήτησης, των Veronica και Jughead. Όπως το Archie, τα δύο αυτά προγράμματα αναζητούσαν ονόματα και τίτλους αρχείων που αποθηκεύονταν στα συστήματα ευρετηρίασης του Gopher. Το Veronica (Very Easy Rodent – Oriented Net-wide / index to Computerized Archives) παρείχε μία φόρμα αναζήτησης μίας λέξης – κλειδιού μεταξύ των περισσότερων τίτλων καταλόγου σε ολόκληρο το ευρετήριο Gopher. Το Jughead (Jonzy's Universal Gopher Hierarchy Excavation And Display) αποτελούσε ένα εργαλείο για την απόκτηση πληροφοριών καταλόγου από συγκεκριμένους εξυπηρετητές του Gopher. Παρότι το όνομα της μηχανής αναζήτησης «Archie» δεν ήταν εμπνευσμένο από τη διάσημη ομώνυμη σειρά κόμικ βιβλίων, τα ονόματα «Veronica» και «Jughead» αποτελούν χαρακτήρες της σειράς, κάνοντας παραπομπή στον προκάτοχό τους.

Έως το καλοκαίρι του 1993, καμία μηχανή αναζήτησης δεν υπήρχε για τον Ιστό, παρότι αρκετοί ειδικευμένοι κατάλογοι διατηρούνταν με το χέρι. Ο Oscar Nierstrasz, στο Πανεπιστήμιο της Γενεύης, έγραψε μία σειρά από αρχεία δέσμης ενεργειών Perl που, κατά περιόδους, αντέγραφαν τις σελίδες αυτές και τις ξαναέγραφαν σε πρότυπη μορφή, που αργότερα αποτέλεσε τη βάση του W3Catalog, την πρώτη αρχέγονη μηχανή αναζήτησης του Παγκόσμιου Ιστού, η λειτουργία της οποίας ανακοινώθηκε στις 2 Σεπτεμβρίου 1993.

Τον Ιούνιο του 1993, ο Matthey Gray, στο Τεχνολογικό Ίδρυμα της Μασσαχουσέτης (MIT – Massachusetts Institute of Technology), παρήγαγε το πρώτο διαδικτυακό ρομπότ (web robot), το βασισμένο σε Perl «World Wide Web Wanderer» ή αλλιώς «Περιηγητή του Παγκόσμιου Ιστού». Ο σκοπός του «Περιηγητή» ήταν η καταμέτρηση του μεγέθους του Παγκόσμιου Ιστού και επετεύχθη έως τα τέλη του 1995. Η δεύτερη μηχανή αναζήτησης του web «Aliweb» εμφανίστηκε το Νοέμβριο του 1993 και η λειτουργία της δε βασιζόταν σε

διαδικτυακό ρομπότ, αλλά στην ενημέρωσή του από τους διαχειριστές της ιστοσελίδας για την ύπαρξη σε κάθε σελίδα ενός αρχείου ευρετηρίασης, σε συγκεκριμένη μορφή.

Το Jumpstation, που εκδόθηκε τον Δεκέμβριο του 1993, βάσιζε τη λειτουργία του στη χρήση ενός διαδικτυακού ρομπότ για την εύρεση σελίδων και την κατασκευή του ευρετηρίου του, καθώς και στη χρήση μίας διαδικτυακής φόρμας, ως διασύνδεσης με το πρόγραμμα ερωτημάτων. Η λειτουργία του αυτή καθιστά το Jumpstation το πρώτο εργαλείο αναζήτησης του Παγκόσμιου Ιστού που συνδυάζει τα τέσσερα βασικά χαρακτηριστικά μιας διαδικτυακής μηχανής αναζήτησης (όπως αυτά αναπτύχθηκαν παραπάνω). Επειδή οι διαθέσιμες πηγές ήταν περιορισμένες, όμως, η ευρετηρίαση και αναζήτηση (επεξεργασία ερωτημάτων) περιορίζονταν στους τίτλους και τις ενότητες που υπήρχαν στις ιστοσελίδες που ο ανιχνευτής (crawler) συναντούσε.

Μία από τις πρώτες crawler - based μηχανές αναζήτησης ήταν ο WebCrawler, ο οποίος εκδόθηκε το 1994. Αντίθετα από τους προκατόχους του, επέτρεπε στους χρήστες να αναζητήσουν για οποιαδήποτε λέξη σε μία ιστοσελίδα, χαρακτηριστικό που συνοδεύει, ως πρότυπο, πλέον, όλες τις μεγάλες μηχανές έκτοτε, γι' αυτό και η λειτουργία του WebCrawler εξετάζεται λεπτομερώς. Ήταν, επίσης, και η πρώτη μηχανή που έγινε ευρέως γνωστή στο κοινό. Επίσης, το 1994, το Πανεπιστήμιο Carnegie Mellon λάνσαρε την Lycos, η οποία εξελίχθηκε σε τεράστια εμπορική επιτυχία.

Λίγο αργότερα, αρκετές μηχανές αναζήτησης έκαναν την εμφάνισή τους, αναζητώντας δημοσιότητα. Τέτοιες είναι οι Magellan, Excite, Infoseek και Altavista. Η Yahoo! αποτέλεσε έναν από τους πιο δημοφιλείς τρόπους εύρεσης ιστοσελίδων ενδιαφέροντος, αλλά η λειτουργία αναζήτησης της βασιζόταν στον δικό της διαδικτυακό κατάλογο. Όσοι χρήστες αναζητούσαν πληροφορίες μπορούσαν, επίσης, να περιηγηθούν στον κατάλογο, αντί να αναζητήσουν λέξεις - κλειδιά.

Το 1996, η Netscape αναζητούσε αποκλειστική συμφωνία με μία μηχανή αναζήτησης, γεγονός που προκάλεσε ιδιαίτερα μεγάλο ενδιαφέρον, με αποτέλεσμα να πραγματοποιηθεί συμφωνία της Netscape με πέντε εκ των μεγαλύτερων μηχανών αναζήτησης της αγοράς, σύμφωνα με την οποία οι πέντε μηχανές θα φιλοξενούνταν, με κυκλική εναλλαγή, στην σελίδα αναζήτησης της Netscape, για $5,000,000 το χρόνο. Οι μηχανές αυτές ήταν οι Yahoo!, Magellan, Lycos, Infoseek και Excite.

Το 2000, η μηχανή αναζήτησης της Google είχε ξεκινήσει την κυριαρχία της στην αγορά. Ο αλγόριθμος της εταιρείας επιτύγχανε καλύτερα αποτελέσματα για πολλές αναζητήσεις, λόγω μιας καινοτομίας, του βαθμού PageRank. Αυτός ο επαναληπτικός αλγόριθμος ταξινομεί τις ιστοσελίδες, με κριτήριο τον αριθμό και το βαθμό PageRank άλλων ιστοτόπων και ιστοσελίδων που παρείχαν σύνδεσμο σε αυτές, έχοντας ως υπόβαθρο την υπόθεση ότι οι καλές ή επιθυμητές, άρα και πλέον χρηστές, σελίδες στο Διαδίκτυο θα δέχονταν

περισσότερους συνδέσμους απ' ότι άλλες. Η Google, επίσης, διατηρούσε, όπως άλλωστε εξακολουθεί να διατηρεί, ένα μινιμαλιστικό και λιτό περιβάλλον χρήσης, σε αντίθεση με τους βασικότερους ανταγωνιστές της που ενσωμάτωναν τη διασύνδεση με τη μηχανή αναζήτησης σε κάποια κεντρική διαδικτυακή πύλη.

Από την άλλη πλευρά, η Yahoo! παρείχε υπηρεσίες αναζήτησης, βασιζόμενη στη μηχανή αναζήτησης της Inktomi, την οποία και αγόρασε το 2002, ενώ ένα χρόνο αργότερα αγόρασε και την Overture (που κατείχε τις μηχανές AlltheWeb και AltaVista). Στη συνέχεια, η Yahoo! Χρησιμοποίησε τη μηχανή αναζήτησης της Google, μέχρι το 2004 που ξεκίνησε να λειτουργεί τη δική της μηχανή, βάσει των συνδυασμένων τεχνολογιών των αποκτημάτων της.

Η Microsoft ανακοίνωσε τη λειτουργία της μηχανής MSN Search, το φθινόπωρο του 1998, χρησιμοποιώντας αποτελέσματα αναζήτησης από την Inktomi. Στις αρχές του 1999, ο ιστότοπος της Microsoft ξεκίνησε να παρέχει εγγραφές από τη Looksmart, συνδυάζοντάς τα με τα αποτελέσματα της Inktomi, με εξαίρεση ένα μικρό διάστημα του ίδιου έτους που χρησιμοποίησε αποτελέσματα της μηχανής AltaVista. Το 2004, η Microsoft ξεκίνησε μία μετάβαση στη δική της τεχνολογία αναζήτησης, κάνοντας χρήση του δικού της ανιχνευτή Ιστού (web crawler) που έφερε το όνομα msnbot.

Τέλος, η Microsoft ανακοίνωσε τη λειτουργία του Bing, της νέας μηχανής αναζήτησης της, την 1η Ιουνίου 2009, ενώ στις 29 Ιουλίου του ίδιου έτους, η Yahoo! ήρθε σε συμφωνία με τη Microsoft, σύμφωνα με την οποία η ιστοσελίδα αναζήτησης Yahoo! Search θα ενισχυόταν με την τεχνολογία του Microsoft Bing.

2.3 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. Άλλες λέξεις, συνώνυμα, που χρησιμοποιούνται είναι τα «ants», «automatic indexers», «bots», «Web spiders» ή «Web robots».

Η διαδικασία αυτή ονομάζεται «Ανίχνευση Ιστού» (web crawling ή spidering). Πολλές ιστοσελίδες, στην πλειοψηφία τους μηχανές αναζήτησης, χρησιμοποιούν την Ανίχνευση για να παρέχουν στους χρήστες ανανεωμένα δεδομένα. Οι ανιχνευτές Ιστού χρησιμοποιούνται κυρίως για τη δημιουργία ενός αντιγράφου 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 να αυτοματοποίηση διαδικασιών συντήρησης σε δεδομένη ιστοσελίδα, όπως ο έλεγχος των σύνδεσμος ή η επικύρωση του HTML κώδικα.

Τέλος, οι ανιχνευτές μπορούν να χρησιμοποιηθούν για να συγκεντρώσουν συγκεκριμένους τύπους πληροφοριών από ιστοσελίδες, όπως η συγκομιδή διευθύνσεων ηλεκτρονικής αλληλογραφίας (συνήθως για πρακτικές «spam», δηλαδή αποστολής ανεπιθύμητων ηλεκτρονικών μηνυμάτων).

Ο ανιχνευτής Ιστού 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.

2.3.1 Πολιτικές ανίχνευσης

Το μεγάλο μέγεθός 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 μέγεθος 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 έχει επίσης καταστήσει δύσκολη την αποφυγή ανάκτησης διπλού περιεχομένου. Υπάρχουν ατελείωτοι συνδυασμοί παραμέτρων HTTP GET, εκ των οποίων μόνο ένα μικρό ποσοστό θα επιστρέψει πραγματικά μοναδικό περιεχόμενο. Για παράδειγμα, μία απλή φωτογραφική γκαλερί μπορεί να προσφέρει στους χρήστες τρεις επιλογές, όπως καθορίζεται από τις παραμέτρους HTTP GET στην τοποθεσία URL. Εάν υπάρχουν τέσσερις τρόποι ταξινόμησης των εικόνων, τρεις επιλογές μεγέθους, δύο υποστηριζόμενοι τύποι αρχείου και μία επιλογή απενεργοποίησης περιεχομένου παρεχόμενου από τον χρήστη, τότε στο ίδιο σύνολο περιεχομένου μπορεί να δίνεται πρόσβαση με 48 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.

Όπως οι Edwards, McCurley και Tomley (2001) σημείωναν, δεδομένου ότι το εύρος διασύνδεσης (bandwith) για τη διεξαγωγή ανιχνεύσεων δεν είναι δωρεάν ή απεριόριστο, γίνεται αναγκαίο να ανιχνεύεται ο Παγκόσμιος Ιστός με τρόπο όχι μόνο κλιμακωτό, αλλά και αποδοτικό, εάν ο στόχος είναι η επίτευξη κάποιου εύλογου επιπέδου ποιότητας ή φρεσκάδας. 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. (Castillo, 2005)

2.3.1.1 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, 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 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. Ο σχεδιασμός μίας καλής πολιτικής επιλογής ενέχει μία επιπρόσθετη δυσκολία: πρέπει να λειτουργεί με μερική πληροφορία, καθώς το σύνολο των ιστοσελίδων δεν είναι γνωστό κατά τη διάρκεια της ανίχνευσης.

Οι Cho, Garcia-Molina και Page (1998) διεξήγαγαν την πρώτη μελέτη σε πολιτικές για τον προγραμματισμό ανίχνευσης. Το σύνολο δεδομένων τους αφορούσε μία ανίχνευση 180,000 σελίδων από τον διαδικτυακό τομέα (domain) του Πανεπιστημίου Stanford, Stanford.edu, η

προσομοίωση της οποίας πραγματοποιήθηκε με διαφορετικές στρατηγικές (breadth – first search, καταμέτρηση των συνδέσμων προς την ιστοσελίδα και υπολογισμός του βαθμού PageRank). Ένα από τα συμπεράσματα που εξήχθησαν ήταν εάν στόχος του ανιχνευτή είναι να μεταφορτώσει σελίδες με υψηλό βαθμό PageRank νωρίς στη διαδικασία της ανίχνευσης, τότε η στρατηγική του μερικού PageRank είναι βέλτιστη, ακολουθούμενη από τις άλλες δύο στρατηγικές. Ωστόσο, τα αποτελέσματα αυτά αφορούσαν έναν μόνο τομέα (domain).

Οι Najork και Wiener (2001) διεκπεραίωσαν μία ανίχνευση σε 328 εκατομμύρια σελίδες, χρησιμοποιώντας την στρατηγική breadth-first αναζήτησης. Βρήκαν ότι μία breadth – first ανίχνευση καταγράφει σελίδες με υψηλό PageRank νωρίς στην ανίχνευση (χωρίς, όμως, να συγκρίνουν τη στρατηγική αυτή έναντι άλλων στρατηγικών). Η εξήγηση που δόθηκε για το αποτέλεσμα αυτό είναι ότι «οι πλέον σημαντικές σελίδες δέχονται πολλούς συνδέσμους προς αυτές από πολυάριθμους εξυπηρετητές (hosts) και αυτοί οι σύνδεσμοι θα ευρεθούν σύντομα, ανεξάρτητα από τον εξυπηρετητή ή τη σελίδα απ' όπου ο ανιχνευτής προέρχεται».

Η Abiteboul σχεδίασε μία στρατηγική ανίχνευσης βασισμένη σε έναν αλγόριθμο, που ονομάζεται OPIC (On – line Page Importance Computation ή Σύγχρονος Υπολογισμός της Σπουδαιότητας της Σελίδας), με τον οποίο κάθε ιστοσελίδα λαμβάνει ένα αρχικό σύνολο από πόντους «μετρητών», οι οποίοι κατανέμονται εξίσου στις σελίδες στις οποίες αυτή κατευθύνει. Είναι παρόμοιος με τη διαδικασία υπολογισμού του βαθμού PageRank, με τη διαφορά ότι είναι ταχύτερος και πραγματοποιείται σε ένα μόλις βήμα. Ένας ανιχνευτής που οδηγείται από τον αλγόριθμο OPIC μεταφορτώνει πρώτα τις σελίδες στο σύνορο ανίχνευσης (crawl frontier) με τους μεγαλύτερους αριθμούς των πόντων «μετρητών».

Σε παρόμοια μελέτη, χρησιμοποιήθηκε προσομοίωση σε υποσύνολα του Παγκόσμιου Ιστού των 40 εκατομμυρίων σελίδων από τα ιταλικά ονόματα τομέα (domain names .it) και των 100 εκατομμυρίων σελίδων από την ανίχνευση WebBase, δοκιμάζοντας κατά πλάτος διάσχιση γράφου έναντι της κατά βάθος διάσχισης, τυχαίας κατάταξης. Η σύγκριση εξέταζε πόσο κοντά ο βαθμός PageRank που υπολογιζόταν σε μία μερική ανίχνευση πλησίαζε τον πραγματικό βαθμό. Προς έκπληξη των ερευνητών, ορισμένες επισκέψεις που υπολογίζουν το βαθμό PageRank ταχύτατα (κυρίως οι breadth-first) παρέχουν πολύ κακές εκτιμήσεις (Boldi et al., 2004). Τέλος, οι Baeza-Yates et al. (2005) προσομοίωσαν σε δύο υποσύνολα του Ιστού των 3 εκατομμυρίων σελίδων από ελληνικούς και χιλιανούς domains, εξετάζοντας διάφορες στρατηγικές ανίχνευσης. Τα αποτελέσματα της προσομοίωσης, συνοπτικά, έδειξαν ότι ο αλγόριθμος OPIC είναι καλύτερος από την ανίχνευση breadth – first, αλλά και ότι είναι ιδιαίτερα αποδοτικό να χρησιμοποιείται προηγούμενη ανίχνευση που θα καθοδηγήσει την παρούσα.

Επικεντρωμένη ανίχνευση

Η σπουδαιότητα μίας σελίδας για έναν ανιχνευτή μπορεί επίσης να εκφραστεί σαν μία συνάρτηση της ομοιότητας μίας σελίδας με ένα δεδομένο ερώτημα (όρο αναζήτησης). Οι ανιχνευτές Ιστού που προσπαθούν να μεταφορτώσουν σελίδες όμοιες η μία με την άλλη ονομάζονται επικεντρωμένοι ή τοπικοί ανιχνευτές.

Το κύριο πρόβλημα στην επικεντρωμένη ανίχνευση είναι ότι στα πλαίσια ενός ανιχνευτή Ιστού, είναι επιθυμητή η πρόβλεψη της ομοιότητας του κειμένου μιας σελίδας με τους όρους αναζήτησης, πριν την πραγματική μεταφόρτωση της σελίδας αυτής. Ένας πιθανός παράγοντας αυτής της πρόβλεψης είναι το anchor text των συνδέσμων, όπως προσέγγισε ο Pinkerton (2000), στον ανιχνευτή WebCrawler. Οι Diligenti et al. (2000) προτείνουν τη χρήση του συνολικού περιεχομένου των σελίδων που έχουν ήδη ανιχνευθεί για την υπόθεση της ομοιότητας μεταξύ του ερωτήματος (όρων αναζήτησης) και των σελίδων που δεν έχουν ανιχνευθεί ακόμα. Η αποτελεσματικότητα μίας επικεντρωμένης ανίχνευσης εξαρτάται κυρίως από τον πλούτο των συνδέσμων στο συγκεκριμένο θέμα υπό αναζήτηση, και μία επικεντρωμένη ανίχνευση συνήθως βασίζεται σε μία γενική Διαδικτυακή μηχανή αναζήτησης για την παροχή σημείων – αφετηριών.

Περιορισμός των συνδέσμων που ακολουθούνται

Ένας ανιχνευτής μπορεί να αναζητήσει μόνο HTML πόρους και να αποφύγει άλλους MIME τύπους αρχείων. Για την αίτηση μόνο HTML πηγών, ένας ανιχνευτής μπορεί να πραγματοποιεί μία αίτηση HTTP HEAD για να καθορίσει τον τύπο MIME του πόρου προτού πραγματοποιήσει αίτηση GET για τη λήψη ολόκληρης της απόκρισης. Για την αποφυγή πολλών αιτήσεων HEAD, ένας ανιχνευτής μπορεί να εξετάζει τη διεύθυνση URL και να αιτηθεί τον πόρο εάν αυτή καταλήγει σε συγκεκριμένους χαρακτήρες (επεκτάσεις αρχείων), όπως .html, .htm, .asp, .aspx, .php, .jsp, .jspx, ή κάθετο. Αυτή η στρατηγική μπορεί να προκαλέσει την ακούσια παράλειψη ανίχνευσης πολλών άλλων χρήσιμων πόρων.

Ορισμένοι ανιχνευτές μπορούν επίσης να αποφεύγουν την αίτηση πόρων που περιλαμβάνουν «?» (δηλαδή παράγονται δυναμικά και όχι στατικά), με στόχο την αποφυγή παγίδων για ανιχνευτές (spider traps). Η στρατηγική αυτή δεν είναι πλήρως έμπιστη, επίσης, καθώς ο ιστότοπος ενδέχεται να απλοποιεί τις URL διευθύνσεις του, με επανονομασία.

Κανονικοποίηση τοποθεσιών URL

Οι ανιχνευτές συνήθως εκτελούν κάποιο τύπο κανονικοποίησης (URL normalization ή canonicalization) για να αποφύγουν την ανίχνευση των ίδιων πόρων για περισσότερες από μία φορές. Ο όρος κανονικοποίηση των URL αναφέρεται στη διαδικασία τροποποίησης και

προτυποποίησης της URL με τρόπο σταθερό. Υπάρχουν αρκετοί τύποι κανονικοποίησης που μπορούν να χρησιμοποιηθούν, συμπεριλαμβανομένων αυτών της μετατροπής μίας τοποθεσίας URL σε πεζούς χαρακτήρες, της αφαίρεσης των χαρακτήρων «.» και «..», καθώς και της προσθήκης καταληκτικής καθέτου στο μονοπάτι του αρχείου.

Ανίχνευση ανάβασης μονοπατιού

Ορισμένοι ανιχνευτές σκοπεύουν να μεταφορτώνουν όσους το δυνατόν περισσότερους πόρους από μία συγκεκριμένα ιστοσελίδα. Επομένως, ο ανιχνευτής ανάβασης μονοπατιού εισήχθη για την ανάβαση και ανίχνευση κάθε φακέλου του μονοπατιού της διεύθυνσης URL που σκοπεύει να ανιχνεύσει. Για παράδειγμα, όταν ο ανιχνευτής επιχειρήσει να ανιχνεύσει τη σελίδα http://www.ntua.gr/files/gnu/linux/debian/latest.html, θα προσπαθήσει να ανιχνεύσει τους φακέλους /files, /files/gnu, /files/gnu/linux. Μάλιστα, ένας ανιχνευτής ανάβασης μονοπατιού θα μπορούσε να είναι ιδιαίτερα αποτελεσματικός στην εύρεση περιθωριοποιημένων πληροφοριών, ή πόρων για τους οποίους κανένας εισερχόμενος σύνδεσμος δεν θα είχε ευρεθεί με την κλασική ανίχνευση (Cothey, 2004).

2.3.1.2 Πολιτική επανεπίσκεψης

Ο Παγκόσμιος Ιστός έχει μία ιδιαίτερα δυναμική φύση και η ανίχνευση ενός μόλις μέρους του μπορεί να διαρκέσει πολύ. Τη στιγμή που ένας ανιχνευτής Ιστού έχει διεκπεραιώσει την ανίχνευση που του έχει ανατεθεί, πολλά γεγονότα μπορεί να έχουν συμβεί, συμπεριλαμβανομένων της δημιουργίας νέων, της ανανέωσης και της διαγραφής αρχείων.

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, οπότε υποθέτουμε ότι τουλάχιστον μία ανανέωση σελίδας, η οποία συμπεριλαμβάνει την προσθήκη συνδέσμου προς τη νέα ιστοσελίδα, πρέπει να πραγματοποιηθεί προτού η δημιουργία μίας ιστοσελίδας να είναι ορατή. Όπως αναφέρθηκε, 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. Επίσης, η ενημέρωση ενός ευρετηρίου και η δραστηριότητα ενός ανιχνευτή Ιστού μπορεί να εξαρτηθεί και από τις αιτήσεις της Ιστοσελίδας, βάσει μίας υγιούς σχέσης συνεργασίας μεταξύ του διακομιστή και του ανιχνευτή.

Αντίστοιχα, όταν μία σελίδα *ανανεώνεται*, η ενημέρωση μπορεί να είναι κύρια ή δευτερεύουσα. Η ειδοποιός διαφορά για τον χαρακτηρισμό της ανανέωσης δεν είναι πάντα

ευδιάκριτη. 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 a priori 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. Είναι αξιοσημείωτο το γεγονός ότι ακόμη κι αν όλοι οι (εσωτερικοί κι εξωτερικοί) σύνδεσμοι προς μία σελίδα αφαιρεθούν, η σελίδα είναι αόρατη και θεωρητικά μη προσβάσιμη μέσα στον ιστότοπο, αλλά παραμένει ορατή για τον ανιχνευτή ο οποίος, πλέον, γνωρίζει ότι πρέπει να την επανεπισκεφθεί, εφόσον αυτή έχει προστεθεί στο crawl frontier. Επίσης, είναι σχεδόν αδύνατο για τον ανιχνευτή να κρίνει αλλά και να θεωρηθεί ότι δύναται να κρίνει εάν μία σελίδα έχει χάσει όλους τους συνδέσμους προς αυτήν, καθώς ο ανιχνευτής δε μπορεί να καταγράφει όλες τις σελίδες που συνδέουν προς αυτήν ή εάν υπάρχουν σύνδεσμοι σε σελίδες που δεν έχουν ανιχνευθεί, προς το παρόν. Οι διαγραφές που δεν έχουν εντοπισθεί αποτελούν μεγαλύτερη ζημιά για τη φήμη μίας μηχανής αναζήτησης, καθώς είναι πιο εμφανείς στο χρήστη. Έρευνα που διεξήγαγαν οι Lawrence και Giles στην απόδοση των μηχανών αναζήτησης δείχνει ότι το 5,3% των συνδέσμων που επιστρέφουν οι μηχανές αναζήτησης, κατά μέσο όρο, οδηγούν σε διαγραμμένες σελίδες. (Castillo, 2005)

Συναρτήσεις κόστους

Για τη μηχανή αναζήτησης, υπάρχει ένα κόστος που αφορά τη μη ανίχνευση ενός γεγονότος. Οι πλέον διαδεδομένες συναρτήσεις κόστους είναι αυτές της φρεσκάδας και της ηλικίας.

Η φρεσκάδα αφορά ένα δυαδικό μέγεθος που υποδεικνύει εάν το τοπικό αντίγραφο είναι ακριβές ή όχι. Η φρεσκάδα μίας αποθηκευμένης σελίδας p, τη χρονική στιγμή p, προσδιορίζεται από τη σχέση

$$F_p(t) = \begin{cases} 1 & \text{εάν το p ισούται με το τοπικό αντίγραφο, τη στιγμή t} \\ 0 & \text{διαφορετικά} \end{cases}$$

Η ηλικία είναι ένα μέτρο που υποδεικνύει πόσο απαρχαιωμένο είναι το τοπικό αντίγραφο. Η ηλικία μίας αποθηκευμένης σελίδας p, την στιγμή t, για χρονική στιγμή τροποποίησης της σελίδας m, προσδιορίζεται από την σχέση

$$A_p(t) = \begin{cases} 0 & \text{εάν το p δεν έχει τροποποιηθεί, τη στιγμή t} \\ \text{t - m} & \text{διαφορετικά} \end{cases}$$

Οι Coffman et al. (1998) εργάστηκαν με έναν ορισμό του σκοπού ενός ανιχνευτή Ιστού που είναι ισοδύναμος με τον ορισμό της φρεσκάδας, αλλά με διαφορετική διατύπωση : Προτείνουν ότι ένας ανιχνευτής πρέπει να ελαχιστοποιεί το χρονικό διάστημα για το οποίο οι σελίδες παραμένουν απαρχαιωμένες. Σημείωσαν, επίσης, ότι το πρόβλημα της ανίχνευσης του Παγκόσμιου Ιστού μπορεί να μοντελοποιηθεί ως ένα πολλαπλής ουράς, μονής εξυπηρέτησης σύστημα ερωταπαντήσεων, στο οποίο ο ανιχνευτής αποτελεί τον εξυπηρετητή και οι ιστοσελίδες τις ουρές. Οι τροποποιήσεις των σελίδων αποτελούν την άφιξη των πελατών και οι χρόνοι αντικατάστασης είναι το διάστημα μεταξύ των προσβάσεων σε μία ιστοσελίδα. Στο μοντέλο αυτό, ο μέσος χρόνος αναμονής για έναν πελάτη στο σύστημα είναι ισοδύναμος με τη μέση ηλικία του ανιχνευτή.

Ο στόχος του ανιχνευτή είναι να διατηρήσει τη μέση φρεσκάδα των σελίδων στη συλλογή του όσο το δυνατόν υψηλότερη, ή να διατηρήσει τη μέση ηλικία των σελίδων όσο το δυνατό χαμηλότερη. Αυτοί οι δύο στόχοι δεν είναι ισοδύναμοι. Στην πρώτη περίπτωση, ο ανιχνευτής ενδιαφέρεται για τον αριθμό των σελίδων που έχουν απαρχαιωθεί, ενώ, στη δεύτερη περίπτωση, ο ανιχνευτής ενδιαφέρεται με το πόσο απαρχαιωμένα είναι τα τοπικά αντίγραφα των σελίδων.

Στρατηγικές

Οι Cho και Garcia – Molina (2003) μελέτησαν δύο απλές πολιτικές επανεπίσκεψης:

- Την ενιαία πολιτική που περιλαμβάνει την επανεπίσκεψη όλων των σελίδων της συλλογής του ανιχνευτή με την ίδια συχνότητα, ανεξάρτητα από τους ρυθμούς αλλαγής τους.
- Την αναλογική πολιτική που περιλαμβάνει την επανεπίσκεψη συχνότερα των σελίδων που μεταβάλλονται ταχύτερα. Η συχνότητα επίσκεψης είναι ευθέως ανάλογη της συχνότητας αλλαγής.

Και στις δύο περιπτώσεις, η επαναλαμβανόμενη σειρά ανίχνευσης των σελίδων μπορεί να γίνει είτε τυχαία είτε αυστηρώς καθορισμένα.

Κατόπιν αυτής της μελέτης, κατάφεραν να αποδείξουν το αναμενόμενο, πλην αξιοσημείωτο, αποτέλεσμα ότι, σε όρους μέσης φρεσκάδας, η ενιαία πολιτική αποδίδει καλύτερα από την αναλογική, τόσο σε πραγματικές όσο και σε συνθήκες προσομοίωσης. Η εξήγηση για το αποτέλεσμα αυτό προέρχεται από το γεγονός ότι, όταν μία σελίδα αλλάζει πολύ συχνά, ο ανιχνευτής θα ξοδέψει χρόνο προσπαθώντας να την επανανιχνεύσει πολύ γρήγορα και, ταυτόχρονα, όχι αρκετά γρήγορα για να διατηρεί το αντίγραφο φρέσκο, ενώ είναι δεδομένο ότι, καθώς ασχολείται με τις ίδιες συχνά μεταβαλλόμενες σελίδες, θα αγνοήσει ή δε θα προλάβει να ασχοληθεί με τις υπόλοιπες.

Για να βελτιώσει την φρεσκάδα, ο ανιχνευτής πρέπει να επιβάλλει ποινή στα στοιχεία εκείνα που μεταβάλλονται με μεγάλη συχνότητα. Η βέλτιστη πολιτική επανεπίσκεψης δεν είναι ούτε η ενιαία πολιτική ούτε η αναλογική πολιτική. Η βέλτιστη μέθοδος για να διατηρείται η μέση φρεσκάδα σε υψηλά επίπεδα απαιτεί από τον ανιχνευτή να αγνοεί τις σελίδες που μεταβάλλονται πολύ συχνά, ενώ η βέλτιστη μέθοδος για να διατηρείται η μέση ηλικία σε πολύ χαμηλά επίπεδα είναι να χρησιμοποιούνται συχνότητες πρόσβασης που μονοτονικά (και υπογραμμικά) αυξάνουν με το ρυθμό αλλαγής κάθε σελίδας. Και στις δύο περιπτώσεις, το βέλτιστο σημείο συγκλίνει περισσότερο στην ενιαία πολιτική. Όπως οι Coffman et al. (1998) παρατηρούν, για την ελαχιστοποίηση του αναμενόμενου χρόνου απαρχαίωσης, οι προσβάσεις σε οποιαδήποτε συγκεκριμένη σελίδα θα έπρεπε να όσο το δυνατόν πιο ισοκατανεμημένες. Σαφείς σχέσεις για την πολιτική επανεπίσκεψης δεν είναι γενικά εφικτές, αλλά ευρίσκονται με αριθμητικές μεθόδους, καθώς εξαρτώνται από την κατανομή των μεταβολών των σελίδων. Παρατηρείται ότι η εκθετική κατανομή αποτελεί ένα ταιριαστό μοντέλο για την περιγραφή των μεταβολών των σελίδων και αναπτύσσονται τρόποι αξιοποίησης των στατιστικών εργαλείων για την ανακάλυψη παραμέτρων που την επηρεάζουν (Ipeirotis et al., 2005).

Σημειώνεται ότι οι πολιτικές επανεπίσκεψης που διατυπώθηκαν αντιμετωπίζουν όλες τις σελίδες ως ποιοτικά ομοιογενείς («όλες οι σελίδες στο Διαδίκτυο είναι ίσης αξίας και ποιότητας»), υπόθεση που δεν αντιπροσωπεύει την πραγματικότητα, επομένως η γνώση της ποιότητας των σελίδων του Ιστού είναι πολύτιμη για την επίτευξη μίας καλύτερης και πιο αποδοτικής πολιτικής ανίχνευσης.

2.3.1.3 Πολιτική ευγένειας

Οι ανιχνευτές μπορούν να ανακτούν δεδομένα πολύ ταχύτερα και σε μεγαλύτερο βάθος από ότι οι άνθρωποι, επομένως μπορούν να έχουν εξοντωτικές επιπτώσεις στην επίδοση ενός ιστοτόπου. Επομένως, εάν 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. Εμπορικές μηχανές αναζήτησης, όπως οι Ask Jeeves, MSN (Bing) και Yahoo, προσφάτως έχουν τη δυνατότητα να χρησιμοποιούν μία παραπάνω παράμετρο, την «Crawl-delay», στο αρχείο robots.txt για να υποδείξουν τον αριθμό των δευτερολέπτων του διαστήματος μεταξύ αιτημάτων.

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) χρησιμοποιούν διάστημα μεταξύ προσβάσεων 10 δευτερολέπτων και ο ανιχνευτής WIRE χρησιμοποιεί 15 δευτερόλεπτα (Baeza-Yates & Castillo, 2002). Ο ανιχνευτής MercatorWeb 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. Αντίθετα, οι Dill et al. (2002) χρησιμοποιούν 1 δευτερόλεπτο. Οι τιμές του διαστήματος αυτού κυμαίνονται, συνήθως, μεταξύ 20 και 200 δευτερολέπτων.

2.3.1.4 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) μελέτησαν δύο τύπους πολιτικών παραλληλοποίησης:

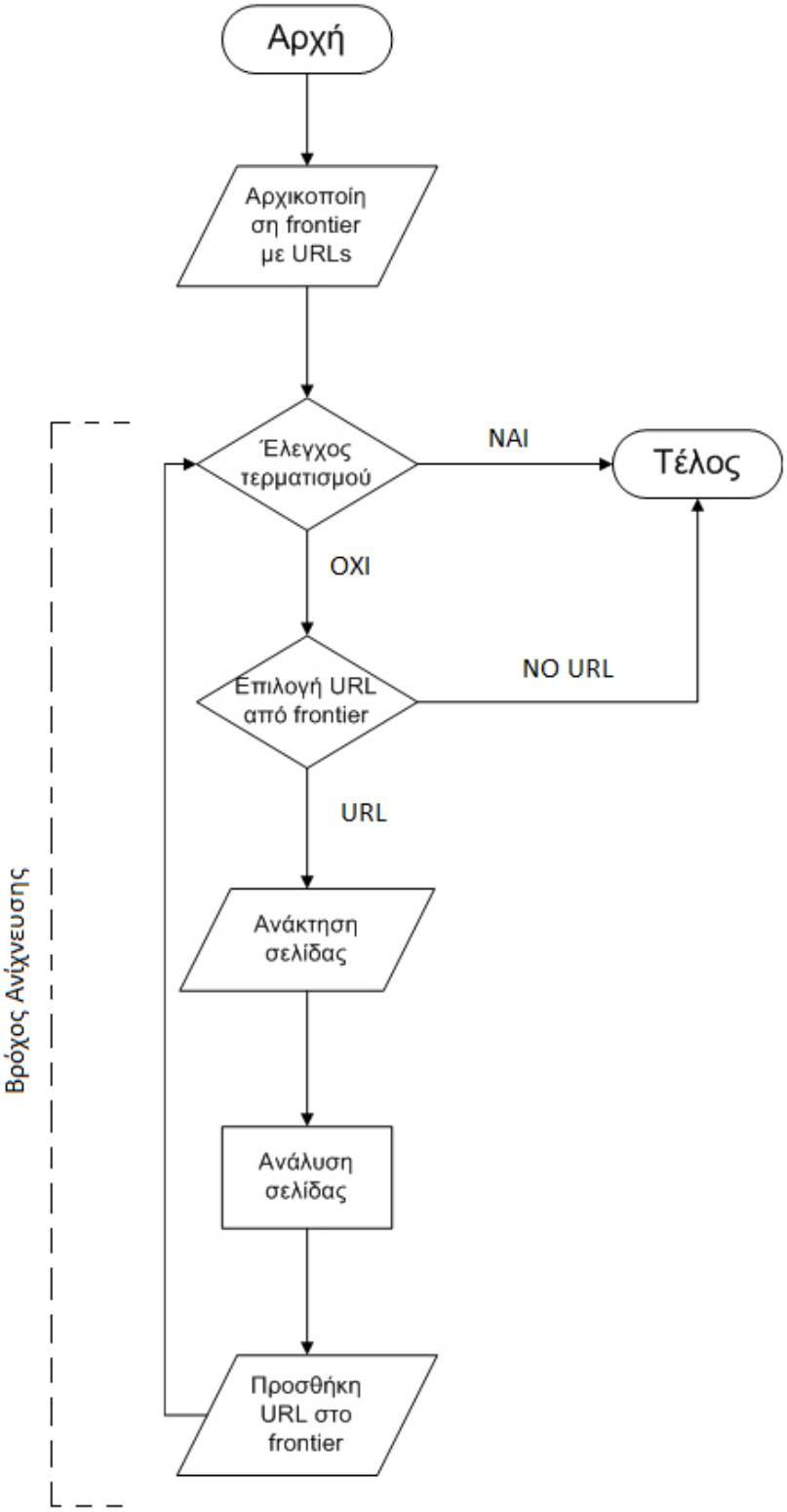
- Την πολιτική δυναμικής ανάθεσης, με την οποία ένας κεντρικός διακομιστής αναθέτει νέες URL διευθύνσεις σε διαφορετικούς ανιχνευτές δυναμικά, επιτρέποντας, έτσι, τη δυναμική εξισορρόπηση της φόρτωσης κάθε ανιχνευτή.
- Την πολιτική στατικής ανάθεσης, σύμφωνα με την οποία υπάρχει ένας αρχικός κανόνας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

2.3.2 Αρχιτεκτονικές ανίχνευσης

Ένας ανιχνευτής δεν πρέπει να περιλαμβάνει μόνο πολιτικές για την στρατηγική ανίχνευσης, αλλά να δομείται με μία βέλτιστη αρχιτεκτονική.

Οι Shkapenyuk και Suel (2002) σημειώνουν ότι ενώ είναι εξαιρετικά εύκολο να κατασκευάσεις έναν αργό ανιχνευτή που μεταφορτώνει μερικές σελίδες το δευτερόλεπτο, για ορισμένο χρονικό διάστημα, η κατασκευή ενός συστήματος υψηλών επιδόσεων που δύναται να μεταφορτώσει εκατοντάδες εκατομμύρια σελίδες, εντός ολίγων εβδομάδων, παρουσιάζει έναν σημαντικό αριθμό προκλήσεων και προβλημάτων στον συστημικό σχεδιασμό, την αποδοτικότητα δικτύου και εισόδου – εξόδου, καθώς και στην αυτοδυναμία και τη διαχείριση.

Οι ανιχνευτές Ιστού αποτελούν ένα βασικό μέρος των μηχανών αναζήτησης και λεπτομέρειες για τους αλγορίθμους και την αρχιτεκτονική διαφυλάσσονται ως επιχειρηματικά μυστικά. Όταν δημοσιεύονται τα αλγοριθμικά σχέδια ενός ανιχνευτή, αυτά παρουσιάζονται λιτά και χωρίς πολλές λεπτομέρειες για να αποτρέπεται η αναπαραγωγή του έργου. Σε καμία περίπτωση, όμως, οι μηχανές αναζήτησης δε διαφωτίζουν τα σαφή κριτήρια και τους αλγορίθμους κατάταξης λεπτομερώς.



Εικόνα 3 Γενικό διάγραμμα ροής βασικού διαδοχικού ανιχνευτή

2.3.3 Βασικοί αλγόριθμοι ανίχνευσης

Οι βασικοί αλγόριθμοι ανίχνευσης που παρουσιάζονται αποτελούν παραλλαγές του best – first σχεδίου (αναζήτηση πρώτα στο καλύτερο). Η βασική διαφορά βρίσκεται στην ευρετική 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.

«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 τιμής αυτής ομοιότητας. Οι URLs, στη συνέχεια, προστίθενται σε ένα crawl frontier με σειρά προτεραιότητας, βάσει αυτών των τιμών. 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, $v_q * v_p$ είνα το εσωτερικό γινόμενο των δύο διανυσμάτων, ενώ $\| v_q \|$ και $\| v_p \|$ είναι οι Ευκλείδειες νόρμες των δύο διανυσμάτων. Πλέον εξελιγμένες διανυσματικές αναπαραστάσεις των σελίδων δεν χρησιμοποιούνται σε εφαρμογές ανίχνευσης, καθώς δεν υπάρχει a priori γνώση της κατανομής των όρων μέσα στις μη επισκεφθείσες σελίδες. Στην περίπτωση ανιχνευτή πολλαπλών νημάτων, ο ανιχνευτής συμπεριφέρεται ως μηχανισμός ανίχνευσης πρώτα στα N καλύτερα, όπου το N είναι συνάρτηση του αριθμού των συγχρόνως απασχολούμενων νημάτων. Έτσι, η πρώτα στα N καλύτερα ανίχνευση αποτελεί μία γενικευμένη εκδοχή της πρώτα στο καλύτερο ανίχνευσης που επιλέγει N καλύτερες διευθύνσεις URL για ταυτόχρονη ανίχνευση. Σύμφωνα με τους G. Pant, P. Srinivasan, F. Menczer, ο αλγόριθμος ανίχνευσης πρώτα στα N καλύτερα (με N = 256) βρίσκεται ανάμεσα στους καλύτερους, με εμφανή υπεροχή στην ανάκτηση ομοίων σελίδων, ενώ μπορεί να διατηρήσει το μέγεθος του crawl frontier εντός του άνω ορίου διατηρώντας μόνο τις X καλύτερες τοποθεσίες URL.

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. Το anchor text και το διαπερνόν αποτέλεσμα της προηγούμενης σελίδας (που παρείχε σύνδεσμο προς την εξεταζόμενη τοποθεσία URL) επηρεάζουν το δυνητικό αποτέλεσμα των συνδέσμων. Ο SharkSearch δεν εφαρμόζει απόλυτα κατά βάθος διάσχιση, αλλά διατηρεί όριο βάθους, με αποτέλεσμα, εάν ο ανιχνευτής φθάνει σε ασήμαντες σελίδες σε ένα υπό ανίχνευση μονοπάτι, να 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:

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

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

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

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

Το neighborhood score χρησιμοποιεί το anchor text και το κείμενο πλησίον του anchor text, σε μία προσπάθεια να τελειοποιηθεί το συνολικό αποτέλεσμα της URL, για λόγους διαφοροποίησης μεταξύ των διαφόρων συνδέσμων που υπάρχουν στην ίδια σελίδα. Για το σκοπό αυτό, ο ανιχνευτής SharkSearch αναθέτει ένα anchor score και ένα context score σε κάθε URL. Το anchor score αποτελεί απλώς την ομοιότητα του anchor text του υπερσυνδέσμου που περιλαμβάνει τη διεύθυνση URL με το ερώτημα q, π.χ. sim(q, anchor_text). Το context score (αποτέλεσμα πλαισίου), από την άλλη, διευρύνει το πλαίσιο του συνδέσμου, περιλαμβάνοντας ορισμένες γειτονικές λέξεις. Το συνολικό πλαίσιο που προκύπτει, το aug_context, χρησιμοποιείται για τον υπολογισμό του context score ως εξής:

$$context(url) = \begin{cases} 1 & \text{εάν anchor(url)>0} \\ \text{sim(q,aug\_context)} & \text{αλλιώς} \end{cases}$$

Τελικά, υπολογίζεται το neighborhood score από τα anchor και context scores, από την σχέση

$$neighborhood(url) = \beta * anchor(url) + (1-\beta) * context(url)$$

όπου β<1 είναι μία ακόμη παράμετρος.

Σημειώνεται ότι η εφαρμογή του SharkSearch αλγορίθμου επιβάλλει την ανάθεση τιμών στις τέσσερις παραμέτρους d, γ, δ και β.

Προτάθηκε ως βελτιωμένη έκδοση του προκατόχου αλγόριθμου Fish Search, ξεπερνώντας κάποιους περιορισμούς του πρώτου και παρέχοντας μία σχετικά καλή εκτίμηση της ομοιότητας των γειτονικών σελίδων προτού αυτές αναλυθούν. Παρ' όλα αυτά, προϋποθέτει τη σωστή ανάθεση τιμών στις παραμέτρους που χρησιμοποιούνται.

Αλγόριθμος επικεντρωμένου ανιχνευτή

Όπως έχει ήδη αναφερθεί, ο αλγόριθμος της επικεντρωμένης ανίχνευσης βασίζεται στην a priori γνώση της θεματικής ομοιότητας που συνδέει δύο σελίδες. Η βασική ιδέα ενός τέτοιου ανιχνευτή ήταν να ταξινομεί σελίδες που έχουν ανιχνευθεί σε μία 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). Τα παραδείγματα αυτά καταχωρούνται αυτόματα σε διάφορες κατηγορίες της ταξινομίας. Μέσα από μία διαδραστική διαδικασία, ο χρήστης μπορεί να διορθώσει την αυτόματη ταξινόμηση, να προσθέσει νέες κατηγορίες στην ταξινομία 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 $\mathrm{P}(c \mid p)$ να ανήκει μία σελίδα p που έχει ανιχνευθεί σε μία κατηγορία c στον κατάλογο. Εξ ορισμού είναι $\mathrm{P}(r \mid p) = 1$, όπου r είναι η αρχική κατηγορία της ταξινομίας (root category). Ένας βαθμός συσχέτισης συνοδεύει κάθε σελίδα που υπολογίζεται από τη σχέση:

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

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. Στην «αυστηρώς» επικεντρωμένη λειτουργία, για μία σελίδα p που έχει ήδη ανιχνευθεί, ο ταξινομητής πρώτα ευρίσκει τον κόμβο c* (στην ταξινομία) με τη μέγιστη πιθανότητα να εμπεριέχει την p. Εάν οποιοδήποτε σημείο προ του κόμβου c* έχει αξιολογηθεί ως «good»

από το χρήστη, τότε οι διευθύνσεις, οι οποίες εξάγονται από τη σελίδα p, προστίθενται στο frontier.

Ένα ακόμη ενδιαφέρον στοιχείο της επικεντρωμένης ανίχνευσης αποτελεί η χρήση μίας μεθόδους απόσταξης. Η απόσταξη εφαρμόζει μία τροποποιημένη εκδοχή του αλγορίθμου του Kleinberg (1999) για να βρίσκει τοπικούς κόμβους. Οι κόμβοι παρέχουν συνδέσμους σε πολλές έγκυρες πηγές σχετικές με το θέμα. Η απόσταξη ενεργοποιείται σε διάφορες στιγμές, κατά τη διάρκεια της ανίχνευσης, και ορισμένοι από τους κορυφαίους κόμβους προστίθενται στο crawl frontier.

Αλγόριθμος InfoSpiders

Στον αλγόριθμο InfoSpiders των Menczer, Pant και Srinivasan (2004), 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. Στον αρχικό αλγόριθμο, κάθε πράκτορας διατηρούσε το frontier του περιορισμένο στους συνδέσμους της σελίδας που ανακτήθηκε τελευταία από τον πράκτορα. Εξαιτίας αυτής της προσέγγισης που βασιζόταν σε περιορισμένη μνήμη, ο ανιχνευτής περιοριζόταν στην ανίχνευση των συνδέσμων της πιο πρόσφατης σελίδας, με αποτέλεσμα να υπερέχει αυτού η «αφελής» πρώτα στο καλύτερο ανίχνευση. Έκτοτε, μία σειρά από βελτιώσεις στον αρχικό αλγόριθμο έχουν σχεδιασθεί. Στην πραγματικότητα, η επανασχεδιασμένη έκδοση του αλγορίθμου έχει αποδειχθεί πως υπερέχει πολλών εκδοχών της «αφελούς» πρώτα στο καλύτερο ανίχνευσης σε συγκεκριμένες εργασίες ανίχνευσης, σε δείγμα ανιχνεύσεων περισσότερων από 10,000 σελίδων.

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

$$P(\lambda) = \frac{e^{\beta\sigma(\lambda)}}{\sum_{\lambda' \in \varphi} e^{\beta\sigma(\lambda')}}$$

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. (Pant, Srinivasan & Menczer, 2004)

2.4 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) αποτελεί μία διεπιστημονική διαδικασία, ενσωματώνοντας 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. Οι ιδιαίτερα μεγάλες υπηρεσίες εκτελούν τη διαδικασία της ευρετηρίασης σε τακτά

προκαθορισμένα διαστήματα, εξαιτίας του απαιτούμενου χρόνου περάτωσης και του λειτουργικού κόστους, ενώ άλλες, που βασίζονται σε ευφυείς πράκτορες, ευρετηριάζουν σε πραγματικό χρόνο.

Η διαδικασία συνοψίζεται στα επιμέρους τμήματα της κατασκευής του ευρετηρίου και της ανάλυσης.

2.4.1 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.

2.4.1.1 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). Η ταχύτητα εύρεσης μίας καταχώρησης σε μία δομή, συγκριτικά με το πόσο γρήγορα μπορεί να ενημερώνεται ή να αφαιρείται, αποτελεί ένα τετριμμένο θέμα της επιστήμης των υπολογιστών.

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.

2.4.1.2 Δομή Ευρετηρίου

Οι αρχιτεκτονικές των μηχανών αναζήτησης ποικίλλουν στον τρόπο με τον οποίο η ευρετηρίαση πραγματοποιείται και στις μεθόδους αποθήκευσης για να ανταποκρίνονται σε διάφορους παράγοντες σχεδίασης. Τύποι δεικτών περιλαμβάνουν:

Δέντρο καταλήξεων

Μεταφορικά σχεδιασμένο σα δέντρο, υποστηρίζει γραμμική χρονική αναζήτηση. Το δέντρο κατασκευάζεται με την ταξινόμηση των καταλήξεων των λέξεων και ανήκει στην κατηγορία των δέντρων που εμπεριέχουν συμβολοσειρές, τα οποία υποστηρίζουν εκτεταμένη διαχείριση πολύπλοκων εγγράφων που είναι σημαντική για την ευρετηρίαση των μηχανών αναζήτησης. Ένα σημαντικό μειονέκτημα είναι ότι η αποθήκευση μίας λέξης σε ένα δέντρο ενδέχεται να απαιτεί περισσότερο χώρο από την αποθήκευση της ίδιας της λέξης. Μία εναλλακτική αναπαράσταση είναι ο πίνακας καταλήξεων, ο οποίος θεωρείται ότι απαιτεί λιγότερη εικονική μνήμη και υποστηρίζει συμπίεση δεδομένων.

Ανεστραμμένο ευρετήριο

Όπως αναφέρθηκε και παραπάνω, η δομή αυτή αποθηκεύει μία λίστα συμβάντων (εμφανίσεων) κάθε ατομικού κριτηρίου αναζήτησης, συνήθως στη μορφή ενός δυαδικού δέντρου.

Για παράδειγμα, έστω ότι έχουμε τα έγγραφα $T_0 = "\text{είναι αυτό που είναι.}"$, $T_1 = "\text{τι είναι αυτό;}"$ και $T_2 = "\text{αυτό είναι ένα ευρετήριο}"$, το ανεστραμμένο ευρετήριο είναι το εξής:

$$"\text{ένα}": \{2\}$$
$$"\text{ευρετήριο}": \{2\}$$
$$"\text{είναι}": \{0,\ 1,\ 2\}$$
$$"\text{αυτό}": \{0,\ 1,\ 2\}$$
$$"\text{τι}": \{1\}$$
$$"\text{που}": \{0\}$$

Μία αναζήτηση, λοιπόν, της φράσης «ένα ευρετήριο» θα επέστρεφε, ως αποτέλεσμα, το έγγραφο $\{2\} \cap \{2\} = \{2\}$, ενώ μία αναζήτηση της φράσης «τι είναι αυτό» θα επέστρεφε το έγγραφο $\{1\} \cap \{0,1,2\} \cap \{0,1,2\} = \{1\}$, εφόσον, φυσικά, η αντιστοιχία γινόταν απόλυτα, κάτι που δε θα βοηθούσε στην αναζήτηση «τι είναι ευρετήριο» που δε θα επέστρεφε κανένα έγγραφο $(\{1\} \cap \{0,1,2\} \cap \{2\} = \{\varnothing\})$.

Ευρετήριο παραπομπών

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

Ευρετήριο Ngram

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

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:

$T_0 = "\text{μου αρέσουν πολύ τα πράσινα μήλα}"$
$T_1 = "\text{υπάρχουν τα κόκκινα μήλα και τα πράσινα μήλα}"$

Ο πίνακας που προκύπτει είναι ο εξής:

	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

Πίνακας 1 Παράδειγμα ευρετηρίου όρων - εγγράφων

Παραλληλισμός

Μία βασική πρόκληση στη σχεδίαση των μηχανών αναζήτησης, που σχετίζεται αποκλειστικά με το ευρετήριο αυτών, αποτελεί η διαχείριση σειριακών υπολογιστικών διαδικασιών. Ελλοχεύουν πολλοί κίνδυνοι από υποθέσεις, η έκβαση (το υπολογιστικό αποτέλεσμα) των οποίων εξαρτάται άμεσα από την περάτωση ταυτόχρονων διαδικασιών, και λάθη συνοχής. Για παράδειγμα, ένα νέο έγγραφο προστίθεται στον κώδικα και το ευρετήριο οφείλει να ενημερωθεί, αλλά, την ίδια στιγμή, το ευρετήριο πρέπει να συνεχίσει να αποκρίνεται σε ερωτήματα αναζήτησης. Αυτό αποτελεί σύγκρουση μεταξύ δύο ανταγωνιστικών διαδικασιών. Εάν οι συγγραφείς κειμένων ή εγγράφων θεωρηθούν παραγωγοί πληροφοριών και ένας ανιχνευτής Ιστού θεωρηθεί ο καταναλωτής της πληροφορίας που πρέπει να λάβει το κείμενο και να το αποθηκεύσει σε μία κρυφή μνήμη (ή σε κώδικα), το ευθύ ευρετήριο αποτελεί τον καταναλωτή της πληροφορίας που προέρχεται από τον κώδικα και το ανεστραμμένο ευρετήριο αποτελεί τον καταναλωτή της πληροφορίας που παράγεται από το ευθύ ευρετήριο. Αυτό ευρέως αναφέρεται ως μοντέλο παραγωγού – καταναλωτή. Ο indexer είναι ο παραγωγός της ερευνήσιμης πληροφορίας και οι χρήστες είναι οι καταναλωτές που χρειάζονται να αναζητήσουν. Η πρόκληση είναι ακόμη μεγαλύτερη, όταν πρόκειται για διανεμημένη αποθήκη κι επεξεργασία διαδικασιών. Γι' αυτό το λόγο, η αρχιτεκτονική της μηχανής αναζήτησης ενδέχεται να περιλαμβάνει τεχνικές διανεμημένης πληροφορικής, όπου η μηχανή περιλαμβάνεται από πολλές συνεργαζόμενες μηχανές που συναποτελούν μία μονάδα. Έτσι αυξάνονται οι πιθανότητες για σύγχυση και μη συνοχή των διαδικασιών, ενώ καθίσταται δυσκολότερο να διατηρηθεί μία πλήρως συγχρονισμένη, διανεμημένη, παράλληλη αρχιτεκτονική.

Ανεστραμμένοι δείκτες

Πολλές μηχανές αναζήτησης επιστρατεύουν τη δομή του ανεστραμμένου ευρετηρίου καθώς αξιολογούν ένα ερώτημα αναζήτησης, με σκοπό τον ταχύτερο προσδιορισμό εγγράφων που περιλαμβάνουν τις λέξεις του ερωτήματος και την ταξινόμηση των εγγράφων αυτών, με

κριτήριο, προφανώς, το βαθμό συσχέτισης. Επειδή το ανεστραμμένο ευρετήριο αποθηκεύει μία λίστα των εγγράφων που περιλαμβάνουν κάθε λέξη, η μηχανή αναζήτησης μπορεί να έχει άμεση πρόσβαση για την εύρεση των εγγράφων που σχετίζονται με κάθε λέξη του ερωτήματος, με στόχο την ανάκτηση των αντίστοιχων εγγράφων γρήγορα. Μία απλοποιημένη αναπαράσταση ενός ανεστραμμένου ευρετηρίου αναφέρθηκε προηγουμένως.

Το ευρετήριο μπορεί μόνο να αποφασίσει εάν μία λέξη υπάρχει μέσα σε ένα συγκεκριμένο έγγραφο, εφόσον δεν αποθηκεύει άλλες χρήσιμες πληροφορίες, σχετικές με τη συχνότητα ή τη θέση της λέξης μέσα στο έγγραφο. Επομένως, αποτελεί ένα ευρετήριο Boole (δυαδικό). Αυτή η ταυτοποίηση εγγράφων και αντιστοίχισή των με τους όρους αναζήτησης αποφαίνεται ποιά έγγραφα αντιστοιχίζονται σε ένα ερώτημα, αλλά δε δύναται να ταξινομήσει τα έγγραφα αυτά. Σε κάποιους σχεδιασμούς, το ευρετήριο περιλαμβάνει επιπρόσθετες πληροφορίες, όπως η συχνότητα κάθε λέξης σε κάθε έγγραφο ή οι θέσεις κάθε λέξης σε κάθε έγγραφο. Πληροφορίες για τη θέση δίνουν τη δυνατότητα στον αλγόριθμο αναζήτησης να προσδιορίζει την εγγύτητα των λέξεων για την υποστήριξη αναζητήσεων με όρους – φράσεις, ενώ η συχνότητα μπορεί να χρησιμοποιηθεί για την υποστήριξη της ταξινόμησης των εγγράφων, με κριτήριο το βαθμό συσχέτισης με το ερώτημα.

Συγχώνευση ευρετηρίων

Το ανεστραμμένο ευρετήριο συμπληρώνεται μέσω μίας συγχώνευσης ή ανακατασκευής. Η ανακατασκευή είναι μία διαδικασία παρόμοια της συγχώνευσης, αλλά πρώτα διαγράφει τα περιεχόμενα του ανεστραμμένου ευρετηρίου. Η αρχιτεκτονική μπορεί να σχεδιαστεί να υποστηρίζει στοιχειώδη ευρετηρίαση, όπου μία συγχώνευση θα ταυτοποιεί τα έγγραφα που πρόκειται να προστεθούν ή ενημερωθούν κι έπειτα θα αναλύει κάθε έγγραφο σε λέξεις. Για λόγους τεχνικής ακρίβειας, μία συγχώνευση εξομοιώνει προσφάτως ευρετηριασμένα έγγραφα, διαμένοντας σε εικονική μνήμη, με την κρυφή μνήμη του ευρετηρίου να εδρεύει σε έναν ή περισσότερους σκληρούς δίσκους.

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:

$$T_0 : \{\text{"είναι"}, \text{"αυτό"}, \text{"που"}\}$$
$$T_1 : \{\text{"είναι"}, \text{"αυτό"}, \text{"τι"}\}$$
$$T_2 : \{\text{"ένα"}, \text{"ευρετήριο"}, \text{"είναι"}, \text{"αυτό"}\}$$

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.

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.

2.4.2 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.

2.4.2.1 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 υπακούν 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.

2.4.2.2 Διαχωρισμός λέξεων ή ενδείξεων

Σε αντίθεση με τον άνθρωπο, οι υπολογιστές δεν καταλαβαίνουν τη δομή ενός κειμένου γραμμένου σε φυσική γλώσσα και δε μπορούν αυτόματα να αναγνωρίσουν λέξεις ή προτάσεις. Συγκεκριμένα, κάθε έγγραφο σημαίνει μόνο μία ακολουθία από bytes. Οι υπολογιστές δε 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.

2.4.2.3 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.

2.4.2.4 Ανάλυση τύπου αρχείων

Εάν η μηχανή αναζήτησης υποστηρίζει πολλαπλούς τύπους εγγράφων, τα έγγραφα πρέπει να προετοιμάζονται για τη διαδικασία του διαχωρισμού και της ανάλυσης, ως συνόλου. Η πρόκληση έγκειται στο γεγονός ότι πολλοί τύποι εγγράφων περιλαμβάνουν διαμόρφωση πληροφοριών, πέραν των κειμενικών περιεχομένων. Για παράδειγμα, τα έγγραφα HTML περιλαμβάνουν HTML ετικέτες, οι οποίες προσδιορίζουν τη διαμόρφωση πληροφοριών, όπως, για παράδειγμα, την έναρξη νέας γραμμής, την έμφαση κειμένου (bold) και το μέγεθος της γραμματοσειράς. Εάν η μηχανή αναζήτησης αγνοούσε τη διαφορά μεταξύ περιεχομένου και σήμανσης, εξωτερικές πληροφορίες θα εισέρχονταν στο ευρετήριο, με αποτέλεσμα τη λανθασμένη επιστροφή αποτελεσμάτων σε σχετικές αναζητήσεις. Η ανάλυση του τύπου αρχείων αφορά στην αναγνώριση και το χειρισμό της διαμόρφωσης του περιεχομένου που είναι ενσωματωμένο μέσα σε έγγραφα, ενώ ελέγχει τον τρόπο με τον οποίο το έγγραφο αποτυπώνεται στην οθόνη του υπολογιστή ή ερμηνεύεται από ένα πρόγραμμα ή λογισμικό. 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).

Ορισμένες, μάλιστα, μηχανές αναζήτησης υποστηρίζουν την ανίχνευση αρχείων που έχουν αποθηκευτεί σε συμπιεσμένη ή κωδικοποιημένη μορφή. Ο μηχανισμός ευρετηρίασης οφείλει, σε αυτήν την περίπτωση, πρώτα να αποσυμπιέσει το έγγραφο, μια διαδικασία από την οποία ενδέχεται να εξαχθούν περισσότερα από ένα αρχεία, και στη συνέχεια να εξετάσει και να ευρετηριάσει κάθε εξαγμένο αρχείο ξεχωριστά. Συμπιεσμένοι τύποι αρχείων που υποστηρίζονται ευρέως περιλαμβάνουν αρχεία ZIP, RAR, CAB (Microsoft Windows), GZIP, BZIP, TAR και παράγωγα της συμπίεσης κατά Unix λειτουργικά συστήματα.

Η ανάλυση του τύπου αρχείου ενδέχεται να περιλαμβάνει μεθόδους ποιοτικής βελτίωσης για την αποφυγή περιπτώσεων όπου συμπεριλαμβάνονται «κακά» δεδομένα στο ευρετήριο της μηχανής. Το περιεχόμενο, πάντως, ενδέχεται να καταχράται τις πληροφορίες διαμόρφωσης για να συμπεριλάβει επιπρόσθετα δεδομένα. Τέτοια παραδείγματα υπάγονται στην κατηγορία του Black Hat SEO και είναι προς αποφυγίν, καθώς δεν προσφέρουν τίποτα σε όρους βελτιστοποίησης, ενώ θα αναφερθούν λεπτομερώς σε αντίστοιχο κεφάλαιο.

2.4.2.5 Ευρετηρίαση META ετικετών

Ορισμένα έγγραφα συχνά περιλαμβάνουν ενσωματωμένες meta πληροφορίες, όπως «συγγραφέας», «λέξεις – κλειδιά», «περιγραφή» και «γλώσσα». Όσον αφορά τις HTML σελίδες, η META ετικέτα περιλαμβάνει λέξεις κλειδιά που συμπεριλαμβάνονται στο ευρετήριο, όπως το περιεχόμενο της σελίδας. Η πρώιμη τεχνολογία των διαδικτυακών μηχανών αναζήτησης ευρετηρίαζε μόνο τις λέξεις – κλειδιά (META keywords) των μεταετικετών για το ευθύ ευρετήριο, ενώ το σύνολο του εγγράφου δεν αναλυόταν. Την εποχή εκείνη, η ευρετηρίαση πλήρους κειμένου δεν είχε καθιερωθεί, ούτε το υλικό (hardware) ήταν δυνατό να υποστηρίξει τέτοιες τεχνολογίες. Ο σχεδιασμός της γλώσσας HTML αρχικά περιελάμβανε υποστήριξη για μεταετικέτες, με σκοπό να ευρετηριάζονται εύκολα και σωστά, χωρίς τη διαδικασία του tokenization ως προϋπόθεση.

Καθώς το Διαδίκτυο μεγάλωνε τη δεκαετία του 1990, πολλοί οργανισμοί κι επιχειρήσεις με φυσική παρουσία συνδέθηκαν σε αυτό και εγκαθίδρυσαν εταιρικούς ιστοτόπους. Οι λέξεις κλειδιά που χρησιμοποιήθηκαν για να περιγράψουν ιστοσελίδες άλλαξαν από περιγραφικές σε όρους marketing, σχεδιασμένους να οδηγούν πωλήσεις, επιτυγχάνοντας υψηλές θέσεις σε σχετικές αναζητήσεις. Η τάση αυτή οδήγησε στην κατάχρηση των πληροφοριών που ήδη αναφέρθηκε, με αποτέλεσμα να αναπτυχθούν και να υιοθετηθούν τεχνολογίες ευρετηρίασης πλήρους κειμένου. Οι εταιρείες μπορούσαν να «γεμίσουν» τη σελίδα τους με περιορισμένο αριθμό προωθητικών λέξεων – κλειδιών, προκειμένου να μην αφαιρεθεί το ενδιαφέρον και χρήσιμο περιεχόμενο. Δεδομένης της σύγκρουσης αυτής ενδιαφέροντος με τον επιχειρηματικό στόχο του σχεδιασμού προσανατολισμένων γύρω από τον χρήστη ιστοσελίδων, η ευρετηρίαση πλήρους κειμένου ήταν περισσότερο αντικειμενική και αύξησε την ποιότητα των αποτελεσμάτων των μηχανών αναζήτησης, καθώς ήταν ένα βήμα παραπέρα από τον υποκειμενικό έλεγχο της θέσης των αποτελεσμάτων.

2.5 Επεξεργασία ερωτημάτων

Υποθέτουμε ότι διαθέτουμε μία συλλογή από έγγραφα $D = \{d_0, d_1, d_2, ..., d_{n-1}\}$ που περιέχονται σε n ιστοσελίδες, οι οποίες έχουν ήδη ανιχνευθεί και είναι διαθέσιμες στο ευρετήριο. Έστω ότι $W = \{w_0, w_1, w_2, ..., w_{m-1}\}$ είναι οι m διαφορετικές λέξεις που εμφανίζονται οπουδήποτε στη συλλογή. Τυπικά, προφανώς 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.

Όπως έχει ήδη αναφερθεί, το ανεστραμμένο ευρετήριο I για τη συλλογή αποτελείται από ένα σύνολο ανεστραμμένων λιστών $I_{w0}, I_{w1}, I_{w2}, ..., I_{wm-1}$, όπου η λίστα I_w 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 $q = \{t_0, t_1, t_2, ..., t_{d-1}\}$ 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 $q = \{t_0, t_1, t_2, ..., t_{d-1}\}$, αναθέτει σε κάθε έγγραφο D, ένα score $F(D, q)$. 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.

2.5.1 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:

Βασικοί Τελεστές		
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 από τους όρους του ερωτήματος.	σταλακτίτες |σταλαγμίτες
""	Η φράση που περικλείεται από τα εισαγωγικά θα αναζητηθεί επακριβώς στο ευρετήριο.	"μηχανές αναζήτησης"
-	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.	"Εθνικό * Πολυτεχνείο"
~	Η μηχανή θα επιστρέψει τις σελίδες που περιλαμβάνουν τον όρο και όλα τα συνώνυμά του.	~job
..	Ο τελεστής αυτός δηλώνει το εύρος μεταξύ δύο αριθμών, περιορίζοντας έτσι τα αποτελέσματα της αναζήτησης.	"έκλειψη ηλίου 2000..2010"

Πίνακας 2 Οι βασικοί τελεστές αναζήτησης

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. Τέτοιοι τελεστές χρησιμοποιούνται καθημερινά στη βιομηχανία της Οργανικής Βελτιστοποίησης των Ιστοσελίδων, με στόχο τον έλεγχο της επίδοσης ενός project, ή τον έλεγχο του ανταγωνισμού, και είναι οι εξής:

Τελεστές εναλλακτικής αναζήτησης		
Τελεστής	**Περιγραφή**	**Παράδειγμα**
site:	Η μηχανή αναζητά μόνο στον συγκεκριμένο ιστότοπο.	"Diploma Thesis" site:ntua.gr
cache:	Η μηχανή αναζητά αποθηκευμένα "στιγμιότυπα" μίας ιστοσελίδας στη μνήμη της.	cache:ieee.ntua.gr
inurl:	Η μηχανή αναζητά τους όρους του ερωτήματος εντός της διεύθυνσης URL.	inurl:ntua
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:imu.ntua.gr
inanchor:	Η μηχανή αναζητά τους όρους του ερωτήματος εντός των anchor texts που περιλαμβάνουν τα έγγραφα.	inanchor:"οδηγός αγοράς ηλιούπολης"
link:	Η αναζήτηση επιστρέφει τις σελίδες εκείνες που συνδέουν στον συγκεκριμένο ιστότοπο. Παρέχει ένα κατώτατο όριο των πραγματικών συνδέσμων προς τον ιστότοπο.	link:imu.ntua.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:dmoz.org
filetype:	Η αναζήτηση πραγματοποιείται μόνο σε συγκεκριμένους τύπους αρχείων.	"crawling and indexing" filetype:pdf

Πίνακας 3 Τελεστές προχωρημένης αναζήτησης

Τέλος, οι σπουδαιότερες μηχανές αναζήτησης παρέχουν επιπρόσθετες δυνατότητες αναζήτησης, καθώς και δυνατότητες λειτουργίας ως αριθμομηχανών. Οι αριθμητικοί και λοιποί συνήθεις τελεστές είναι οι εξής:

Αριθμητικοί Τελεστές		
Τελεστής	**Περιγραφή**	**Παράδειγμα**
+, -, *, /	Πρόσθεση, αφαίρεση, πολλαπλασιασμός, διαίρεση αριθμών.	5 + 12,4 * (0,2/21)
^ ή **	Ύψωση αριθμού σε δύναμη	5 ^ (32/3)
sqrt()	Τετραγωνική ρίζα αριθμού	sqrt(162)
sin(), cos(), tan()	Τριγωνομετρικοί αριθμοί γωνίας	sin(35), cos(5*pi)
ln()	Λογάριθμος με βάση το e	ln(5)
log()	Λογάριθμος με βάση το 10	log(5)
lg()	Λογάριθμος με βάση το 2	lg(5)
!	Παραγοντικό αριθμού	!5
% of	Ποσοστό επί τοις εκατό αριθμού	23,5% of 201
in	Μετατροπή συναλλάγματος, φυσικών μεγεθών.	14 euro in USD 13,5kg in lb

Πίνακας 4 Τελεστές αριθμητικών πράξεων και υπολογισμών

Λοιποί Τελεστές		
Τελεστής	**Περιγραφή**	**Παράδειγμα**
define:	Η μηχανή επιστρέφει τον ορισμό του όρου αναζήτησης.	define:algorithm
weather	Η μηχανή επιστρέφει πρόγνωση καιρού για τη συγκεκριμένη πόλη	weather Athens
time	Η μηχανή παρουσιάζει την ώρα στην συγκεκριμένη πόλη	time New York

Πίνακας 5 Τελεστές ορισμών, καιρικών προγνώσεων και ώρας

Αξίζει να σημειωθεί ότι η μηχανή της Google, προς το παρόν, έχει κάνει ορισμένα βήματα προόδου προς το σημασιολογικό ιστό, αναγνωρίζοντας και απαντώντας σε ορισμένα ερωτήματα αναζήτησης, εκφρασμένα στη μορφή ερώτησης.

3 *Προσδιορισμός παραγόντων κατάταξης στις μηχανές αναζήτησης*

3.1 Οι αρχικοί παράγοντες κατάταξης

Στην αρχική έρευνα που πραγματοποίησαν, οι ιδρυτές της μηχανής αναζήτησης Google, Lawrence Page και Sergey Brin, περιγράφουν τον τρόπο με τον οποίο η Google ταξινομεί τα αποτελέσματα αναζήτησης. Συγκεκριμένα, δημιουργεί μία βάση δεδομένων που περιλαμβάνει τη θέση, τη γραμματοσειρά, τη διάκριση των χαρακτήρων σε πεζά και κεφαλαία, την ύπαρξη στο anchor text των λέξεων που υπάρχουν από κοινού στο ερώτημα αναζήτησης και το κάθε έγγραφο (παράγοντες εξαρτημένοι από τον όρο αναζήτησης), καθώς και το βαθμό PageRank του ίδιου του εγγράφου (παράγοντας ανεξάρτητος από τον όρο αναζήτησης). Η συνάρτηση κατάταξης των αποτελεσμάτων που σχεδιάσθηκε δεν έδινε ιδιαίτερη βαρύτητα σε έναν από τους παραπάνω παράγοντες έναντι των υπολοίπων.

Στην περίπτωση μίας λέξης αναζήτησης, η Google καταμετρούσε τις αναφορές της λέξης στο έγγραφο και, με κριτήριο τη θέση, συμπλήρωνε την παραπάνω βάση δεδομένων, κατηγοριοποιώντας τις αναφορές αυτές στις διαφορετικές κατηγορίες – παράγοντες, κάθε μία από τις οποίες είχε τη δική της βαρύτητα. Κάθε αναφορά μετατρεπόταν, έτσι, σε έναν αριθμό ίσο με το βάρος του αντίστοιχου τύπου, ενώ κάθε προσαύξηση κατά μία φορά του αριθμού κάθε τύπου γινόταν γραμμικά μέχρις ενός συγκεκριμένου σημείου, μετά το οποίο οι αναφορές έπαυαν να συμβάλλουν. Ο βαθμός που συγκεντρωνόταν από τις αναφορές της λέξης – κλειδιού συνδυαζόταν με το βαθμό PageRank για να αποδώσουν έναν τελικό βαθμό κατάταξης του εγγράφου στα αποτελέσματα της Google.

Στην περίπτωση μίας φράσης αναζήτησης, αποτελούμενης από πολλαπλές λέξεις, η διαδικασία αυτή επαναλαμβανόταν για κάθε λέξη της φράσης και δινόταν προτεραιότητα στις αναφορές των διαφορετικών λέξεων που γίνονταν γειτονικά, εν αντιθέσει με τις λέξεις που βρίσκονταν μακριά η μία από την άλλη μέσα στο έγγραφο. Καθώς γινόταν σύγκριση των διαφόρων βάσεων για κάθε μία λέξη, ώστε να ευρεθούν οι γειτονικές λέξεις που συνιστούσαν αυτούσια ή μέρος της φράσης εντός του εγγράφου, αποδίδονταν 10 διαφορετικοί χαρακτηρισμοί της απόστασης μεταξύ των λέξεων, σε μία κλίμακα από «τέλειο ταίρι» έως «καμία σχέση». Έτσι, ο βαθμός που συγκεντρωνόταν από τις αναφορές των λέξεων για την φράση – κλειδί ήταν αποτέλεσμα τόσο του βάρους των αναφορών όσο και της απόστασης μεταξύ των λέξεων. Εν συνεχεία, όπως πριν, ο βαθμός αυτός και ο βαθμός PageRank συνδυάζονταν για την απόδοση του τελικού βαθμού κατάταξης του εγγράφου για τη συγκεκριμένη αναζήτηση (Brin & Page, 1998).

Είναι δεδομένο ότι, παρότι οι συγκεκριμένοι παράγοντες που επηρεάζουν τη θέση των ιστοσελίδων στα αποτελέσματα αναζήτησης εξακολουθούν να έχουν βαρύτητα στους αλγορίθμους των μηχανών αναζήτησης, οι συνθήκες και οι χρησιμοποιούμενες τεχνολογίες, σε όρους γλωσσών προγραμματισμού, πολυμέσων και αντιλήψεων, στο Διαδίκτυο έχουν μεταβληθεί εξαιρετικά, με τις μηχανές αναζήτησης να προσθέτουν ολοένα και περισσότερους παράγοντες για την όσο το δυνατό πιο αντικειμενική και ποιοτική προβολή αποτελεσμάτων στους χρήστες. Άλλωστε, η ίδια η μηχανή της Google δε θα μπορούσε να αφήσει απαράλλαχτο τον αλγόριθμο έκτοτε για να μην επιτρέψει στους διαχειριστές να χειραγωγήσουν τις υπηρεσίες αναζήτησης προς ίδιον όφελος.

3.2 Προσέγγιση των παραγόντων

Τον Απρίλιο του 2005, οι Bifet, Castillo και Chirita (2005) δημοσίευσαν μία ανάλυση των παραγόντων που χρησιμοποιούνται από τις μηχανές και επηρεάζουν άμεσα την απόδοση και τη θέση των ιστοσελίδων στα (οργανικά) αποτελέσματα αναζήτησης.

Σκοπός της εργασίας ήταν να μελετήσει την επιρροή των διαφόρων χαρακτηριστικών στοιχείων, παραγόντων, παραμέτρων των σελίδων στην κατάταξή τους στα αποτελέσματα της αναζήτησης. Χρησιμοποιώντας τη μηχανή αναζήτησης της Google, ως πλατφόρμα δοκιμών, αναλύθηκαν οι θέσεις των αποτελεσμάτων για αρκετούς όρους αναζήτησης, διαφορετικών κατηγοριών, με στατιστικές μεθόδους. Έπειτα, επαναδιατυπώθηκε το πρόβλημα προσδιορισμού και ανάλυσης των βαρών που αποδίδονται στους παράγοντες αυτούς ως ένα δυαδικής ταξινόμησης, στο οποίο εφαρμόστηκαν γραμμικές και μη γραμμικές μέθοδοι, διαχωρίζοντας τα δεδομένα σε ένα σύνολο εκπαίδευσης κι ένα δεδομένων εξέτασης.

Είχε προηγηθεί αντίστοιχη ανάλυση από τους Pringle, Allison και Dowe (1998), η οποία, όμως, επιχειρήθηκε με τη χρήση δέντρων αποφάσεων, για την εμφάνιση ή όχι της σελίδας ανεξαρτήτως θέσης στα αποτελέσματα αναζήτησης, και γραμμικής παλινδρόμησης στα δεδομένα που εξήχθησαν από τη μηχανή της InfoSeek. Παράλληλα, βασίστηκε σε σελίδες που δημιουργήθηκαν για τους σκοπούς της εργασίας και όχι ήδη υπάρχουσες, αγνοώντας έτσι τους πλέον ιδιαίτερα βεβαρημένους παράγοντες του επαρκούς, πραγματικού περιεχομένου και του χρόνου ωρίμανσης της ιστοσελίδας και του εξυπηρετητή που τη φιλοξενεί. Επίσης, έχουν, κατά καιρούς, επιχειρηθεί αντίστοιχες προσπάθειες, οι οποίες, όμως, περιορίζονταν σε ορισμένες μόνο μαθηματικές μεθόδους που εφαρμόστηκαν στο κοινό πρόβλημα, ενώ δεν διαχώριζαν τα δεδομένα στα δύο σύνολα που αναφέρθηκαν (training set, test set).

3.2.1 Προσομοίωση μοντέλου κατάταξης

Ο βασικός άξονας της έρευνας περιελάμβανε τον προσδιορισμό μίας συνάρτησης f, ως εκτίμηση της συνάρτησης κατάταξης μίας μηχανής αναζήτησης, σε πρώτη φάση, και τη σύγκριση των προβλεπόμενων, βάσει αυτής, αποτελεσμάτων με τα πραγματικά, στη συνέχεια.

3.2.1.1 Σύνολα δεδομένων

Οι Bifet et al. (2005) χρησιμοποίησαν διάφορα σύνολα από ομογενείς όρους αναζήτησης, σχετικούς με ένα συγκεκριμένο θέμα, θεωρώντας ότι τους επεξεργαζόταν η ίδια συνάρτηση.

Τα ερωτήματα διαχωρίστηκαν σε 4 κατηγορίες, βάσει αντικειμένου: Τέχνες, Πολιτείες, Spam και Πολλαπλές. Στην πρώτη κατηγορία εντάσσονταν ονόματα διάσημων καλλιτεχνών, στη δεύτερη ονόματα πολιτειών Αμερικής, η τρίτη περιελάμβανε τετριμμένες φράσεις – κλειδιά σχετικές με την αγορά ή δωρεάν μεταφόρτωση αρχείων πολυμέσων (ταινιών, μουσικής) και λογισμικού που ανταποκρίνονται σε πολύ μεγάλο αριθμό αναζητήσεων αλλά και ιστοσελίδων στο Διαδίκτυο και για τις οποίες, επομένως, έχουν χρησιμοποιηθεί ευρέως τεχνικές βελτιστοποίησης ιστοσελίδων. Τέλος, η τέταρτη κατηγορία περιελάμβανε μεγάλες φράσεις που αποτελούνταν από 6 όρους της στατιστικής.

Στη συνέχεια, οι 12 όροι κάθε κατηγορίας χωρίστηκαν σε 3 σύνολα με τα εξής χαρακτηριστικά:

- Σύνολο εκπαίδευσης (7 όροι): Σύνολο για τον προσδιορισμό μίας γραμμικής συνάρτησης κατάταξης ή ενός δέντρου αποφάσεων, τα δεδομένα που συλλέχθηκαν

από το οποίο αποτελούνται από διανύσματα που αποδίδουν έναν υπό εξέταση παράγοντα.

- Σύνολο επαλήθευσης (2 όροι): Σύνολο για τη σταδιακή, σειριακή επιλογή παραγόντων και το διαχωρισμό του δέντρου αποφάσεων.
- Σύνολο εξέτασης (3 όροι): Σύνολο για τον υπολογισμό του σφάλματος γενίκευσης της κατασκευασθείσας συνάρτησης, που προκύπτει από τη σύγκριση των προβλεπόμενων, βάσει αυτής, αποτελεσμάτων με τα πραγματικά αποτελέσματα αναζήτησης. Το σύνολο αυτό αποτελεί και τη βασική υπεροχή της έρευνας έναντι των προγενέστερων.

3.2.1.2 Δυαδική ταξινόμηση

Αρχικά, το πρόβλημα κατάταξης των αποτελεσμάτων μηχανής αναζήτησης αναδιατυπώθηκε ως δυαδικής ταξινόμησης.

Έστω q ένας όρος αναζήτησης, και **u, v** διανύσματα παραμέτρων των σελίδων, όπου κάθε συνιστώσα του διανύσματος αντιστοιχεί σε μία συγκεκριμένη παράμετρο, π.χ. πόσο συχνά εμφανίζεται ο όρος μέσα στη σελίδα, ή εάν εμφανίζεται στη διεύθυνση URL.

Έστω, λοιπόν, ότι ισχύει **u** < **v** εάν η σελίδα που χαρακτηρίζεται από το διάνυσμα **u** επιστρέφεται στα αποτελέσματα πιο κάτω από τη σελίδα διανύσματος **v**, για δεδομένο ερώτημα αναζήτησης. Ο σκοπός της μεθόδου είναι να βρεθεί μία πραγματική συνάρτηση ώστε f(**u**) < f(**v**) όταν **u** < **v**. Υποθέτοντας ότι η f είναι γραμμική, υπάρχει ένα διάνυσμα **w** τέτοιο ώστε η f(**u**) να ισούται με το εσωτερικό γινόμενο των διανυσμάτων **w, v.**

Το διάνυσμα **w**, από γεωμετρική άποψη, φέρει την κατεύθυνση της αύξησης της «βαθμολογίας» που αποδίδεται στην κάθε σελίδα (δηλαδή, στο κάθε διάνυσμα), ενώ μελετήθηκε η συμπεριφορά της διαφοράς των διανυσμάτων παραγόντων / παραμέτρων των δύο σελίδων **v** – **u**, εφαρμόζοντας διάφορα γραμμικά και μη μοντέλα, καθώς και των πιθανών συνδυασμών τους. Ενδεικτικά, αναφέρεται η μέθοδος λογιστικής παλινδρόμησης.

Έστω ότι το διάνυσμα **v** – **u**, που αναφέρθηκε προηγουμένως, φέρει την τιμή «+» εάν η σελίδα **v** κατατάσσεται πάνω από τη σελίδα **u**, ή «-» εάν συμβαίνει το αντίθετο. Η λογιστική παλινδρόμηση μοντελοποιεί τις πιθανότητες να ανήκει ένα στοιχείο σε συγκεκριμένη κλάση, δεδομένων των συντεταγμένων **x**, μέσω γραμμικών συναρτήσεων στο διάνυσμα συντεταγμένων. Στην περίπτωση των δύο κλάσεων «+» και «-», το μοντέλο έχει την εξής μορφή:

$$\log \frac{P(class = "+" \mid X = x)}{P(class = "-" \mid X = x)} = \beta_0 + \boldsymbol{W} \cdot \boldsymbol{X}$$

Λόγω συμμετρίας των δεδομένων, το β_0 θα είναι μηδέν και το διάνυσμα w δίνει τα επιθυμητά βάρη στους παράγοντες που μελετώνται, υποδεικνύοντας την κατεύθυνση από χαμηλές προς υψηλές θέσεις, στον άξονα κατάταξης των αποτελεσμάτων. Για τις δοκιμές τους, οι ερευνητές χρησιμοποίησαν την εφαρμογή της gmlfit (γενικευμένο γραμμικό μοντέλο) εντολής στο λογισμικό Matlab για τον υπολογισμό των μοντέλων λογιστικής παλινδρόμησης.

3.2.2 Αρχιτεκτονική υλοποίησης

Για την εξαγωγή των συνόλων με τα χαρακτηριστικά (παράγοντες, παραμέτρους) που χρησιμοποίησαν στην ανάλυσή τους, δημιούργησαν ένα σύστημα που αποτελούνταν από τρία μέρη: έναν μεταφορτωτή, έναν εξαγωγέα χαρακτηριστικών και έναν αναλυτή.

3.2.2.1 Μεταφόρτωση

Χρησιμοποιήθηκε ένα λογισμικό για να εκτελεί αναζητήσεις και να μεταφορτώνει τις σελίδες που επιστρέφονταν, χρησιμοποιώντας τα σύνολα με τους όρους αναζήτησης. Για κάθε αναζήτηση, ο μεταφορτωτής ακολουθούσε τα εξής βήματα:

1. Καταχώρηση του ερωτήματος στη μηχανή αναζήτησης, μεταφόρτωση των διευθύνσεων u_i που επιστρέφονταν ως ένα σύνολο αποτελεσμάτων κι έλεγχος εάν κάθε μία από αυτές τις URL διευθύνσεις έχει καταχωρηθεί στον κατάλογο Open Directory Project (DMOZ). Η καταχώρηση στον τελευταίο θεωρήθηκε τότε από τους ερευνητές ότι ίσως επηρέαζε την κατάταξη των αποτελεσμάτων των μηχανών αναζήτησης, παρότι σήμερα θεωρείται δεδομένο.

2. Αποστολή ενός ερωτήματος, για κάθε μία URL u_i, στη μηχανή αναζήτησης με τον τελεστή «link:», όπως έχει αναλυθεί και στο 2^o κεφάλαιο της παρούσας Διπλωματικής, για την εύρεση του αριθμού (και όχι της ποιότητας, σε αντίστοιχους όρους) των συνδέσμων προς την URL αυτή, καθώς και των 5 ισχυρότερων, ως προς την κατάταξη, σελίδων που συνέδεαν προς την υπό εξέταση ιστοσελίδα.

3. Μεταφόρτωση των 5 αυτών σελίδων με σκοπό την ανάλυση του anchor text του υπό εξέταση συνδέσμου.

Τέλος, για κάθε ξεχωριστό όρο αναζήτησης, ο μεταφορτωτής καταχωρεί ένα ερώτημα και εξαγάγει ον αριθμό των αποτελεσμάτων, κάτι που αργότερα θα χρησιμοποιηθεί για τον υπολογισμό των tf – idf ομοιοτήτων.

3.2.2.2 Εξαγωγή χαρακτηριστικών

Στο σημείο αυτό, λαμβάνονται ως είσοδοι οι μεταφορτωμένες σελίδες και υπολογίζονται τα χαρακτηριστικά των σελίδων αυτών, τα οποία χωρίζονται σε περιεχομένου, μορφοποίησης, συνδέσμου και meta δεδομένων. Οι παράγοντες που εξετάστηκαν φαίνονται παρακάτω και αποτελούν την πλειοψηφία των παραγόντων που επηρεάζουν τους αλγορίθμους κατάταξης των μηχανών αναζήτησης ακόμη και σήμερα.

Παράγοντες περιεχομένου, ανεξάρτητοι από το ερώτημα

- Αριθμός των διαφορετικών όρων εντός του εγγράφου.
- Ποσοστό των όρων εντός των εγγράφων που δε μπορούν να ευρεθούν σε ένα Αγγλικό λεξικό.
- Μέγεθος του εγγράφου, σε bytes.
- Μέγεθος του κειμένου του εγγράφου, σε bytes.
- Σχετική συχνότητα του πιο συχνού όρου (για παράδειγμα, εάν το έγγραφο περιλαμβάνει το κείμενο «electrical engineering, computer engineering», τότε η ζητούμενη τιμή είναι 2/4, δηλαδή 0,5).
- Μέση τιμή μεγέθους λέξεων.

Παράγοντες περιεχομένου, εξαρτημένοι από το ερώτημα

- Συχνότητα του όρου αναζήτησης εντός του κειμένου.
- Ομοιότητα του όρου με το έγγραφο (tf – idf).
- Μέση θέση των όρων αναζήτησης εντός του εγγράφου (1 στην αρχή, 0 στο τέλος και ενδιάμεσες τιμές εντός του κειμένου).
- Μέση ακριβής αντιστοιχία των όρων αναζήτησης.
- Σχετική απόσταση μεταξύ των όρων αναζήτησης εντός του εγγράφου.

Παράγοντες μορφοποίησης, εξαρτημένοι από το ερώτημα

- Ποσοστό εμφανίσεων του όρου αναζήτησης εντός των HTML ετικετών <b>, <i>, <u>, <h1> - <h6>, <a>, <li> και <title>.
- Ποσοστό εμφανίσεων του όρου αναζήτησης εντός των στοιχείων <img alt=””> και <img title=””>.

- Ποσοστό εμφανίσεων του όρου μέσα στις meta ετικέτες (λέξεις – κλειδιά και περιγραφή).
- Ποσοστό εμφανίσεων του όρου αναζήτησης με κεφαλαία γράμματα.

Παράγοντες συνδέσμου, ανεξάρτητοι από το ερώτημα

- Αριθμός των σελίδων που συνδέουν προς την ιστοσελίδα, με τη χρήση του τελεστή «link:».
- Βαθμός PageRank της σελίδας, ή μία εκτίμηση αυτού από την εργαλειοθήκη της Google, στη λογαριθμική κλίμακα 0-10.
- Αριθμός των εξερχόμενων συνδέσμων της σελίδας.
- Ποσοστό των εξερχόμενων συνδέσμων προς τις σελίδες στις οποίες συνέδεαν.

Παράγοντες meta δεδομένων

- Εμφάνιση του όρου στη διεύθυνση URL της σελίδας ή όχι.
- Εμφάνιση του όρου σε διαδικτυακούς καταλόγους.

3.2.2.3 Ανάλυση

Πρόκειται για το τελευταίο στάδιο του συστήματος ανάλυσης των τριών ερευνητών, κατά το οποίο υπολογίζεται η συνάρτηση κατάταξης, με τη χρήση των στατιστικών μεθόδων που αναλύθηκαν παραπάνω. Για την αξιολόγηση της συνάρτησης αυτής χρησιμοποίησαν τρία μεγέθη ακρίβειας των συνδυασμών σελίδων (διανυσμάτων παραγόντων).

3.2.3 Περιορισμοί ανάλυσης

Περιγράφηκαν παραπάνω οι παράγοντες που μελετήθηκαν και εκτιμήθηκε ότι έχουν άμεση σχέση με την επίδοση των ιστοσελίδων στα αποτελέσματα των μηχανών αναζήτησης, για δεδομένους όρους. Τα χαρακτηριστικά αυτά αποτελούν και σήμερα βασικούς παράγοντες τεχνικής βελτιστοποίησης μίας ιστοσελίδας, εντός κι εκτός αυτής.

Όπως, όμως, συμφωνούν και οι τέσσερις ερευνητές, πολλά στοιχεία που θα έπρεπε να εξετασθούν, καθώς εικάζεται ή έχει ήδη διευκρινιστεί από τους εκπροσώπους βασικών μηχανών αναζήτησης ότι σχετίζονται άμεσα ή έμμεσα με τη θέση που καταλαμβάνουν οι σελίδες στα αποτελέσματα σχετικών αναζητήσεων, αποκρύπτονται και δε μπορούν να εξετασθούν. Τέτοιοι παράγοντες είναι οι εξής:

- Ιστορικό αναζητήσεων ιστού
- Ηλικία των εισερχόμενων συνδέσμων
- Δυναμικότητα των ιστοσελίδων (ρυθμός ανανέωσης ή πρόσθεσης περιεχομένου)
- Πραγματική εικόνα των εισερχόμενων συνδέσμων προς τις εξεταζόμενες ιστοσελίδες (καθώς, όπως έχει αναφερθεί, ο τελεστής «link:» δίνει, ακόμη και σήμερα, μία πολύ κακή εκτίμηση αυτών)
- Η πραγματική τρέχουσα τιμή του βαθμού PageRank που χρησιμοποιούν οι μηχανές αναζήτησης, καθώς η τιμή που δίνεται στις εργαλειοθήκες είναι μία μόνο (ανηγμένη σε γραμμική κλίμακα) εκτίμηση αυτής που ανανεώνεται, συνήθως, κάθε 3 – 6 μήνες.

3.2.4 Συμπεράσματα

Παρατηρήθηκε ότι όλοι αυτοί οι παράγοντες είχαν άμεση επίδραση στην κατάταξη των αποτελεσμάτων, όπως αρχικά είχε εκτιμηθεί, ενώ, όπως ήταν αναμενόμενο, παράγοντες όπως το μέγεθος του εγγράφου, ο αριθμός των λέξεων που δεν ερμηνεύονται στο λεξικό αλλά και το μήκος των λέξεων έχουν, κατά βάση, αρνητική επίδραση, σε αντίθεση με τους υπόλοιπους παράγοντες που είχαν κυρίως θετική.

Τέλος, αξίζει να σημειωθεί πως η πραγματική αξία αυτής της ανάλυσης δεν περιορίζεται στην προσπάθεια εκτίμησης των αλγορίθμων – συναρτήσεων που χρησιμοποιούνται από τις μηχανές αναζήτησης για την κατάταξη των αποτελεσμάτων, αλλά έγκειται κυρίως στην κατεύθυνση που δόθηκε, σε ακαδημαϊκό επίπεδο, ως προς τους παράγοντες αυτούς καθ' αυτούς που επηρεάζουν την επίδοση των ιστοσελίδων στις μηχανές αναζήτησης.

4 *Τεχνικές βελτιστοποίησης εντός της ιστοσελίδας*

Κατά τον σχεδιασμό και την εφαρμογή τεχνικών βελτιστοποίησης εντός του ιστοχώρου, μεθόδων, δηλαδή, που εφαρμόζονται στον κώδικα και τη δομή μιας σελίδας καθαυτής, καθώς και στην οργάνωση του ιστοχώρου σε επίπεδο εξυπηρετητή, προσπαθούμε να δημιουργήσουμε έγγραφα που αφενός διευκολύνουν την ακριβή ανίχνευσή τους από τις μηχανές αναζήτησης, αναδεικνύοντας το περιεχόμενό τους, αφετέρου κρίνονται ιδιαίτερα σχετικά με ορισμένους όρους αναζήτησης.

Όπως έχει ήδη αναφερθεί, το πλέον βασικό και σημαντικό κριτήριο σχετικότητας μίας ιστοσελίδας με μία ορισμένη λέξη ή φράση είναι η συχνότητα με την οποία εμφανίζεται η λέξη ή φράση αυτή στο περιεχόμενο της σελίδας. Παρ' όλα αυτά, υπάρχουν πολλοί άλλοι παράγοντες αξιολόγησης της σελίδας από τις μηχανές αναζήτησης, οι οποίοι, βέβαια, στην πλειοψηφία τους συνδέονται άμεσα με την εμφάνιση του όρου αναζήτησης και αφορούν στον προσδιορισμό της βαρύτητας της εμφάνισης αυτής.

Η Google (1998), μάλιστα, αναφέρει πως «ασχολείται με θέματα που δεν περιορίζονται στον αριθμό των εμφανίσεων ενός όρου σε μία σελίδα και εξετάζει λεπτομερώς όλες τις πτυχές του περιεχομένου της ιστοσελίδας (και των σελίδων που συνδέουν προς αυτήν) για να συμπεράνει εάν αποτελεί καλό αποτέλεσμα για την αναζήτηση του χρήστη».

Το παρόν κεφάλαιο πραγματεύεται όλους εκείνους τους παράγοντες κατάταξης που σχετίζονται με τη χρήση των HTML ετικετών σήμανσης και της δομής, καθώς επίσης και με το περιεχόμενο των σελίδων καθαυτό.

4.1 Μέγεθος σελίδας και συχνότητα όρων

Η συχνότητα με την οποία μία λέξη ή φράση – κλειδί εμφανίζεται και επαναλαμβάνεται μέσα στο κείμενο μίας ιστοσελίδας υπήρξε ο κυρίαρχος, αν όχι μοναδικός, παράγοντας για τον προσδιορισμό της θέσης της σελίδας στα αποτελέσματα, για τη συγκεκριμένη λέξη ή φράση αναζήτησης. Όμως, όπως ήταν φυσικό, με την ανάπτυξη του Παγκόσμιου Ιστού και τη συμμετοχή ολοένα και περισσότερων κερδοσκοπικών οργανισμών κι επιχειρήσεων στο Διαδίκτυο, το γεγονός αυτό οδήγησε στην κατάχρηση και άσκοπη επανάληψη των λέξεων (spamming) για τις οποίες μία επιχείρηση ενδιαφερόταν να φαίνεται υψηλά στις μηχανές αναζήτησης, όπως για παράδειγμα *«κινητή τηλεφωνία... κινητή τηλεφωνία... κινητή τηλεφωνία.. κινητή τηλεφωνία...»*.

Στη συνέχεια, οι αλγόριθμοι των μηχανών αναζήτησης έσπευσαν να ευνοήσουν την πυκνότητα αντί της συχνότητας των λέξεων ή φράσεων μέσα στις σελίδες, κάτι το οποίο, όμως, είχε σαν αποτέλεσμα την τιμωρία αρκετών ιστοτόπων για υπερβολική και άκομψη προσπάθεια να «χωρέσουν» λέξεις και φράσεις μέσα στη σελίδα. Η επίτευξη μίας καλής αλλά μετριοπαθούς πυκνότητας λέξεων –κλειδιών αποτελεί και σήμερα θετικό κριτήριο και υποδεικνύει στις μηχανές ότι η σελίδα, όντως, σχετίζεται με τις συγκεκριμένες λέξεις – κλειδιά.

Η **πυκνότητα μίας λέξης – κλειδιού**, επομένως, ορίζεται ως η συχνότητα με την οποία εμφανίζεται αυτή εντός της σελίδας ως ποσοστό του συνολικού μεγέθους της σελίδας.

$$\text{πυκνότητα λέξης - κλειδιού} = \frac{\text{συχνότητα λέξης - κλειδιού}}{\text{μέγεθος σελίδας}}$$

Για παράδειγμα, μία λέξη που εμφανίζεται 10 φορές σε μία σελίδα μεγέθους 400 λέξεων (πυκνότητα = 2,5%) έχει μεγαλύτερη πυκνότητα από μία λέξη που εμφανίζεται 30 φορές σε μία σελίδα μεγέθους 3000 λέξεων (πυκνότητα = 1%).

Αντίστοιχα, το μέγεθος μίας σελίδας ορίζεται από τις μηχανές αναζήτησης ως το σύνολο των λέξεων του εγγράφου, χωρίς τα scripts, τις ετικέτες σήμανσης και τον τομέα <head> (δηλαδή, το κομμάτι του κώδικα που εμπεριέχεται στις ετικέτες <head>).

Σε περίπτωση που ο περιορισμός του μεγέθους είναι επιθυμητός, συνίσταται ο διαχωρισμός του περιεχομένου σε διαφορετικά έγγραφα με σελιδοποίηση και η επιλογή της συνολικής προβολής του περιεχομένου σε διαφορετικό έγγραφο που προορίζεται, ενδεχομένως, για σκοπούς εκτύπωσης. Στην περίπτωση αυτή, οφείλουμε να ορίσουμε το κανονικοποιημένο έγγραφο (canonicalization), μέσα από μία πρακτική που θα αναλυθεί σε επόμενη παράγραφο.

Η επίτευξη μίας ορισμένης πυκνότητας συγκεκριμένης λέξης – κλειδιού εντός ενός εγγράφου είναι κρίσιμη για τη συσχέτιση της λέξης με το έγγραφο αυτό, κατά την ανίχνευση και ευρετηρίαση αυτού από τις μηχανές αναζήτησης. Πρέπει να τονιστεί, όμως, ότι ο παράγοντας της πυκνότητας, πλέον, δεν προσφέρει περιθώρια τεχνικής βελτιστοποίησης, αλλά σημασιολογικής, και δεν πρέπει να αποτελεί κριτήριο διαμόρφωσης του κώδικα, ενώ αναφέρεται ως παρελθοντικός παράγοντας αξιολόγησης και ως μέτρο που έχει ορισθεί και αναλυθεί πολλάκις στη βιομηχανία της κατασκευής ιστοσελίδων. Μάλιστα, για πυκνότητα λέξεων μεγαλύτερη του 10%, έχει παρατηρηθεί επιβολή ποινής στη σελίδα και πτώση της θέσης αυτής στα αποτελέσματα των δημοφιλέστερων μηχανών αναζήτησης.

Συγκεκριμένα, σύμφωνα με την ανάλυση των Dr. E. Garcia και Mike Grehan (Mi Islita, 2005), η βελτιστοποίηση της πυκνότητας των λέξεων – κλειδιών, συναρτήσει του συνολικού μεγέθους ενός εγγράφου, αποτελεί μύθο στη βιομηχανία του online marketing. Η θέση αυτή αποδεικνύεται ως εξής:

Συμβολίζοντας την πυκνότητα μίας λέξης i με *KD*, τη συχνότητα αυτής σε ένα έγγραφο j με $tf_{i,j}$ και το συνολικό μέγεθος του εγγράφου j σε λέξεις με l_j, η παραπάνω εξίσωση γράφεται:

$$KD = \frac{tf_{i,\,j}}{l_j}$$

Όμως, η βαρύτητα ενός όρου σε ένα έγγραφο αποτελείται από τρεις τύπους βαρών: το τοπικό, το γενικό και την κανονικοποίηση. Επομένως, το συνολικό βάρος γράφεται:

$$w_{i,\,j} = L_{i,\,j} * G_i * N_j$$

Στην προηγούμενη εξίσωση, το $L_{i,j}$ είναι το τοπικό βάρος για τον όρο i στο έγγραφο j, G_i είναι το γενικό βάρος για τον όρο i και N_j είναι ο παράγοντας κανονικοποίησης για το έγγραφο j. Τα τοπικά βάρη είναι συναρτήσεις συχνότητας εμφάνισης ενός όρου σε ένα έγγραφο, τα γενικά βάρη είναι συναρτήσεις συχνότητας των εγγράφων που περιλαμβάνουν τον όρο μέσα σε μία συλλογή (έστω το ευρετήριο της μηχανής αναζήτησης), ενώ ο παράγοντας κανονικοποίησης διορθώνει τυχόν αποκλίσεις μεταξύ των μεγεθών των εγγράφων της συλλογής.

Επομένως, τα βάρη, υπό ιδανικές συνθήκες, μπορούν να γραφούν και ως εξής:

$$L_{i,\,j} = tf_{i,\,j}$$

$$G_i = \log\left(\frac{D}{d_i}\right)$$

$$N_j = 1$$

Και το συνολικό βάρος μπορεί να εκφρασθεί συναρτήσει των δεύτερων μερών των παραπάνω εξισώσεων:

$$w_{i,j} = tf_{i,j} * \log\left(\frac{D}{d_i}\right)$$

Σε αυτή την εξίσωση, ο όρος $\log\left(\frac{D}{d_i}\right)$ ονομάζεται και αντίστροφη συχνότητα εγγράφου (Inverse Document Frequency – IDF), το *D* εκφράζει τον αριθμό των εγγράφων της συλλογής (το μέγεθος, δηλαδή, του ευρετηρίου σε έγγραφα) και το d_i εκφράζει τον αριθμό των εγγράφων της συλλογής που περιέχουν τον όρο i (Manning, Raghavan & Schütze, 2008). Η εξίσωση αυτή είναι μία από τις πολλές εξισώσεις που συναντώνται στη βιβλιογραφία.

Ο μόνος τρόπος, επομένως, που θα μπορούσε η πυκνότητα της λέξης i στο έγγραφο j να ισούται με τη βαρύτητα του όρου $(KD = w_{i,j} = \frac{tf_{i,j}}{l_j})$ είναι να αγνοηθούν τα γενικά βάρη και ο παράγοντας της κανονικοποίησης να εξαρτάται αποκλειστικά και να είναι ανάλογος του συνολικού μεγέθους του εγγράφου l_j:

$$G_i = IDF = 1$$

$$N_j = \frac{1}{l_j}$$

Η πρώτη σχέση συνεπάγεται ότι το μέγεθος της συλλογής D θα είναι ίσο με 10 φορές τον αριθμό των εγγράφων που περιλαμβάνουν τον όρο i και η δεύτερη ότι ο παράγοντας της κανονικοποίησης δε λαμβάνει υπόψη το διαχωρισμό των λέξεων.

Επομένως, γίνεται σαφές ότι δε θα έπρεπε να τίθεται θέμα τεχνικής βελτιστοποίησης της συχνότητας των σημαντικών λέξεων και του μεγέθους του κάθε εγγράφου, αρκεί να διατηρείται μια ικανοποιητική εμφάνιση των όρων που μας ενδιαφέρουν.

4.2 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.ntua.gr/schools.html. Προτού την επισκεφθεί, ελέγχει τη διεύθυνση http://www.ntua.gr/robots.txt , στην οποία διαβάζει τα εξής:

```
User-agent: *
Disallow:   /cgi-bin/

User-agent: Mozilla/3.01 (hotwired-test/0.1)
Disallow:   /cgi-bin/

User-agent: Slurp
Disallow:   /cgi-bin/

User-agent: Scooter
Disallow:   /cgi-bin/

User-agent: Ultraseek
Disallow:   /cgi-bin/

User-agent: smallbear
Disallow:   /cgi-bin/

User-agent: GoogleBot
Disallow:   /cgi-bin/

User-agent: *
Disallow:   /pub/
```

Με το παραπάνω αρχείο, οι διαχειριστές του ιστοχώρου ntua.gr επιθυμούν να αποκλείσουν την είσοδο όλων των ανιχνευτών στους φακέλους /cgi-bin/ και /pub/, ενώ έχουν απευθυνθεί κι ειδικότερα σε ορισμένους βασικούς ανιχνευτές (των Google, Yahoo, Altavista), αποκλείοντάς τους από τον φάκελο /cgi-bin/ και επιτρέποντάς τους την πρόσβαση σε όλο τον υπόλοιπο ιστοχώρο, συμπεριλαμβανομένου και του φακέλου /pub/.

Συγκεκριμένα, με την εντολή "User-agent: *" του παραπάνω αρχείου, απευθυνόμαστε σε όλους τους ανιχνευτές, από οποιαδήποτε μηχανή αναζήτησης ή ιστοχώρο, δίνοντάς τους τις οδηγίες που βρίσκονται ακριβώς μετά από αυτήν και μέχρι την κενή γραμμή που ακολουθείται από άλλη αντίστοιχη εντολή.

Η εντολή "Disallow: /cgi-bin/" δηλώνει στους ανιχνευτές να μην επισκεφθούν τον συγκεκριμένο φάκελο, καθώς και όλα τα αρχεία και τους υποφακέλους που υπάρχουν μέσα σε αυτόν.

Αντίστοιχα, η εντολή "Disallow: /cgi-bin/", μετά την εντολή "User-agent: Slurp" δηλώνει στον ανιχνευτή της μηχανής αναζήτησης Yahoo! να μην επισκεφθεί το φάκελο /cgi-bin/.

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.ntua.gr/sitemap.xml
Sitemap: http://www.ntua.gr/updates/new_sitemap.xml


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

Καλύτερες πρακτικές

α) Με τη χρήση του αρχείου robots.txt μπορούμε να εμποδίσουμε την ανίχνευση φακέλων και υποφακέλων των οποίων την ανίχνευση δεν επιθυμούμε. Για παράδειγμα, είναι σύνηθες φαινόμενο να μην επιτρέπεται η πρόσβαση στους φακέλους που αφορούν τη διαχείριση ενός ιστοτόπου (/administrator, /admin, /administration) οι οποίοι δημιουργούνται δυναμικά σε συστήματα διαχείρισης περιεχομένου, τους φακέλους με τα διάφορα templates, plugins, modules που χρησιμοποιήθηκαν στην κατασκευή της ιστοσελίδας, τον φάκελο με τα αρχεία της εγκατάστασης του συστήματος διαχείρισης περιεχομένου, καθώς και οποιονδήποτε άλλον φάκελο επιθυμούμε να διατηρήσουμε μακριά από τις μηχανές αναζήτησης.

β) Είναι σαφές πως με το αρχείο αυτό έχουμε τη δυνατότητα να εμποδίσουμε την πρόσβαση των ανιχνευτών σε έγγραφα των οποίων τα δικαιώματα δε μας ανήκουν (π.χ. εικόνες) ή σε περιεχόμενο που ανήκει σε ξένο ιστότοπο και αναπαράγουμε, επιθυμώντας, φυσικά, να μη τιμωρηθεί η σελίδα μας για διπλότυπο περιεχόμενο από τον αλγόριθμο της μηχανής αναζήτησης. Αυτή η μέθοδος βρίσκει εφαρμογή και σε περιεχόμενο του ίδιου του ιστοτόπου μας που αναπαράγεται δυναμικά από πρόσθετες εφαρμογές ανάθεσης σημασιολογικά εντάξει ονομάτων στα παραγόμενα έγγραφα, των οποίων η διεύθυνση URL εξ αρχής αποτελείται από πολλές παραμέτρους, όπως θα αναλυθεί σε διαφορετικό κεφάλαιο της παρούσας διπλωματικής. Τέλος, υπάρχει περίπτωση να αναπαραγάγουμε εκούσια διπλότυπο περιεχόμενο σε διαφορετικές μορφές (όπως, για παράδειγμα, μία εναλλακτική εκδοχή των κειμένων σε ευνοϊκά εκτυπώσιμη μορφή) και, για λόγους προστασίας από ποινές διπλότυπου περιεχομένου, το αρχείο robots.txt μας δίνει αυτή ακριβώς τη δυνατότητα. Υπό την ίδια έννοια, μπορούμε να εμποδίσουμε την πρόσβαση των robots σε σελίδες σφάλματος 404 (έχοντας ορίσει αυτές τις σελίδες, με ανακατεύθυνση από σελίδες με σφάλμα στην πλοήγηση του χρήστη στον ιστότοπο), καθώς επίσης και σε σελίδες με αποτελέσματα εσωτερικής αναζήτησης, οι οποίες παράγονται δυναμικά κατά την αναζήτηση ενός χρήστη σε κάποιον υποφάκελο (πλην της περίπτωσης όπου χρησιμοποιείται τεχνολογία AJAX για την διεξαγωγή εσωτερικών αναζητήσεων, οπότε τα αποτελέσματα εμφανίζονται στο ίδιο έγγραφο με αυτό της φόρμας αναζήτησης).

γ) Το αρχείο robots.txt οφείλει να βρίσκεται στον αρχικό φάκελο (root directory) ‘/’. Αυτό σημαίνει ότι οποιαδήποτε άλλη τοποθεσία δεν είναι έγκυρη, καθώς οι ανιχνευτές δεν πρόκειται να ανιχνεύσουν τον ιστοχώρο με σκοπό να το βρουν, καθώς δε θα έχουν τις αντίστοιχες οδηγίες για τα μονοπάτια που επιτρέπεται να ακολουθήσουν. Ακόμη κι αν ο ανιχνευτής παραπεμφθεί από κάποιον σύνδεσμο σε μία σελίδα, έστω http://www.ntua.gr/schools/ece/courses.html, δεν θα αναζητήσει κάποιο robots αρχείο στον υποφάκελο /ece/, αλλά θα ανατρέξει στον αρχικό φάκελο όπου και θα αναζητήσει το αρχείο. Στην περίπτωση που αυτό δεν υπάρχει, ή είναι κενό, ο ανιχνευτής θα θεωρήσει ότι όλες οι περιοχές του διακομιστή είναι προσβάσιμες. Παρ’ όλα αυτά, η ύπαρξη του αρχείου αυτού, έστω και κενού, καθιστά την ιστοσελίδα πιο φιλική στις μηχανές αναζήτησης, διευκολύνοντας τη λειτουργία των ανιχνευτών τους, επομένως και θα εκμαιεύσει μια πιο ευνοϊκή αντιμετώπιση από τους αλγορίθμους αυτών.

δ) Η ονομασία των εγγράφων ενός διακομιστή είναι case sensitive και το ίδιο ισχύει για το αρχείο robots.txt. Αυτό σημαίνει ότι ο ανιχνευτής δεν θα αναζητήσει το robots.TXT, ή το ROBOTS.txt, ή οποιονδήποτε συνδυασμό πεζών και κεφαλαίων λατινικών χαρακτήρων, αλλά το robots.txt. Εάν, δηλαδή, υπάρχει το RoBoTs.txT αρχείο, ο ανιχνευτής θα το αγνοήσει και θα προχωρήσει στην ανίχνευση ολόκληρου του ιστοχώρου.

ε) Μεγάλη προσοχή πρέπει να δοθεί στις πληροφορίες που ο διαχειριστής θεωρεί πως μπορεί να αποκρύψει. Στο σημείο αυτό, τονίζεται ότι το αρχείο robots.txt είναι διαθέσιμο για ανάγνωση σε οποιονδήποτε επισκεφθεί τη διεύθυνση URL /robots.txt (καθώς και οι πληροφορίες που αναγράφονται μέσα σε αυτό μέχρι και οι πληροφορίες που ενδεχομένως ο επισκέπτης θα ανακαλύψει ακολουθώντας τους φακέλους των οποίων την ανίχνευση απαγορεύουμε στους ανιχνευτές.

στ) Τέλος, οι οδηγίες που δίνονται στο robots.txt δεν έχουν καμία αξία εάν η σύνταξη των εντολών είναι λανθασμένη και δεν ακολουθεί τα πρότυπα, ενώ υπάρχει η περίπτωση ο ίδιος ο ανιχνευτής να μην συμβιβαστεί με τη διαδικασία ανάγνωσης του αρχείου robots, να το αγνοήσει ή να προσπεράσει τις εντολές. Παρόλο που η συνεργασία του ανιχνευτή απαιτείται για τη διαδικασία αυτή, έχει παρατηρηθεί η δραστηριότητα κακόβουλων ανιχνευτών που αγνοούν το αρχείο γιατί στοχεύουν στην εξαγωγή πελατολογίων, διευθύνσεων ηλεκτρονικής αλληλογραφίας και άλλων προσωπικών κι ευαίσθητων στοιχείων των οποίων την έκθεση στα αποτελέσματα των μηχανών αναζήτησης ο διαχειριστής ενδέχεται να επιθυμεί να αποτρέψει.

4.3 Meta – Ετικέτες

Οι Meta – ετικέτες (Meta tags) αποτελούν κομμάτι του αρχείου HTML (ή του κώδικα HTML, αν πρόκειται για κάποιο άλλο αρχείο, π.χ. PHP), το οποίο τοποθετείται στον τομέα <HEAD> μίας σελίδας και διαβάζεται από το φυλλομετρητή και τις μηχανές αναζήτησης.

Αυτές οι ετικέτες αποκρύπτονται αποτελεσματικά 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. Παλαιότερα, όπως θα δούμε, εκτός των άλλων, πολύ σπουδαίο ρόλο και σημασία κατείχε η ετικέτα των λέξεων – κλειδιών (Meta keywords tag), η οποία, όμως, έχασε την αξία της κατά τον ίδιο τρόπο που έχασαν την αξία τους η συχνότητα και η πυκνότητα των λέξεων – κλειδιών.

Η αξιοποίηση των ετικετών, όσον αφορά την τεχνική βελτιστοποίηση της ιστοσελίδας στα αποτελέσματα των μηχανών αναζήτησης, είναι αμφιβόλου αξίας και αμφισβητείται, ήδη από τον Σεπτέμβριο του 2009, τόσο από επαγγελματίες SEOs όσο και κυρίως από τους μηχανικούς της Google (Google Webmaster Central Blog, 2009).

Παρά την έλλειψη γνώσης της πραγματικής τους αξίας για τους αλγορίθμους κατάταξης, η εμπειρία έχει αποδείξει πως η ορθή χρήση και βελτιστοποίηση των Meta – ετικετών, όπως

αυτή αναλύεται στο παρόν κεφάλαιο, 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. Το γνώρισμα HTTP-EQUIV αφορά αποκλειστικά σε πληροφορίες που απευθύνονται στον διακομιστή (όπως η κωδικοποίηση χαρακτήρων, η ημερομηνία λήξης, κ.λπ.) κι, επομένως, σπάνια αφορούν τεχνικές βελτιστοποίησης των ιστοσελίδων. Η σύνταξη των δύο περιπτώσεων φαίνεται παρακάτω:

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

4.3.1 Meta ετικέτα περιγραφής

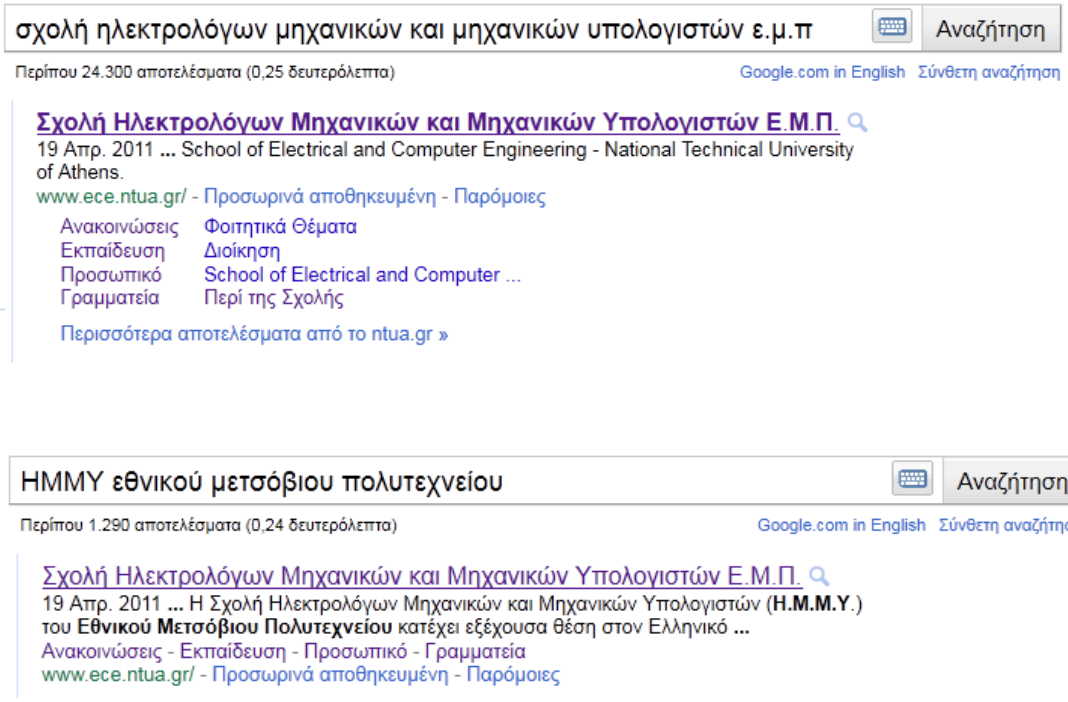
Οι Meta – περιγραφές, οι οποίες αποτελούν HTML γνωρίσματα (attributes) που παρέχουν ακριβείς και συνοπτικές επεξηγήσεις του περιεχομένου των ιστοσελίδων στους χρήστες, χρησιμοποιούνται ευρέως από τις μηχανές αναζήτησης κατά την παρουσίαση των αποτελεσμάτων αναζήτησης με σκοπό την υπόδειξη ενός αποσπάσματος προεπισκόπησης μίας δεδομένης σελίδας.

Οι ετικέτες αυτές, παρότι, σύμφωνα με αρκετές πηγές πληροφοριών στο Διαδίκτυο αλλά και επίσημες πηγές ενημέρωσης των μηχανών αναζήτησης, όχι ιδιαίτερα σημαντικές για την κατάταξη των αποτελεσμάτων, είναι εξαιρετικά σημαντικές για την προσέλκυση επισκεπτών από τις σελίδες των αποτελεσμάτων των μηχανών (Search Engine Result Pages – SERPs). 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.

Τονίζεται εδώ ότι οι μηχανές αναζήτησης δεν δεσμεύονται να χρησιμοποιήσουν την ετικέτα περιγραφής για να περιγράφουν την ιστοσελίδα, σε κάθε αναζήτηση που αυτή επιστρέφει ως αποτέλεσμα, αλλά τείνουν να την χρησιμοποιούν εφόσον το ερώτημα της αναζήτησης βρίσκεται μέσα στο περιεχόμενο που αποδίδεται σε αυτήν. 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, όπως αναλύεται παρακάτω, οι META ετικέτες δίνουν τη δυνατότητα στον διαχειριστή να ελέγξει λίγο περισσότερο την περιγραφή που συνοδεύει τα αποτελέσματα αναζήτησης, ή και, εάν επιθυμεί, να μη συμπεριλάβει καμία περιγραφή.

Για παράδειγμα, η ιστοσελίδα http://www.ece.ntua.gr με την ετικέτα `<meta name="description" content="School of Electrical and Computer Engineering - National Technical University of Athens" />`, για δύο διαφορετικές αναζητήσεις, επιστρέφει διαφορετική περιγραφή:

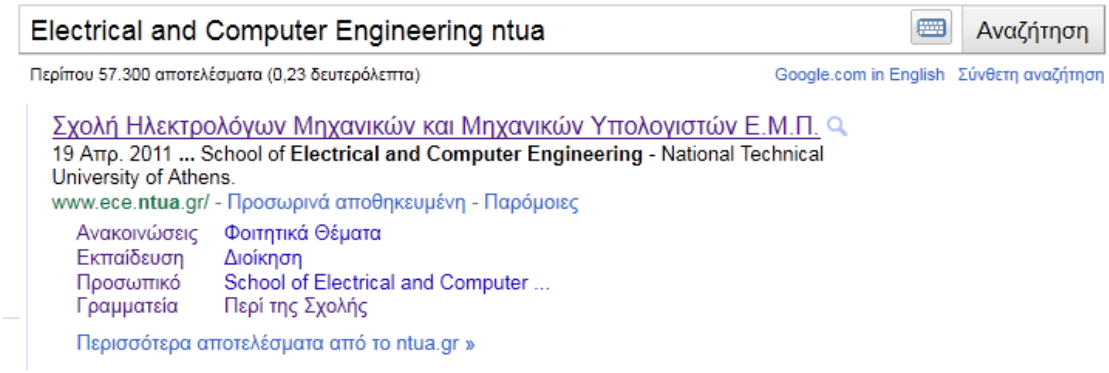


Εικόνα 4 Περιγραφή αποτελεσμάτων για ίδια σελίδα, διαφορετικά ερωτήματα

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. Το ίδιο θα είχε συμβεί εάν ο όρος της αναζήτησης υπήρχε μέσα στο περιεχόμενο της meta ετικέτας περιγραφής, ή εφόσον οι λέξεις

του ερωτήματος μοιράζονταν στον τίτλο και την meta περιγραφή, ανεξάρτητα από την εμφάνισή του ή όχι και στο περιεχόμενο της σελίδας, όπως φαίνεται εδώ:



Εικόνα 5 Διαφοροποίηση περιγραφής αποτελέσματος από τη meta ετικέτα περιγραφής

Παρατηρούμε, λοιπόν, ότι οι μηχανές αναζήτησης, όσον αφορά την περιγραφή των ιστοσελίδων που επιστρέφονται ως αποτελέσματα, δίνουν προτεραιότητα στον τίτλο και τη meta ετικέτα περιγραφής. Εάν δεν έχει ευρετηριασθεί ο όρος αναζήτησης σε κανένα μέρος εκ των δύο, τότε αποκόπτεται το πλέον σχετικό με τον όρο απόσπασμα κειμένου από το περιεχόμενο του εγγράφου.

Τέλος, παρατηρείται ότι, είτε εμφανισθεί απόσπασμα από το κείμενο είτε από την περιγραφή, οι μηχανές αναζήτησης θα παρουσιάσουν περιγραφές που, συνήθως, αποτελούνται από σύντομες φράσεις που εμπεριέχουν (κυρίως κεντρικά) λέξεις ή φράσεις του συνολικού ερωτήματος αναζήτησης, στην περίπτωση που δεν γειτνιάζουν όλες οι λέξεις του ερωτήματος, όπως φαίνεται στην παρακάτω αναζήτηση, όπου εμφανίζεται μέρος της ετικέτας meta description που εμπεριέχει μέρος του όρου και απόσπασμα του περιεχομένου της σελίδας της ΣΗΜΜΥ ΕΜΠ το οποίο εμπεριέχει το υπόλοιπο κομμάτι του ερωτήματος:

Εικόνα 6 Συνδυασμός meta ετικέτας περιγραφής και περιεχομένου στα αποτελέσματα

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>
```

Καλύτερες πρακτικές

α) Οι meta περιγραφές μπορούν να έχουν απεριόριστο μήκος, σε χαρακτήρες. Όμως, οι μηχανές αναζήτησης δεν εμφανίζουν περισσότερους από 160 χαρακτήρες. Επομένως, είναι σαφές ότι κάθε διαχειριστής έχει τη δυνατότητα να ορίσει όσο μακροσκελή περιγραφή επιθυμεί, στις meta ετικέτες, αλλά ο βέλτιστος έλεγχος ως προς το παρεχόμενο απόσπασμα επιτυγχάνεται περιορίζοντας τις περιγραφές στους 155 – 160 χαρακτήρες. Αυτό συμβαίνει διότι δε μπορεί να γνωρίζει εξ αρχής ποιο μέρος της περιγραφής θα αποκόψουν και ποιο θα εμφανίσουν.

β) Όπως και με άλλα ιδιαίτερα στοιχεία ενός HTML αρχείου, οι meta ετικέτες περιγραφής οφείλουν να είναι μοναδικές σε κάθε σελίδα αλλά και μοναδικές σε σχέση με άλλους ιστοχώρους. Αυτό συμβαίνει διότι η ετικέτα περιγραφής αποτελεί ένα από τα κριτήρια εντοπισμού σελίδων με διπλότυπο περιεχόμενο, οι οποίες τιμωρούνται με ποινές υποβάθμισης στα αποτελέσματα αναζήτησης ή και με ποινή αφαίρεσης από το ευρετήριο μίας μηχανής. Στην περίπτωση μεγάλων δυναμικών ιστοτόπων, όπου οι σελίδες παράγονται δυναμικά έως και αυτόματα και η ανάθεση μοναδικού περιεχομένου στις διάφορες ετικέτες περιγραφής είναι δύσκολη, μία λύση είναι η δημιουργία ενός δυναμικού προγραμματιστικού τρόπου ανάθεσης ετικετών περιγραφής (π.χ. με την ανάθεση, ως περιγραφής, της σύντομης περιγραφής ενός προϊόντος σε ένα ηλεκτρονικό κατάστημα ή της πρώτης πρότασης του περιεχομένου της σελίδας σε ένα blog, ή της περιγραφής της χαρακτηριστικής εικόνας ή του υπότιτλου ενός άρθρου σε μία δημοσιογραφική διαδικτυακή πύλη).

γ) Μπορεί, όντως, η ετικέτα αυτή να μην έχει τον καθοριστικό ρόλο που κάποτε είχε, όσον αφορά την κατάταξη των αποτελεσμάτων, ή να μην χρησιμοποιείται και καθόλου στους αλγορίθμους κατάταξης, όπως διατείνονται οι μηχανικοί των μηχανών αναζήτησης, όμως πρακτικά εξυπηρετεί τον σκοπό της πιο άμεσης διαφήμισης μίας σελίδας στους χρήστες των μηχανών αναζήτησης. Όπως ακριβώς ισχύει και με τον τίτλο της σελίδας, ένα αποτελεσματικό call-to-action, σε συνδυασμό με μία περιεκτική περιγραφή που θα πετύχει τα πιο σχετικά με τη σελίδα ερωτήματα αναζήτησης, θα έχει τα καλύτερα δυνατά αποτελέσματα.

Στο σημείο αυτό, τονίζεται ότι, όπως και με τον τίτλο της σελίδας, οι λέξεις του ερωτήματος που υπάρχουν στην περιγραφή που εμφανίζεται στα αποτελέσματα τονίζονται με bold.

δ) Η μηχανή αναζήτησης της Google αποκόπτει τις περιγραφές που ακολουθούν διπλά εισαγωγικά (“). Επομένως, προτείνεται να μη χρησιμοποιούνται γενικότερα αλφαριθμητικοί χαρακτήρες, ενώ, εάν απαιτούνται τα διπλά εισαγωγικά στην ετικέτα της περιγραφής, είναι προτιμότερο να χρησιμοποιούνται δύο απλά εισαγωγικά (‘ ’).

ε) Τέλος, υπάρχει η δυνατότητα, όπως αναφέρθηκε, μέσα από τη meta ετικέτα ανιχνευτών, να μη χρησιμοποιείται κάποια περιγραφή στα αποτελέσματα μίας σχετικής αναζήτησης. Εξίσου σημαντική στρατηγική, ορισμένες φορές, αποτελεί η πρακτική να μη χρησιμοποιείται η meta ετικέτα περιγραφής καθόλου στο αρχείο HTML, όταν αυτό κρίνεται σκόπιμο. Η ετικέτα περιγραφής είναι κατάλληλη όταν επιθυμείται η στόχευση 2 – 3 δημοφιλών ή κύριων λέξεων – κλειδιών. Όμως, στην περίπτωση που στοχεύονται πολλαπλές λέξεις της μακροσκελούς ουράς αναζήτησης, η επιλογή της μη ανάθεσης meta περιγραφής προσφέρει περισσότερα. Για παράδειγμα, ορισμένες διαδικτυακές πύλες ενημέρωσης ή μεγάλα blogs που ασχολούνται με πολυποίκιλα θέματα, τουλάχιστον όσον αφορά την αρχική σελίδα όπου πολλές ειδήσεις συνυπάρχουν, είναι σκόπιμο να επιτρέπουν στις μηχανές αναζήτησης να επιλέγουν ένα απόσπασμα από το αντίστοιχο άρθρο ως περιγραφή, για να μην απωθείται ο χρήστης από άσχετες με το ερώτημά του λέξεις.

4.3.2 Meta ετικέτα ανιχνευτών

Η ετικέτα ανιχνευτών (Meta robots tag) μπορεί να χρησιμοποιηθεί 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, ενώ δίνεται η δυνατότητα ελέγχου εξειδικευμένων ανιχνευτών, όπως συμβαίνει και με το πρωτόκολλο αποκλεισμού robots.

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 ενημερώνει τις μηχανές εάν η σελίδα πρέπει να ανιχνευθεί και να ευρετηριασθεί, με σκοπό την ανάκτηση και επιστροφή της στα αποτελέσματα αναζητήσεων, ή όχι, αντίστοιχα. 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. Υποθέτουν, δηλαδή, ότι μπορούν να ευρετηριάσουν όλες τις σελίδες που περιλαμβάνονται ή προστίθενται στο crawl frontier των ανιχνευτών τους, επομένως η ανάθεση της τιμής "index" είναι, σε γενικές γραμμές, περιττή.

- 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 (όπως, δηλαδή, θα αναλυθεί σε άλλο κεφάλαιο, δεν θα δοθεί καμία αξία στον σύνδεσμο και την σελίδα στην οποία αυτός κατευθύνει). 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. Στο σημείο αυτό τονίζουμε ότι, με την οδηγία "nofollow" σε μία σελίδα, δεν απαγορεύεται εξ ολοκλήρου στις μηχανές αναζήτησης η ευρετηρίαση των ιστοσελίδων στις οποίες οδηγούν οι σύνδεσμοι της σελίδας, απλώς απαγορεύεται η μεταβίβαση των ανιχνευτών προς αυτές της ιστοσελίδες από την "nofollow" σελίδα. Εν ολίγοις, οι ιστοσελίδες αυτές μπορούν να ανακαλυφθούν, να ανιχνευθούν και να ευρετηριασθούν από τις μηχανές αναζήτησης από άλλες σελίδες που συνδέουν προς αυτές.

- 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.

- Nosnippet

Με την ανάθεση της τιμής αυτής, οι διαχειριστές μίας σελίδας απαγορεύουν αποκλειστικά στη μηχανή αναζήτησης της Google να παρουσιάσει οποιαδήποτε περιγραφή της σελίδας, κάτω από τον τίτλο, στα αποτελέσματα αναζήτησης. Η χρησιμότητα της εντολής αυτής αναλύθηκε εκτενώς κατά την ανάλυση της meta ετικέτας περιγραφής. Είναι προφανές ότι η εντολή αυτή προηγείται, σε προτεραιότητα, της meta ετικέτας περιγραφής, επομένως, σε ενδεχόμενη ύπαρξη και των δύο στον <HEAD> τομέα του HTML αρχείου, η Google θα υπακούσει την ετικέτα ανιχνευτών, αφαιρώντας οποιαδήποτε περιγραφή της σελίδας στα αποτελέσματα σχετικής αναζήτησης. Οι υπόλοιπες μηχανές αναζήτησης, μέχρι στιγμής, δεν έχουν ανακοινώσει ότι αναγνωρίζουν την τιμή αυτή κι, επομένως, την αγνοούν.

- 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).

Αντί της τιμής “robots”, στο γνώρισμα name=”” της meta ετικέτας, μπορεί να χρησιμοποιηθεί το όνομα του συγκεκριμένου ανιχνευτή στον οποίο ο διαχειριστής επιθυμεί να απευθυνθεί.

Για παράδειγμα, δίνεται η παρακάτω ακολουθία meta ετικετών:

```
<head>
      <meta name="robots" content="index, nofollow" />
      <meta name="GoogleBot" content="noODP" />
      <meta name="Slurp" content="noYDIR" />
</head>
```

Με τις ετικέτες αυτές, δίνεται η οδηγία σε όλους τους ανιχνευτές να ευρετηριάσουν την παρούσα σελίδα, αλλά να μην ακολουθήσουν τους συνδέσμους που εμπεριέχονται σε αυτήν, ενώ δίνεται σαφής οδηγία στις μηχανές Google και Yahoo! να μην αντικαταστήσουν την περιγραφή της σελίδας, στα αποτελέσματα αναζήτησης, με την περιγραφή του ODP και του Yahoo! καταλόγου, αντίστοιχα. Η τελευταία, μάλιστα, εντολή είναι περιττή, καθώς η τιμή "noYDIR" θα μπορούσε να δοθεί στην πρώτη, εφόσον δεν αναγνωρίζεται από και δεν αφορά τις άλλες μηχανές αναζήτησης.

4.3.3 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, ή τη 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, ενώ στη χειρότερη περίπτωση δεν επηρεάζει καθόλου (αρνητικά) την επίδοση ενός συγκεκριμένου ιστοτόπου. Υπάρχουν, όμως, και περιπτώσεις στις οποίες αυτό δεν είναι επιθυμητό. Με την ανάθεση της τιμής "notranslate" στο εξειδικευμένο γνώρισμα "google" της meta ετικέτας, ο διαχειριστής απαγορεύει στη μηχανή της Google την παροχή αυτού του συνδέσμου μετάφρασης της ιστοσελίδας του. Σύμφωνα με τους μηχανικούς της Google, η χρήση της ετικέτας αυτής δεν επηρεάζει την κατάταξη των ιστοσελίδων θετικά ή αρνητικά.

- **Ετικέτα επαλήθευσης κατοχής Google Webmasters**

```
<head>
      <meta name="verify-v1" content="…" />
</head>
```

Πρόκειται για μία εξειδικευμένη ετικέτα επαλήθευσης κατοχής ενός ιστοτόπου στα εργαλεία για webmasters. Η τιμή του γνωρίσματος content="…" δίνεται από το λογαριασμό Google Webmasters του διαχειριστή του ιστοτόπου. Πρόκειται για τη μόνη meta ετικέτα που να έχει ευαισθησία σε πεζούς και κεφαλαίους χαρακτήρες (case sensitive).

- **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 (και μη προτεινόμενη, συγκριτικά με την ανακατεύθυνση 301, όπως θα αναλυθεί σε επόμενο κεφάλαιο) μορφή ανακατεύθυνσης. Αυτή η μορφή ανακατεύθυνσης δεν υποστηρίζεται από όλους τους φυλλομετρητές και ενδέχεται να μπερδεύει τον χρήστη, βλάπτοντας ενδεχομένως

και την εμπειρία του στην ιστοσελίδα. Σύμφωνα, μάλιστα, με το W3C, χαρακτηρίζεται ως αποδοκιμασμένη τεχνική ανακατεύθυνσης (W3C, 2010).

- **Ετικέτες σύνδεσης ιστοσελίδας με το Facebook**

Σήμερα, ολοένα και περισσότερες εταιρείες, που δραστηριοποιούνται στο Διαδίκτυο, ασχολούνται με ιστοσελίδες κοινωνικής δικτύωσης (social networking). Το πλέον δημοφιλές μέσο 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, όπως ακριβώς με το webmaster central λογαριασμό, για την ορθή χρησιμοποίηση των εργαλείων του Facebook Insights, 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.

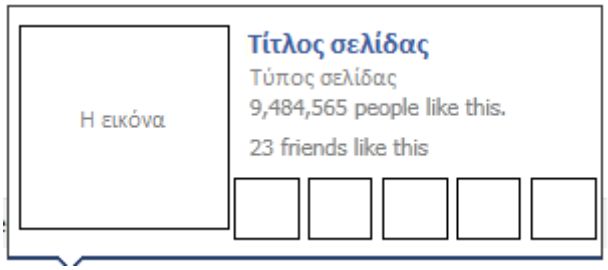
- **Ετικέτες πρωτοκόλλου Open Graph**

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>
```

Όπως φαίνεται στις παραπάνω ετικέτες, όπως και πριν, οφείλει να υιοθετηθεί το αντίστοιχο πρότυπο του Πρωτοκόλλου Open Graph. Στην περίπτωση που δίνεται, ως παράδειγμα, παραπάνω, το αντικείμενο εντός του social network θα φαινόταν ως εξής στα προφίλ των φίλων χρηστών της σελίδας:



4.3.4 Meta ετικέτα λέξεων – κλειδιών

Πρόκειται για μία ετικέτα που, κατά την εμφάνισή της, υποστηρίχθηκε από την πλειοψηφία των σπουδαιότερων τότε μηχανών αναζήτησης, μαζί με τις ετικέτες περιγραφής και ανιχνευτών. Ιδιαίτερη σύγχυση συνοδεύει αυτό το στοιχείο της HTML για περισσότερο από μία δεκαετία, καθώς σε πολλές διαδικτυακές πύλες και πηγές πληροφοριών παρουσιάζεται ως απαραίτητο εργαλείο για τη βελτιστοποίηση των ιστοσελίδων στις μηχανές αναζήτησης.

Η πρώτη επίσημη αναφορά βρίσκεται σε ένα προσχέδιο της IETF (Internet Engineering Task Force) του Δεκέμβριου του 1995, από τον Davide Musella (1995), στο οποίο χρησιμοποιήθηκε η εξής σύνταξη:

```
<head>
      <META HTTP-EQUIV="keywords" CONTENT="Italy Product,
      Italy Tourism" />
</head>
```

Παρατηρούμε ότι ενδεικνυόταν η χρήση του HTTP-EQUIV γνωρίσματος στην ετικέτα, δίνοντας, έτσι, πληροφορίες στο διακομιστή (server) και όχι στις μηχανές αναζήτησης. Επίσης, τα κενά μεταξύ του χαρακτήρα “,” και της λέξης (και ανάποδα) αγνοούνται.

Εκπρόσωποι των μηχανών αναζήτησης συμμετείχαν σε ένα workshop, το Μάιο του 1996, για να πραγματευθούν τα meta δεδομένα, τις πληροφορίες, δηλαδή, που εμπεριέχονταν στις meta ετικέτες. Από τη συνάντηση αυτή, προέκυψε ένα πρότυπο από κοινού αντιμετώπισης των ετικετών ανιχνευτών και περιγραφής, όμως κανένα συμπέρασμα δεν εξήχθη όσον αφορά την ετικέτα λέξεων – κλειδιών, παρόλο που συζητήθηκε.

Παρά τη μη εξαγωγή οδηγιών για την εν λόγω ετικέτα, τόσο η Infoseek όσο και η Altavista (που σήμερα αποτελεί απόκτημα της Yahoo! μηχανής) την υποστήριξαν το 1996, ενθαρρύνοντας μάλιστα τους κατόχους ιστοτόπων να τη χρησιμοποιούν. Η μηχανή αναζήτησης της Inktomi (επίσης απόκτημα της Yahoo!, σήμερα, όπως έχει αναλυθεί στην

ιστορική αναδρομή των μηχανών) επίσης υποστήριξε, με την έναρξη της λειτουργίας της στα τέλη του 1996, την ετικέτα, ενώ σε προσθήκη του στοιχείου στους αλγορίθμους ανίχνευσης και κατάταξης προχώρησε το 1997 και η Lycos.

Αυτή ήταν και η τελευταία χρονιά που η ετικέτα meta keywords tag υποστηριζόταν από την πλειοψηφία των τότε σπουδαιότερων μηχανών αναζήτησης (των 4 από τις 7, καθώς οι Excite, WebCrawler και Northern Light δεν την υποστήριξαν).

Όταν εμφανίσθηκαν η μηχανή αναζήτησης FAST και η κυρίαρχη σήμερα Google, το 1998, δεν υποστήριξαν την ετικέτα καθώς η μέχρι στιγμής εικόνα έδειχνε ότι οι διαχειριστές ιστοχώρων «φούσκωναν» την ετικέτα με τις ίδιες λέξεις ξανά και ξανά, με σκοπό να βρίσκονται υψηλά στα αποτελέσματα σχετικών με τις λέξεις αναζητήσεων. Άλλοι πάλι φρόντιζαν να συσχετίζουν την ιστοσελίδα τους με εξ ολοκλήρου άσχετες λέξεις – κλειδιά, μόνο για να εμφανίζονται σε περισσότερες αναζητήσεις, άρα και περισσότερους χρήστες. Άλλωστε, ήταν πολύ νωρίς για να βασίζονταν οι μηχανές στην ορθή και αποτελεσματική ανάλυση των συνδέσμων, επομένως ο ασφαλής δρόμος ήταν αυτός της αδιαφορίας για την ετικέτα.

Μάλιστα, τον Ιούλιο του 2002, και η AltaVista έπαψε να υποστηρίζει τη χρήση της ετικέτας, αφήνοντας μόνη την Inktomi εκ των κυρίαρχων μηχανών να τη χρησιμοποιεί στους αλγορίθμους της. Έκτοτε, η Inktomi αποκτήθηκε από την Yahoo!, η οποία ακόμη και σήμερα δεν έχει κάνει επίσημη ανακοίνωση για τη σαφή θέση της ως προς το στοιχείο αυτό.

Ο Danny Sullivan (Search Engine Land, 2007) διαπίστωσε πειραματικά ότι η σύγχυση που αναφέρθηκε ότι συνοδεύει τη meta ετικέτα αυτή δικαιολογείται, καθώς από τις τέσσερις, τουλάχιστον, σπουδαιότερες τότε μηχανές αναζήτησης (Google, Yahoo, Microsoft Live και Ask), άλλες δίνουν περιορισμένη βαρύτητα σε αυτήν και άλλες καμία απολύτως σημασία. Συγκεκριμένα, στις 28 Αυγούστου 2007, προσέθεσε την παρακάτω γραμμή στην αρχική σελίδα του ιστοχώρου όπου αρθρογραφεί και που διαχειρίζεται:

```
<meta name="keywords" content="qiskodslajdmnkd,
ddakaieciuaj jkdalladpaoaw, wdaopeqndlkakjad" />
```

Στη συνέχεια, περίμενε να ανανεώσουν οι 4 μηχανές αναζήτησης το ευρετήριό τους, όσον αφορά τη δική του ιστοσελίδα, και στη συνέχεια έκανε αναζητήσεις με τους όρους που είχε αναθέσει στις meta keywords.

Παρατήρησε, λοιπόν, ότι οι Google και Microsoft Live (νυν Bing), δεν εμφάνιζαν την ιστοσελίδα του στα αποτελέσματα για τους παραπάνω όρους αναζήτησης, μετά την

ανανέωση του ευρετηρίου τους, ενώ οι Yahoo και Ask την εμφάνιζαν στα αποτελέσματα της αναζήτησης. Απέδειξε, επομένως, ότι η meta ετικέτα αυτή σίγουρα δεν έχει κάποια βαρύτητα στον αλγόριθμο κατάταξης των μηχανών της Google και της Microsoft, ενώ χρησιμοποιείται από τις μηχανές Yahoo και Ask για την **ανάκτηση** και παρουσίαση της ιστοσελίδας στα αποτελέσματα αναζήτησης. Οι δύο τελευταίες, δηλαδή, μηχανές αναζήτησης χρησιμοποιούν τη meta ετικέτα λέξεων – κλειδιών για να αποφανθούν εάν μία ιστοσελίδα είναι στοιχειωδώς σχετική με κάποια λέξη ή φράση, αντιμετωπίζοντας τους όρους που εμπεριέχονται στην ετικέτα ως λέξεις (ή κάτι λιγότερο) που εμφανίζονται στο περιεχόμενο της σελίδας και τίποτα παραπάνω, χωρίς να δίνουν, δηλαδή, κάποια βαρύτητα στην ετικέτα αυτή.

Φυσικά, ακόμη και αυτό το συμπέρασμα δε μπορεί να είναι απόλυτο, καθώς το πείραμα ήταν απλοϊκό και εξαρτάται από παράγοντες που δε μετρήθηκαν ούτε προσδιορίσθηκαν. Υπάρχει ακόμη και σήμερα μεγάλη διαμάχη σχετικά με τη βαρύτητα και αξία της ετικέτας αυτής (μάλιστα, σε αναζήτηση οποιουδήποτε εκ των τεσσάρων όρων του πειράματος του Danny Sullivan στις μηχανές αναζήτησης, ως πρώτο αποτέλεσμα επιστρέφει ο αντίστοιχος από τους τέσσερις ιστοχώρους που μία εταιρεία έχει δημιουργήσει, διαψεύδοντας το πείραμα και τον ειδικό), αλλά τελευταία όλοι συγκλίνουν στη μη χρησιμότητα της ετικέτας.

Σύνταξη

Όπως παρουσιάζεται και στο προσχέδιο της IETF, η σύνταξη της meta ετικέτας λέξεων – κλειδιών, τροποποιημένη για να απευθύνεται στις μηχανές αναζήτησης και όχι τον εξυπηρετητή, φαίνεται παρακάτω:

```
<head>
      <META NAME="keywords" CONTENT="…, …, …, …, …, …" />
</head>
```

Καλύτερες πρακτικές

Παρότι, όπως υποδεικνύεται παραπάνω, δεν αποτελεί παράγοντα τεχνικής βελτιστοποίησης των ιστοσελίδων, η ετικέτα των λέξεων – κλειδιών μπορεί να χρησιμοποιηθεί για τους εξής λόγους:

α) Όπως φάνηκε, ορισμένες μηχανές αναζήτησης αξιοποιούν την ετικέτα αυτή, αντιμετωπίζοντάς την ως κείμενο εντός του περιεχομένου μίας σελίδας. Όμως, πρόκειται για ένα κείμενο που ο μέσος χρήστης δεν πρόκειται να διαβάσει, καθώς δεν θα ενδιαφερθεί να

αποκτήσει πρόσβαση στον πηγαίο κώδικα της σελίδας, αλλά και για ένα κείμενο που δεν αποκρύπτεται με τεχνικές CSS από τους επισκέπτες, μη εγείροντας ζητήματα ηθικής (black – hat SEO) στις μηχανές αναζήτησης. Επομένως, η ετικέτα μπορεί να αξιοποιηθεί, προσθέτοντας σε αυτήν λέξεις ή φράσεις για την αναζήτηση των οποίων θέλουμε η ιστοσελίδα μας να εμφανίζεται στα αποτελέσματα, όπως ανορθογραφίες ή συνώνυμα των βασικών όρων – κλειδιών με τα οποία σχετίζεται η σελίδα μας. Με τον τρόπο αυτό, φροντίζεται η εμφάνιση της ιστοσελίδας μας σε ορισμένες μηχανές αναζήτησης για κάποιον όρο γραμμένο λανθασμένα εκ παραδρομής ή αγνοίας, καθώς και κάποιον όρο που δεν προβλέψαμε να συμπεριλάβουμε στο ίδιο το περιεχόμενο και σχετίζεται εννοιολογικά με τους επιθυμητούς όρους.

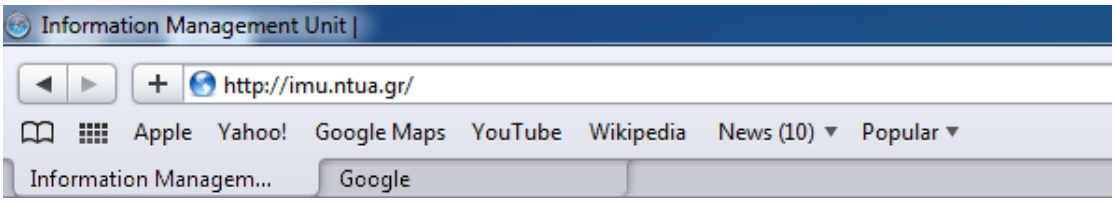
β) Η χρήση της ετικέτας για τους παραπάνω λόγους καθίσταται αναγκαία σε περίπτωση που η ιστοσελίδα είναι κατασκευασμένη με τεχνολογία flash, επομένως χωρίς ιδιαίτερο περιεχόμενο ορατό στις μηχανές αναζήτησης. Η περίπτωση αυτή θα εξεταστεί σε επόμενο κεφάλαιο διεξοδικά.

4.4 Ετικέτες σήμανσης περιεχομένου

4.4.1 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:

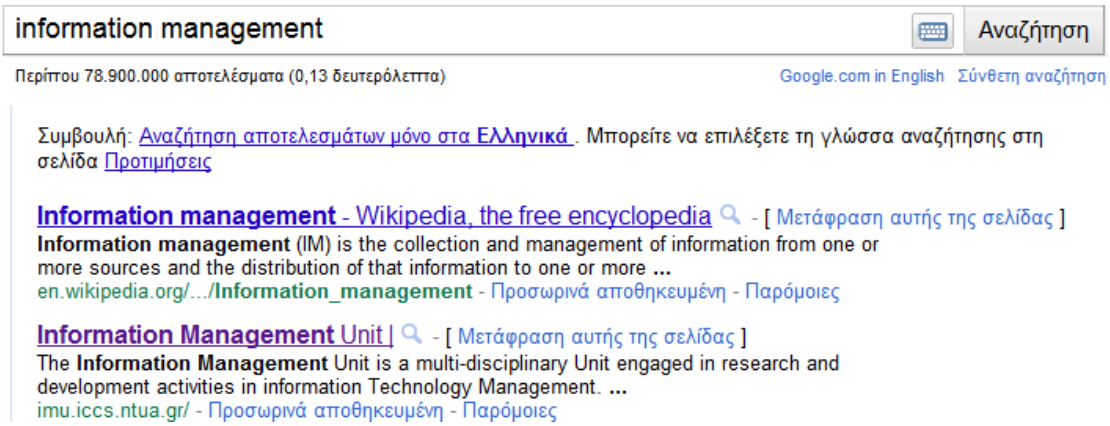
α) Στην κορυφή του φυλλομετρητή του χρήστη, καθώς και ως όνομα της αντίστοιχης καρτέλας του φυλλομετρητή. Έρευνες έχουν δείξει, όμως, ότι ο χρήστης δίνει περιορισμένη προσοχή στα σημεία εκείνα, παρ' όλα αυτά διευκολύνει την εμπειρία του κατά τη διαχείριση πολλαπλών καρτελών.



Εικόνα 7 Εμφάνιση του τίτλου σελίδας στο φυλλομετρητή

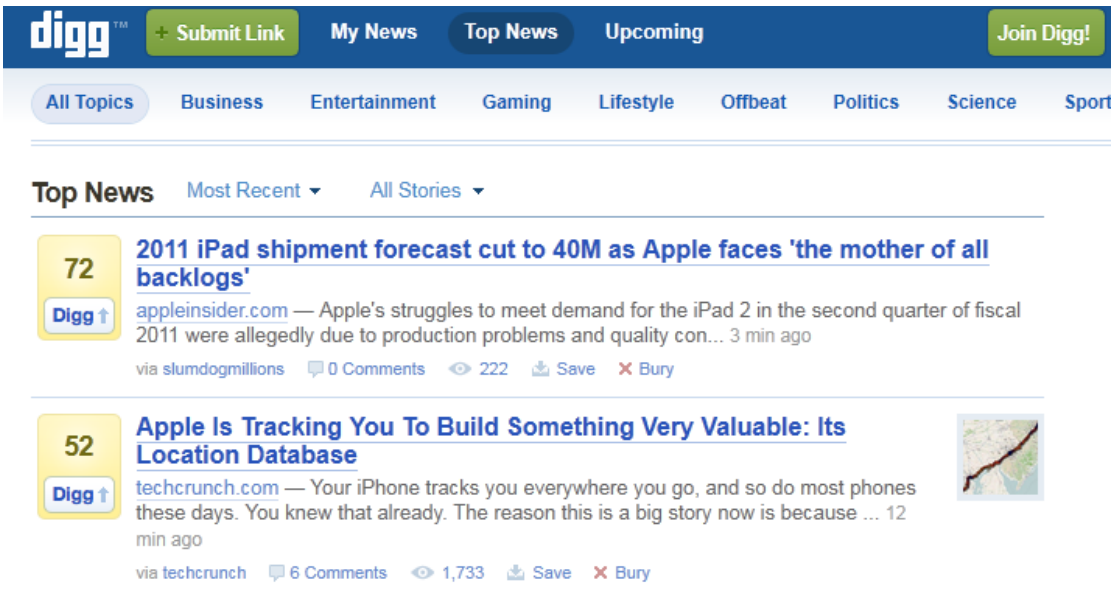
β) Στα αποτελέσματα των μηχανών αναζήτησης, δίνοντας μία πρώτη συνοπτική αλλά σαφή εικόνα του περιεχομένου προτού ο χρήστης εισέλθει στο έγγραφο αυτό. Μάλιστα, σε

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 8 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 9 Εμφάνιση τίτλου σελίδας στο anchor text ορισμένων συνδέσμων

Σύνταξη

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

```
<head>
      <title>Information Management Unit |</title>
</head>
```

Καλύτερες πρακτικές

α) Ο τίτλος της σελίδας συνίσταται, ορισμένες φορές, να αποτελείται από ένα αποτελεσματικό call-to-action μήνυμα, αντί της ακριβούς συνοπτικής περιγραφής της σελίδας με φράσεις – κλειδιά που αντιπροσωπεύουν το περιεχόμενο της σελίδας. Φυσικά, υπάρχει ένα trade-off μεταξύ των δύο πρακτικών, καθώς η πρώτη είναι πιο ελκυστική στο μάτι του χρήστη, εκμαιεύοντας περισσότερες επισκέψεις και, κατ' επέκταση, υποδεικνύοντας στις μηχανές αναζήτησης ότι το αντίστοιχο έγγραφο είναι σχετικό με το εκάστοτε ερώτημα, ενώ η δεύτερη πρακτική είναι πιο φιλική στους αλγορίθμους των μηχανών αναζήτησης.

β) Έχει παρατηρηθεί ότι οι ιστοσελίδες των οποίων ο τίτλος περιλαμβάνει λέξεις και φράσεις κλειδιά στην αρχή (στ' αριστερά, δηλαδή), κι ενδεχομένως το όνομα του ιστοτόπου ή το όνομα κάποιου διαφημιζόμενου προϊόντος προς το τέλος (στα δεξιά) έχουν ευνοϊκότερη αντιμετώπιση από τους αλγορίθμους των μηχανών αναζήτησης, καθώς ο ενδιαφέρων όρος είναι ο πρώτος που θα συναντήσουν κατά την προσπέλαση και ανίχνευση, ενώ, παράλληλα, επιτυγχάνεται και προώθηση του προϊόντος ή της ιστοσελίδας καθαυτής.

γ) Μεγάλη προσοχή πρέπει να δοθεί στην ανάθεση μοναδικών τίτλων σε κάθε σελίδα ενός ιστοτόπου. Ειδικά στην περίπτωση των συστημάτων διαχείρισης περιεχομένου (CMS), όπου πολλές παράμετροι παράγονται δυναμικά, είναι σημαντικό να ανατεθούν διαφορετικοί τίτλοι σε κάθε σελίδα, καθώς ο τίτλος είναι το πρώτο στοιχείο που τα φίλτρα διπλότυπου περιεχομένου των μηχανών θα εξετάσουν. Σημειώνεται εδώ ότι, όπως θα αναφερθεί σε άλλο κεφάλαιο, οι μηχανές αναζήτησης τιμωρούν ιστοσελίδες με ποινές (χαμηλότερη θέση στα αποτελέσματα, αφαίρεση από το ευρετήριο), σε περίπτωση που διαπιστωθεί αυτούσια αναπαραγωγή σελίδων, ακόμη κι αν πρόκειται για σελίδες στον ίδιο ιστοχώρο.

δ) Τέλος, πρέπει να ληφθεί σοβαρά υπόψη το γεγονός ότι οι μηχανές αναζήτησης, στα αποτελέσματα, δεν αναγράφουν περισσότερους από 70 χαρακτήρες (για περισσότερους χαρακτήρες, ο τίτλος κόβεται με το σύμβολο «...» στο τέλος, δημιουργώντας μια ελλιπή εικόνα και εντύπωση στο χρήστη), ενώ στην αγορά του Search Engine Optimization προτείνεται οι λέξεις του τίτλου να μην είναι περισσότερες από 15, για να αποφεύγεται η

μειωμένη πυκνότητα των λέξεων – κλειδιών, καθώς και η κατάχρηση του στοιχείου με άσκοπη καταγραφή ή κι επανάληψη όρων (keyword stuffing).

4.4.2 Επικεφαλίδες

Πάρα πολλοί διαχειριστές και συντάκτες ιστοσελίδων στο Διαδίκτυο αγνοούν την ύπαρξη και την αξία των πρότυπων επικεφαλίδων (h1, h2, h3 headings) και είτε δε 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.

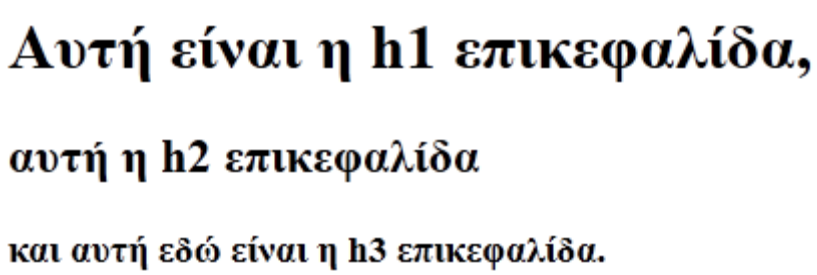
Σημειώνεται, εδώ, ότι ορισμένα συστήματα διαχείρισης περιεχομένου (CMSs) δεν υποστηρίζουν πλήρως τη χρήση επικεφαλίδων, καθώς μεταφράζουν τον υποτιτλισμό σε κείμενο με διαφοροποιημένη, πιο έντονη κι ευδιάκριτη γραμματοσειρά, ή αυτομάτως μεταφέρουν το περιεχόμενο της ετικέτας τίτλου της σελίδας στην ετικέτα 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).

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 10 Οι διάφορες επικεφαλίδες στο φυλλομετρητή

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

Για παράδειγμα, έστω το παρακάτω κομμάτι κώδικα:

```
<html>

	<head>
		<title>Gaming</title>
	</head>

	<body>
		<h1>Παιχνίδια</h1>
			<h2>Παιχνίδια για τον υπολογιστή</h2>
				<h3>Ιστορία</h3>
			<h2>Παιχνίδια για κονσόλες</h2>
				<h3>PlayStation</h2>
				<h3>Dreamcast</h2>
	</body>

</html>
```

Υποθέτουμε ότι οι πληροφορίες κάθε υποκατηγορίας παρεμβάλλονται μεταξύ των επικεφαλίδων, για να έχει νόημα η ιστοσελίδα και η παραπάνω δομή.

Είναι φανερό ότι η εν λόγω σελίδα αφορά τα παιχνίδια για υπολογιστή και κονσόλες, ενώ κάνει μία ιστορική αναδρομή για τη μακρά πορεία των τίτλων παιχνιδιών για Η/Υ, καθώς και ένα διαχωρισμό των παιχνιδιών για κονσόλες, ανάλογα με την κονσόλα.

Οι μηχανές αναζήτησης, όμως, θα δώσουν μεγαλύτερη σημασία στον τίτλο της σελίδας (Gaming) και τις h1 επικεφαλίδες (Παιχνίδια), προτού θεωρήσουν ότι το έγγραφο αφορά, για παράδειγμα, την Ιστορία ή το PlayStation γενικότερα.

Καλύτερες πρακτικές

α) Όπως ήδη αναφέρθηκε, οι ετικέτες επικεφαλίδων έχουν μεγαλύτερη αξία, από την σκοπιά των μηχανών αναζήτησης, σε σχέση με το απλό κείμενο ενός εγγράφου. Επομένως, είναι προτιμότερο να χωριστεί το κείμενο σε υποενότητες, όσο σχετικές κι αν είναι αυτές μεταξύ τους, δίνοντας σε αυτές διαφορετικές, κάθε φορά, επικεφαλίδες, όλες σχετικές, όμως, με την φράση ή λέξη – κλειδί που στοχεύεται.

β) Αξίζει να σημειωθεί ότι μόνο οι πρότυπες επικεφαλίδες <h1>, <h2>, <h3> απολαμβάνουν μεγαλύτερη αξία κατά την ευρετηρίαση ενός εγγράφου, δίνοντας μεγαλύτερη αξία στο ίδιο το έγγραφο για ορισμένες αναζητήσεις και όχι επικεφαλίδες που δημιουργούνται με το CSS αρχείο. Για τη βελτίωση της εμφάνισης του περιεχομένου, όμως, συνίσταται η επεξεργασία της γραμματοσειράς που αντιστοιχεί στο αναγνωριστικό των <h1>, <h2> και <h3> στοιχείων από το CSS αρχείο, αρκεί να πραγματοποιείται η χρήση των επικεφαλίδων αυτών καθαυτών.

γ) Δεδομένης της σειράς προτεραιότητας των ετικετών <h1>, <h2> και <h3>, ένα έγγραφο, εφόσον περιέχει επικεφαλίδες, πρέπει να περιέχει οπωσδήποτε <h1> επικεφαλίδες. Εάν οι ενότητες διαιρούνται σε πολλαπλές υποενότητες, τότε μπορούν να χρησιμοποιηθούν και οι

υπόλοιπες ετικέτες επικεφαλίδων. Δεν έχει νόημα η χρήση της <h3> εάν δε χρησιμοποιείται η <h2>, ή η χρήση της <h2> εάν δε χρησιμοποιείται η <h1> ετικέτα.

δ) Τέλος, μία πολύ αποτελεσματική μέθοδος ανάθεσης αξίας στην αρχική σελίδα, όσο και στους εσωτερικούς συνδέσμους, που ακολουθείται κυρίως από τις μεγαλύτερες ειδησεογραφικές ιστοσελίδες και διαδικτυακές πύλες είναι η χρήση εσωτερικών συνδέσμων ως επικεφαλίδες, εμφωλεύοντας συνδέσμους μέσα στις επικεφαλίδες, ως εξής:

```
<h1><a href="link_url">Περιεχόμενο επικεφαλίδας</a></h1>
```

4.4.3 Μορφοποίηση κειμένου

Παρότι οι διάφορες γραμματοσειρές, σε μία σελίδα, έχουν την ίδια βαρύτητα, υπάρχουν πολλές ετικέτες μορφοποίησης του κειμένου που διαφοροποιούν ορισμένες λέξεις ή φράσεις από τις υπόλοιπες, όπως η έντονη γραφή, η πλάγια γραφή και η υπογράμμιση. Τα βασικά στοιχεία έμφασης κειμένου είναι τα <b> και για έντονη γραφή (bold) και <i> και για πλάγια γραφή (italic).

Σύμφωνα, όμως, με τα πρότυπα της W3C.org (W3C, 2000), τα οποία υιοθετούνται και από τις μηχανές αναζήτησης για την ανάθεση βαρύτητας στο περιεχόμενο των ιστοσελίδων και οφείλουν να υιοθετούνται και από τους διαχειριστές ιστοτόπων, παρότι το οπτικό αποτέλεσμα από τη χρήση των <b> και ή <i> και είναι το ίδιο, για τη βελτίωση της εικόνας του κειμένου οφείλουν να χρησιμοποιούνται οι ετικέτες <b> και <i>, ενώ για την έμφαση και την έντονη έμφαση του κειμένου οι ετικέτες και , αντίστοιχα.

Έτσι, στο παρακάτω παράδειγμα όπου χρησιμοποιούνται και οι τέσσερις ετικέτες για τη μορφοποίηση του κειμένου, οι λέξεις έχουν διαφορετική βαρύτητα:

```
<body>
      <p>Το αντικείμενο του <strong>Ηλεκτρολόγου
Μηχανικού</strong> και <b>Μηχανικού Υπολογιστών</b>
συνδυάζει με γόνιμο τρόπο ένα ευρύ σύνολο περιοχών της
επιστήμης και της τεχνολογίας, όπως η
<em>πληροφορική</em>, οι <i>τηλεπικοινωνίες</i>, η
<i>ηλεκτρονική</i>, ο <i>αυτόματος έλεγχος</i> και η
<em>ενέργεια</em>. Η τεχνολογική επανάσταση που
συντελείται στις μέρες μας βασίζεται κατά μεγάλο μέρος στη
δημιουργική συνεισφορά των ηλεκτρολόγων μηχανικών και
μηχανικών υπολογιστών. </p>

</body>
```

Στο φυλλομετρητή, παρουσιάζεται το παρακάτω κείμενο:

Το αντικείμενο του **Ηλεκτρολόγου Μηχανικού** και **Μηχανικού Υπολογιστών** συνδυάζει με γόνιμο τρόπο ένα ευρύ σύνολο περιοχών της επιστήμης και της τεχνολογίας, όπως η *πληροφορική*, οι *τηλεπικοινωνίες*, η *ηλεκτρονική*, ο *αυτόματος έλεγχος* και η *ενέργεια*. Η τεχνολογική επανάσταση που συντελείται στις μέρες μας βασίζεται κατά μεγάλο μέρος στη δημιουργική συνεισφορά των ηλεκτρολόγων μηχανικών και μηχανικών υπολογιστών.

Εικόνα 11 Η μορφοποίηση του κειμένου ως σήμανση βαρύτητας των λέξεων

Το παραγόμενο αποτέλεσμα φαίνεται οπτικά να δίνει έμφαση και στις δύο πτυχές του τίτλου του μηχανικού, όπως επίσης και στα πέντε διαφορετικά πεδία του ηλεκτρολόγου μηχανικού και μηχανικού υπολογιστών. Όσον αφορά τις μηχανές αναζήτησης, όμως, έμφαση δίνεται μόνο στην ενέργεια και την πληροφορική, ενώ ιδιαίτερη έμφαση δίνεται στο αντικείμενο του ηλεκτρολόγου μηχανικού. Μάλιστα, τα λογισμικά μετατροπής κειμένου σε φωνή (ευρέως χρησιμοποιούμενα από άτομα με ειδικές ανάγκες) δίνουν αντίστοιχη έμφαση στις ετικέτες και , αλλάζοντας τον τόνο και την ένταση της φωνής κατά την ανάγνωση, ενώ τις λέξεις που μορφοποιούνται με τις ετικέτες <b> και <i> τις διαβάζουν με τον ίδιο τόνο που διαβάζουν και το υπόλοιπο κείμενο.

Για λόγους διατήρησης υψηλής ποιότητας του οπτικού αποτελέσματος και της εμπειρίας του χρήστη αλλά και αποφυγής διαμοιρασμού της βαρύτητας σε πολλαπλές λέξεις, πρέπει να αποφεύγεται η κατάχρηση των ετικετών μορφοποίησης.

4.4.4 Σύνδεσμοι (links)

Οι υπερσύνδεσμοι μίας ιστοσελίδας αποτελούν την κινητήρια δύναμη της κατάταξης των ιστοσελίδων στα αποτελέσματα αναζήτησης, όπως εξηγήθηκε κατά την ανάλυση της λειτουργίας των μηχανών αναζήτησης, και διαχωρίζονται σε εισερχόμενους (inbound) και εξερχόμενους (outbound), με κριτήριο την κατεύθυνση της σύνδεσης μεταξύ της υπό εξέταση ιστοσελίδας και κάποιας ιστοσελίδας στο Διαδίκτυο. 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, όπως αναλύεται στο επόμενο κεφάλαιο, 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.

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="link url" title="Περιγραφή">Anchor Text</a>
```

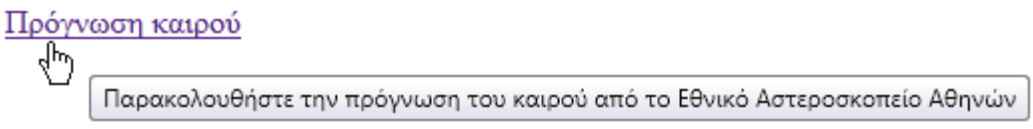
Τέλος, όσον αφορά τους εξωτερικούς εξερχόμενους συνδέσμους και όπως θα αναλυθεί στο επόμενο κεφάλαιο εκτενέστερα, έχουμε τη δυνατότητα να αποκλείσουμε την επίσκεψη των ανιχνευτών μέσω αυτών και να επιτρέψουμε μόνο την ανακατεύθυνση των χρηστών, με την παρακάτω σύνταξη:

```
<a href="link url" 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.

Καλύτερες Πρακτικές

α) Είναι προφανές ότι το anchor text θα πρέπει να περιλαμβάνει μία αποτελεσματική φράση – κλειδί που να περιγράφει ικανοποιητικά την σελίδα στην οποία αναφέρεται, παρά κάτι άσχετο ή τελείως ουδέτερο όπως «Πατήστε εδώ» ή «Περισσότερα..». Παράλληλα, καλό είναι να συμπεριλαμβάνεται και ένα αποτελεσματικό call-to-action, για να προσελκύει τον χρήστη και να τον ενθαρρύνει να προχωρήσει σε κάποια εγγραφή, αγορά ή συνέχεια της πλοήγησής του στον ιστότοπο. Τονίζεται, βέβαια, ότι μεγάλο ρόλο για τις μηχανές αναζήτησης παίζει όλο το κείμενο που πλαισιώνει έναν σύνδεσμο, μαζί με τις γειτονικές λέξεις ή φράσεις, την τιμή του "title=" γνωρίσματος και το συνολικό περιεχόμενο της σελίδας. Η τιμή αυτή του τελευταίου γνωρίσματος είναι ορατή στον ίδιο τον επισκέπτη, όπως φαίνεται εδώ:



Εικόνα 12 Το anchor text ή ο τίτλος καλεί τον χρήστη να δράσει

β) Στην περίπτωση των εξερχόμενων εξωτερικών συνδέσμων, η κατεύθυνση του χρήστη προς μία σελίδα διαφορετικού ιστοτόπου ενδέχεται να διώξει τον χρήστη μόνιμα από τη δική μας ιστοσελίδα (ενώ, εάν η σελίδα εκείνη λειτουργεί με ανακατεύθυνση του επισκέπτη σε επόμενη σελίδα, ο επισκέπτης δε θα μπορεί να επιστρέψει στη δική μας ιστοσελίδα). Για τον σκοπό αυτό, συνίσταται η χρήση του γνωρίσματος "target=", ακολουθούμενο από την τιμή "_blank", μέσα στην ετικέτα υπερσυνδέσμου <a href>. Το στοιχείο αυτό κατευθύνει τον επισκέπτη στον εξωτερικό σύνδεσμο, ανοίγοντας, όμως, νέο παράθυρο ή νέα καρτέλα στον φυλλομετρητή (ανάλογα με τις δυνατότητες του τελευταίου). Η σύνταξη του στοιχείου αυτού φαίνεται στον κώδικα του προηγούμενου παραδείγματος:

```
<a href="http://www.meteo.gr" title="Παρακολουθήστε την πρόγνωση του καιρού από το Εθνικό Αστεροσκοπείο Αθηνών" rel="nofollow" target="_blank">Πρόγνωση καιρού</a>
```

γ) Τέλος, καλό είναι να αξιοποιούνται οι δυνατότητες μορφοποίησης, μέσω του αρχείου CSS, για την βελτιστοποίηση της εμφάνισης των συνδέσμων και την αρμονική ενσωμάτωσή τους (σε χρώμα, μέγεθος, γραμματοσειρά) στο έγγραφο στο οποίο ανήκουν.

4.4.5 Εικόνες

Όπως και με τους συνδέσμους, υπάρχουν ορισμένα γνωρίσματα της ετικέτας <img> τα οποία χαρακτηρίζουν και περιγράφουν τις εικόνες στις μηχανές αναζήτησης, βελτιστοποιώντας όχι μόνο την ευρετηρίαση των εικόνων καθαυτών αλλά και την επίδοση της ιστοσελίδας στα αποτελέσματα αναζήτησης για ορισμένους όρους.

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=”. Το γνώρισμα “title=” έχει ακριβώς την ίδια λειτουργία με αυτή του ίδιου γνωρίσματος στην ετικέτα των συνδέσμων (a href), ενώ το γνώρισμα “alt=” περιγράφει πιο αναλυτικά και κυριολεκτικά το περιεχόμενο της εικόνας. Η τιμή του γνωρίσματος αυτού εμφανίζεται αντί της εικόνας, μέχρι να φορτωθεί πλήρως η εικόνα στον φυλλομετρητή, ενώ αναγιγνώσκεται από αντίστοιχο λογισμικό (screen reader) στα άτομα με ειδικές ανάγκες, όπως ακριβώς και οι λέξεις που μορφοποιούνται με τις ετικέτες και . Η σύνταξη των δύο αυτών στοιχείων φαίνεται στο παρακάτω παράδειγμα:

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

Καλύτερες πρακτικές

α) Όπως και με τις επικεφαλίδες h1, h2, h3, πολύ αποτελεσματική τεχνική έχει αποδειχθεί η χρήση εικόνων – υπερσυνδέσμων. Πρόκειται για εικόνες που συνδέουν προς κάποια ιστοσελίδα, ή την διεύθυνση URL της ίδιας της εικόνας σε μεγαλύτερες (πραγματικές) διαστάσεις. Η σύνταξη που ακολουθείται είναι η εξής:

```
<a href="link_url" title="Περιγραφή συνδέσμου"><img
src="image_url" title="Περιγραφή εικόνας" alt="Εναλλακτικό κείμενο
εικόνας"></a>
```

Στην περίπτωση αυτή, όπως φαίνεται στο παραπάνω παράδειγμα, υπάρχει «σύγχυση» μεταξύ των δύο γνωρισμάτων “title=” στην εικόνα και τη διεύθυνση URL, καθώς οι φυλλομετρητές επιλέγουν μόνο τη μία εκ των δύο τιμών για την περιγραφή της εικόνας στο χρήστη. Για παράδειγμα, έχουμε το ακόλουθο κομμάτι κώδικα:

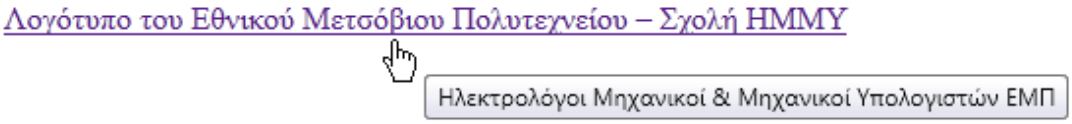
```
<a href="http://www.ece.ntua.gr" title="ΣΗΜΜΥ ΕΜΠ"><img
src="ntua.jpg" title="Ηλεκτρολόγοι Μηχανικοί & Μηχανικοί
Υπολογιστών ΕΜΠ" alt="Λογότυπο του Εθνικού Μετσόβιου Πολυτεχνείου
- Σχολή ΗΜΜΥ"></a>
```

Το αποτέλεσμα στο φυλλομετρητή θα είναι το εξής:

Εικόνα 13 Προβολή του τίτλου των εικόνων και για τον χρήστη

Όπως φαίνεται παραπάνω, ο φυλλομετρητής επιλέγει τον τίτλο της εικόνας και όχι του συνδέσμου για την περιγραφή στο χρήστη. Ο τίτλος του συνδέσμου είναι φαινομενικά άχρηστος, όμως ο ανιχνευτής, κατά την ευρετηρίαση, θα διαβάσει και τον τίτλο του συνδέσμου. Επομένως, η ανάθεση τιμών εναλλακτικού κειμένου και τίτλου, τόσο στην εικόνα όσο και τον σύνδεσμο, αποτελεί ιδιαίτερα αποτελεσματική τεχνική βελτιστοποίησης.

Η αξία του εναλλακτικού κειμένου για το χρήστη φαίνεται στο ίδιο παράδειγμα, όταν ο φυλλομετρητής δε θα έχει ακόμη φορτώσει την εικόνα:

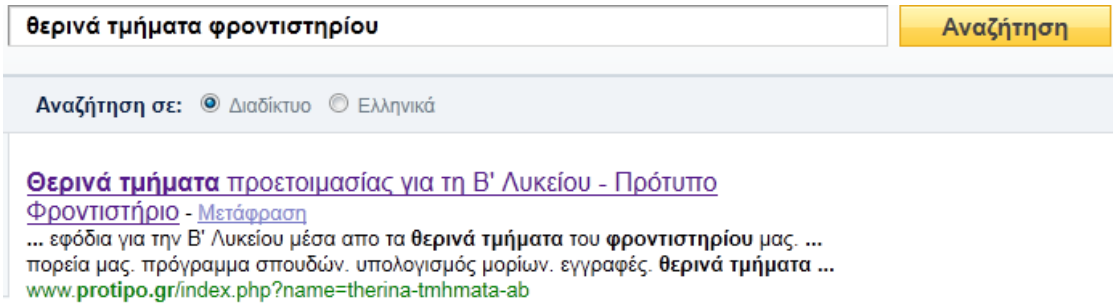


Εικόνα 14 Προβολή του εναλλακτικού κειμένου πριν τη φόρτωση της εικόνας

4.5 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.

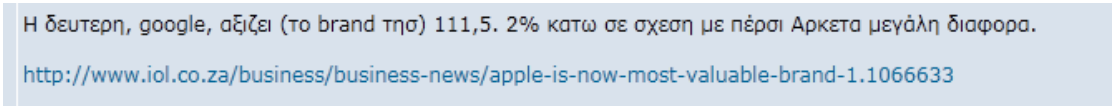


Εικόνα 15 Παράδειγμα εμφάνισης της δομής URL στα αποτελέσματα αναζήτησης

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 16 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 17 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.

Μετατροπή δυναμικών URLs σε στατικές

Οι περισσότερες ιστοσελίδες σήμερα παρουσιάζουν το περιεχόμενο δυναμικά από κάποια βάση δεδομένων. Συνήθη παραδείγματα αποτελούν οι εμπορικοί ή ειδησεογραφικοί ιστοχώροι, οι οποίοι περιλαμβάνουν σελίδες καταλόγων, προϊόντων, κατηγοριών, καλαθιού αγορών ή άρθρων. Παράλληλα, με την επέκταση των Συστημάτων Διαχείρισης Περιεχομένου (CMS – Content Management Systems), ολοένα και περισσότεροι ιστότοποι παράγουν περιεχόμενο δυναμικά. Οι δυναμικές σελίδες αναγνωρίζονται από τον χαρακτήρα "?" της διεύθυνσης URL που ακολουθεί την επέκταση τύπου του αρχείου (π.χ. index.php?...).

Οι σπουδαιότερες μηχανές αναζήτησης ανιχνεύουν και ευρετηριάζουν μία σελίδα, η διεύθυνση URL της οποίας δεν περιλαμβάνει περισσότερες από δύο παραμέτρους, οι οποίες διαχωρίζονται από τον χαρακτήρα "&" (Sisson, 2006). Στην περίπτωση αυτή, οι μηχανές ανιχνεύουν σπάνια τις δυναμικά παραγόμενες σελίδες και με πολύ αργούς ρυθμούς. Ο λόγος για τον οποίο συμβαίνει αυτό είναι ότι οι ανιχνευτές έχουν προγραμματιστεί να αποφεύγουν να εγκλωβίζονται σε ατέρμονες βρόχους, στην προσπάθεια ευρετηρίασης εκατοντάδων χιλιάδων πιθανών σελίδων. Για παράδειγμα, η διεύθυνση http://www.website.com/index.php?categ=4&product=62 αποτελείται από δύο παραμέτρους, την categ και την product.

Συγκεκριμένα, οι ανιχνευτές αποφεύγουν να ευρετηριάσουν σελίδες που περιλαμβάνουν την ακολουθία "id=" μέσα στη διεύθυνση URL, στην προσπάθεια αποφυγής ευρετηρίασης σελίδων που παράγονται με αναγνωριστικά συνόδων (Session IDs). Πρόκειται για αναγνωριστικά που παράγονται σε κάθε επισκέπτη και η διαφορά τους με τα αναγνωριστικά χρηστών (στην περίπτωση ιστοσελίδων που απαιτούν ή επιτρέπουν την εγγραφή του χρήστη) έγκειται στο γεγονός ότι αυτά δε λειτουργούν με εγγραφή και σύνδεση κάποιου χρήστη στην ιστοσελίδα αλλά, μετά από δεδομένο χρονικό διάστημα ή κάποια αλλαγή της διεύθυνσης IP του χρήστη, λήγουν για κάθε επισκέπτη και δημιουργούνται επόμενα. Ένας σωστά δομημένος και τεχνικά βελτιστοποιημένος ιστότοπος οφείλει να αποτρέπει την ευρετηρίαση σελίδων με αναγνωριστικά συνόδων, καθώς αυτό, σε πιθανή επιστροφή των σελίδων στα αποτελέσματα αναζήτησης, τείνει να οδηγήσει σε ανύπαρκτη σελίδα ή σφάλμα 404 (καθώς, μεταξύ της δημιουργίας και ευρετηρίασης της σελίδας έως τη στιγμή της ευρετηρίασης, της αναζήτησης και της επίσκεψης στη σελίδα, το session ID θα έχει ήδη λήξει). Επομένως, καλό είναι να αποφεύγονται παράμετροι τύπου "id", "rid", "uid", "pid", id1", καθώς οι διευθύνσεις URL που τις περιλαμβάνουν τείνουν να αποφεύγονται, ανεξάρτητα από το αν πρόκειται για αναγνωριστικά συνόδου ή για απλές παραμέτρους δυναμικής κατηγοριοποίησης των σελίδων.

Η πιο προχωρημένη σχετική τεχνική περιλαμβάνει την εγκατάσταση μίας δέσμης ενεργειών (script) στο διακομιστή που μετατρέπει τη δυναμική διεύθυνση URL σε στατική σελίδα,

μεταφράζοντας κάθε παράμετρό της σε όνομα φακέλου ή υποφακέλου. Η μέθοδος διαφέρει ανάλογα με το λειτουργικό σύστημα του διακομιστή (Apache ή Windows).

Στην σχετική με το θέμα της επανασυγγραφής διευθύνσεων ιστοσελίδα της Apache (Apache, 2004), αναλύεται διεξοδικά η χρήση και διαχείριση της μονάδας mod_rewrite για τη σύνταξη του κώδικα στατικοποίησης των σελίδων ενός ιστοτόπου, σε επίπεδο διακομιστή που βασίζεται σε λειτουργικό σύστημα Linux.

Αντίστοιχα, πληροφορίες για την ανάλογη διαδικασία σε ιστοσελίδα που φιλοξενείται σε διακομιστή Windows, με το εργαλείο επανασυγγραφής διευθύνσεων ISAPI_Rewrite, παρέχονται στην αντίστοιχη ιστοσελίδα της Microsoft (Microsoft, 2004).

Καλύτερες πρακτικές

α) Το πιο σημαντικό στοιχείο που οφείλει να χαρακτηρίζει μία διεύθυνση URL είναι η περιγραφή του περιεχομένου της. Το ζητούμενο είναι ο χρήστης να μπορεί να κοιτάζει τη μπάρα διευθύνσεων (ή κάποιον σκέτο σύνδεσμο) και να καταλαβαίνει ακριβώς το περιεχόμενο της σελίδας προτού επισκεφθεί το σύνδεσμο. Άλλωστε οι URLs επικολλούνται, διαμοιράζονται, στέλνονται μέσω ηλεκτρονικής αλληλογραφίας, καταγράφονται ως πηγές σε ειδησεογραφικά blogs, φόρουμ, διαδικτυακές πύλες και κοινότητες και, φυσικά, αναγνωρίζονται από τις μηχανές αναζήτησης. Παράλληλα, δεδομένου ότι κάθε σελίδα στοχεύει, συνήθως, σε μία ή δύο συγκεκριμένες λέξεις – κλειδιά, είναι κρίσιμο να συμπεριλαμβάνονται στη διεύθυνση URL αυτής και να ενσωματώνονται ομαλά, χωρίς κατάχρηση, στην περιγραφή του περιεχομένου.

β) Εξίσου σημαντική πρακτική τεχνικής βελτιστοποίησης των διευθύνσεων URL είναι ο περιορισμός των φακέλων και υποφακέλων που βρίσκονται μεταξύ της αρχικής σελίδας (root directory) και του τελικού εγγράφου. Η διεύθυνση δεν πρέπει να περιλαμβάνει αχρείαστους φακέλους (ή λέξεις ή χαρακτήρες), ενώ και το όνομα του τελικού εγγράφου οφείλει να είναι όσο το δυνατόν συντομότερο, άρα και πιο εύκολο να αντιγραφεί κι επικολληθεί, να αναγνωσθεί μέσω τηλεφώνου, να γραφεί σε μία επαγγελματική κάρτα κ.λπ.

γ) Καθίσταται σαφές ότι οι μηχανές αναζήτησης αντιμετωπίζουν τις στατικές διευθύνσεις URL διαφορετικά από τις δυναμικές, σε αναλογία με τους φυσικούς χρήστες που δυσανασχετούν όταν το άγνωστο περιεχόμενο της σελίδας δεν προδίδεται από συμβολοακολουθίες τύπου “?sid=325&pid=12”. Όπως αναλύθηκε παραπάνω, η πιο προχωρημένη και αποτελεσματική μέθοδος μετατροπής των δυναμικών διευθύνσεων URL σε στατικές γίνεται με τα εργαλεία mod_rewrite και ISAPI_rewrite, για διακομιστές Apache και Windows, αντίστοιχα. Τέλος, κατ’ αυτήν την έννοια, η αφαίρεση των παραμέτρων από τη διεύθυνση δεν είναι αρκετή εάν αντικατασταθεί από εξίσου άσχετους και εννοιολογικά

ασύνδετους χαρακτήρες ή αριθμούς (αντί για 105/241/cat25/ είναι σαφώς προτιμότερη η χρήση του Petroupoli/Epicheirisi/Diafimistikes-Etairies/).

δ) Μία αποτελεσματική μέθοδο κατηγοριοποίησης εντός ενός ιστοχώρου αποτελεί η διαίρεση των σελίδων του σε υποτομείς (subdomains), όπως για παράδειγμα η ιστοσελίδα http://www.ece.ntua.gr. Γενικά, όμως, δεν ενδείκνυται η διαίρεση των σελίδων σε πολλαπλούς υποτομείς (π.χ. http://www.mycourses.ece.ntua.gr), ενώ, για τις ανάγκες της ορθής δόμησης των URLs, αντενδείκνυται εξ ολοκλήρου η χρήση υποτομέων όπου αυτοί δε χρειάζονται. Αυτό συμβαίνει, κυρίως, επειδή οι μηχανές αναζήτησης τείνουν να αντιμετωπίζουν τους υποτομείς ξεχωριστά από τον αρχικό τομέα, αγνοώντας για την αξία αυτού (όπως αυτή μετράται από τη θέση του τομέα για ορισμένες αναζητήσεις αλλά και σε PageRank, ένα μετρικό σύστημα που θα αναλυθεί διεξοδικά σε άλλο κεφάλαιο). Επομένως, σε περιπτώσεις ιστοτόπων περιορισμένων απαιτήσεων σε όγκο δεδομένων και πληροφοριών, έχει παρατηρηθεί ότι η κατηγοριοποίηση με υποφακέλους (χρήση, δηλαδή, μόνο του αρχικού διαδικτυακού τομέα) έχει καλύτερα αποτελέσματα.

ε) Μία πολύ σημαντική παράμετρος, όσον αφορά τη δομή των URL διευθύνσεων, είναι ο διαχωρισμός των λέξεων – κλειδιών, κατά την ονομασία των τελικών εγγράφων ή υποφακέλων ενός ιστοτόπου. Όπως φάνηκε και στο παραπάνω παράδειγμα, “Petroupoli/Epicheirisi/Diafimistikes-Etairies/”, ο διαχωρισμός των λέξεων γίνεται με το χαρακτήρα “-” και, σε περίπτωση που αυτό δεν είναι επιθυμητό ή εφικτό για τις ανάγκες μίας ιστοσελίδας, μπορούν να χρησιμοποιηθούν, με σειρά προτεραιότητας, οι χαρακτήρες “_”, “+” ή κανένας χαρακτήρας (ενωμένες λέξεις). Αυτό συμβαίνει διότι δεν αναγνωρίζουν όλες οι μηχανές αναζήτησης τους χαρακτήρες “_”, “+” ή “%20” (κενό).

στ) Τέλος, μία πολύ σημαντική λεπτομέρεια αφορά την ευαισθησία των URLs σε πεζούς και κεφαλαίους χαρακτήρες. Εφόσον ένας ιστότοπος φιλοξενείται σε διακομιστή βασισμένο σε Windows λειτουργικό σύστημα, η επίσκεψη στη σελίδα www.example.com/case-sensitive.html ή την σελίδα www.example.com/characters/case-sensitive.html μπορεί να γίνει εάν κάποιος πληκτρολογήσει ή οδηγηθεί από κάποιον σύνδεσμο www.example.com/cASe-SENSitiVe.html και www.example.com/chaRACteRs/case-sensitive.html. Αυτό, όμως, δε συμβαίνει και για ιστοτόπους που φιλοξενούνται σε διακομιστές βασισμένους σε Linux / Unix λειτουργικό σύστημα, με αποτέλεσμα ο χρήστης να οδηγηθεί σε λανθασμένη σελίδα (σφάλμα 404 ή μη ανταπόκριση στο φυλλομετρητή). Τέτοιο λάθος είναι πολύ εύκολο να συμβεί, είτε εξαιτίας του διαχειριστή μετά από κάποια επεξεργασία της σελίδας, είτε από σφάλμα του χρήστη μετά από «λανθασμένη» πληκτρολόγηση (οι χρήστες δε γνωρίζουν ότι υπάρχει τέτοια ευαισθησία στις διαδικτυακές διευθύνσεις), είτε από σφάλμα κάποιου τρίτου που θέλησε να παρέχει σύνδεσμο προς μία σελίδα και τον πληκτρολόγησε εσφαλμένα. Όπως φαίνεται και στα παραπάνω παραδείγματα, το ίδιο ισχύει τόσο για τις σελίδες όσο και για

τους υποφακέλους ενός ιστοτόπου, όχι, όμως, για τους αρχικούς τομείς (main domain), δηλαδή για τη σελίδα www.example.com του παραδείγματος. Επειδή, όμως, κανένας δεν εγγυάται για τη δια παντός φιλοξενία ενός ιστοχώρου σε συγκεκριμένο διακομιστή (και, συγκεκριμένα, λειτουργικού Windows), η ασφαλής οδός περιλαμβάνει τη δομή όλων των διευθύνσεων URL σε πεζούς χαρακτήρες (και όχι κεφαλαίους για να είναι πιο ευανάγνωστες).

Παράδειγμα δομών URL

Σύμφωνα με τα παραπάνω και δεδομένης της βαρύτητας της δομής των URL διευθύνσεων των εγγράφων ενός ιστοχώρου στην αντίληψη της σχετικότητας των εγγράφων με συγκεκριμένο όρο αναζήτησης, από την πλευρά των μηχανών αναζήτησης αλλά και των χρηστών, ο παρακάτω σύνδεσμος είναι παράδειγμα καλής δομής URL:

- http://www.dmoz.org/World/Greek/Υγεία/Ιατρική/Σύλλογοι

Η διεύθυνση αυτή δείχνει ξεκάθαρα την ιεραρχία της πληροφορίας στη σελίδα. Η πληροφορία αυτή χρησιμοποιείται από τις μηχανές αναζήτησης για να διαπιστώσει τη σχέση μίας δεδομένης ιστοσελίδας με ένα δεδομένο όρο αναζήτησης. Εξαιτίας της δομής αυτής, οι μηχανές αναζήτησης μπορούν να κατανοήσουν ότι η σελίδα αυτή δεν αφορά γενικά τους συλλόγους, όπως αναφέρεται στο τελικό έγγραφο που εξετάζεται, αλλά αναφέρεται σε ιατρικούς συλλόγους. Η πληροφορία αυτή μπορεί να εξαχθεί πριν την επίσκεψη της συγκεκριμένης σελίδας. Αντίστοιχα, παρουσιάζεται κι ένα παράδειγμα κακής δομής URL:

- http://www.amazon.co.uk/gp/product/B002Y27P46/ref=s9_pop_gw_ir04/279-6218584-0042064?pf_rd_m=A3P5ROKL5A1OLE&pf_rd_s=center-2&pf_rd_r=0K52TA0QX42X5J5WPWHK&pf_rd_t=101&pf_rd_p=202007167&pf_rd_i=468294

Αντίθετα από το πρώτο παράδειγμα, η διεύθυνση αυτή δεν αντιπροσωπεύει ούτε περιγράφει συνοπτικά την πληροφορία που περιέχει το έγγραφο, δεν είναι σύντομη, ενώ περιλαμβάνει πολλές παραμέτρους. Οι μηχανές αναζήτησης θα διαπιστώσουν ότι πρόκειται για κάποιο προϊόν (φάκελος product) που βρίσκεται στον ιστότοπο της Amazon, αλλά δε θα διαπιστώσουν περί τίνος πρόκειται έως ότου την επισκεφθούν.

4.6 Χάρτες ιστοτόπων

Οι χάρτες ιστοτόπων (sitemaps), κατά το πλαίσιο προτύπων «sitemaps.org», αποτελούν ένα εύκολο κι εύχρηστο εργαλείο για τους διαχειριστές ιστοσελίδων για την ενημέρωση των μηχανών αναζήτησης σχετικά με τις σελίδες του ιστοτόπου τους που είναι διαθέσιμες προς

ανίχνευση. 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, μέσω ενός προτύπου ανταλλαγής δεδομένων XML. 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, του γεωγραφικού περιεχομένου KML, 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.

4. Ο διαχειριστής, μετά την υποβολή ενός χάρτη, γνωρίζει, ανά πάσα στιγμή, πόσες και ποιες σελίδες έχουν ανιχνευθεί και ενσωματωθεί στο ευρετήριο της μηχανής αναζήτησης της Google, ενώ ενημερώνεται πλήρως σχετικά με τη δραστηριότητα των ανιχνευτών στις σελίδες που έχουν υποβληθεί προς ανίχνευση, όπως φαίνεται παρακάτω:

Εικόνα 18 Εποπτεία της συχνότητας ανίχνευσης σελίδων, μετά την υποβολή χάρτη

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. Κυρίως 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. Το πρόβλημα

έγκειται στο γεγονός ότι είναι πολύ πιο πιθανό η ανίχνευση από τις μηχανές αναζήτησης να είναι πιο αποτελεσματική από την ανίχνευση από τα διάφορα λογισμικά παραγωγής χαρτών XML, με αποτέλεσμα η ίδια η διαδικασία δημιουργίας και υποβολής χάρτη να μετατρέπεται σε περιττή, καθώς υποδεικνύει σελίδες που είναι ούτως ή άλλως ανιχνεύσιμες από τις μηχανές αναζήτησης.

Παρακάτω, αναλύονται οι βασικότεροι τύποι χαρτών ιστοτόπων.

4.6.1 Γενικοί χάρτες XML

Πρόκειται για τους χάρτες ιστοτόπων στην πιο απλή τους μορφή, στην οποία δηλώνουν στις μηχανές αναζήτησης ένα σύνολο από διευθύνσεις URL που απαρτίζουν έναν ιστότοπο. Όπως η κατάρτιση και υποβολή του χάρτη είναι προαιρετική, έτσι και η καταγραφή όλων των διευθύνσεων URL του ιστοτόπου δεν είναι υποχρεωτική. Συνίσταται, μάλιστα, η καταγραφή των διευθύνσεων των σημαντικότερων εγγράφων, ή εκείνων που, σύμφωνα με την εμπειρία, δεν είναι προσβάσιμες από τις μηχανές αναζήτησης, λόγω χαμηλής δραστηριότητας των ανιχνευτών ή κακής αρχιτεκτονικής του ιστοτόπου.

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.

Πίνακας 6 Ετικέτες σύνταξης γενικού χάρτη XML

Ο πρότυπος χώρος ονομάτων που πρέπει να δηλώνεται είναι ο εξής:

xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"

Έστω ο ιστότοπος www.website.com (αρχική σελίδα) ενός καταστήματος, με τρεις στατικές σελίδες (ιστορική πορεία, προϊόντα, επικοινωνία). Ως παράδειγμα σύνταξης, ακολουθεί μία πιθανή εκδοχή του XML χάρτη του ιστοτόπου:

```
<?xml version="1.0" encoding="UTF-8"?>

<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

      <url>
            <loc> http://www.website.com </loc>
            <lastmod> 2011-05-20 </lastmod>
            <changefreq> weekly </changefreq>
            <priority> 1.0 </priority>
      </url>
      <url>
            <loc> http://www.website.com/history.html </loc>
            <lastmod> 2011-05-20 </lastmod>
            <changefreq> monthly </changefreq>
            <priority> 0.7 </priority>
      </url>
      <url>
            <loc> http://www.website.com/products.html </loc>
            <lastmod> 2011-05-21 </lastmod>
            <changefreq> hourly </changefreq>
            <priority> 0.95 </priority>
      </url>
      <url>
            <loc> http://www.website.com/contact.html </loc>
            <lastmod> 2011-03-01 </lastmod>
            <changefreq> monthly </changefreq>
            <priority> 0.6 </priority>
      </url>

</urlset>
```

Στο παραπάνω παράδειγμα, υπονοείται ότι η σελίδα με τα προϊόντα ανανεώνεται πολύ συχνά μέσα στην ημέρα και έχει πολύ μεγάλη σχετική προτεραιότητα, καθώς επιθυμείται οι πληροφορίες που αφορούν στα νέα προϊόντα της εταιρείας να ευρετηριάζονται όσο το δυνατόν συχνότερα. Παράλληλα, η σελίδα της επικοινωνίας ανανεώνεται σπάνια, καθώς τα στοιχεία επικοινωνίας είναι σταθερά ή προτιμάται μία μορφή φόρμας ηλεκτρονικής επικοινωνίας, παράλληλα, όμως, της αποδίδεται πολύ μικρή προτεραιότητα, παρότι αποτελεί το μέσο επικοινωνίας επιχείρησης – καταναλωτών, παραγγελιών άρα και οικονομικής αποδοτικότητας (conversion rate) της ιστοσελίδας.

Ο λόγος για τον οποίον συμβαίνει αυτό είναι ότι η προτεραιότητα που αποδίδεται στις σελίδες αφορά στην αναγκαιότητα και τη συχνότητα ανιχνεύσεων της εκάστοτε διεύθυνσης URL και όχι τη σπουδαιότητα του περιεχομένου της σελίδας αυτό καθαυτό. Σε οποιαδήποτε περίπτωση, άλλωστε, η σελίδα της επικοινωνίας είναι η τελευταία σελίδα η θέση της οποίας στις μηχανές αναζήτησης ενδιαφέρει τον κάτοχο ή διαχειριστή ενός ιστοτόπου να βελτιστοποιήσει, καθώς αυτή αποκτά νόημα όταν ο επισκέπτης έχει πρώτα επισκεφθεί άλλες σελίδες του ιστοχώρου και έχει λόγο να επικοινωνήσει με τη διαχείριση ή διεύθυνση.

Οδηγίες σύνταξης

- Κάθε χάρτης οφείλει να περιέχει το πολύ 50,000 διευθύνσεις URL και να μην ξεπερνάει τα 10MB σε μέγεθος.
- Όλες οι διευθύνσεις URL οφείλουν να καταχωρούνται με την ίδια σύνταξη (π.χ. όλες να καταχωρούνται είτε με είτε χωρίς τη χρήση του 'www')
- Η κωδικοποίηση των χαρακτήρων οφείλει να είναι η UTF-8, ενώ πρέπει οπωσδήποτε να δηλώνεται ο πρότυπος χώρος ονομάτων που αναφέρθηκε στη σύνταξη.
- Οι διευθύνσεις URL οφείλουν να μη περιλαμβάνουν session IDs.

4.6.2 Χάρτες βίντεο

Οι χάρτες βίντεο συμβάλλουν τόσο στην ενίσχυση της διαδικασίας ευρετηρίασης των σελίδων του ιστοτόπου όσο και στη βελτιστοποίηση της θέσης των ιστοσελίδων στα αποτελέσματα των μηχανών αναζήτησης, καθώς 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 ιστοτόπων θέτει υποψηφιότητα για προβολή 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.

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+ΩραΖώνης]
<video:publication_date>	Η ημερομηνία έκδοσης του βίντεο, στη μορφή [Έτος-Μήνας-Ημέρα] ή [Έτος-Μήνας-ΗμέραΤ:Ώρες:Λεπτά:Δευτερόλεπτα+ ΩραΖώνης]
<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.

Πίνακας 7 Ετικέτες σύνταξης XML χάρτη βίντεο

Ο πρότυπος χώρος ονομάτων που πρέπει να δηλώνεται είναι ο εξής:

xmlns:video=http://www.google.com/schemas/sitemap-video/1.1

Έστω ο ίδιος ιστότοπος www.website.com του προηγούμενου παραδείγματος, ο οποίος, στη σελίδα με τα προϊόντα, περιλαμβάνει και ένα βίντεο παρουσίασης και προώθησης προσφοράς για ένα νέο προϊόν του καταστήματος. Ένας νέος, ξεχωριστός χάρτης ιστοτόπου μπορεί να δημιουργηθεί για την καταχώρηση του οπτικοακουστικού υλικού, ως εξής:

```
<?xml version="1.0" encoding="UTF-8"?>

<urlset xmlns=http://www.sitemaps.org/schemas/sitemap/0.9
        xmlns:video=http://www.google.com/schemas/sitemap-
video/1.1>

<url>
<loc> http://www.website.com/products.html </loc>
<video:video>
      <video:content_loc> http://www.website.com/video1.flv
      </video:content_loc>
      <video:thumbnail_loc> http://www.website.com/thumb1.jpg
      </video:thumbnail_loc>
      <video:title> Website's S3O brand new vacuum cleaner
      </video:title>
      <video:description> Enjoy our new vacuum cleaner
      presentation, in full HD. Watch how to use Website S3O and
      make the housework a piece of cake.</video:description>
      <video:publication_date> 2011-05-20T18:50:00+02:00
      </video:publication_date>
      <video:expiration_date> 2011-06-10T23:59:59+02:00
      </video:expiration_date>
      <video:duration> 120 </video:duration>
      <video:tag> vacuum </video:tag>
      <video:tag> vacuum cleaning </video:tag>
      <video:tag> housecleaning </video:tag>
      <video:tag> Website </video:tag>
      <video:tag> S3O cleaner </video:tag>
<video:video>
</url>

</urlset>
```

Στο σημείο αυτό, τονίζεται ότι μπορούμε να ενσωματώσουμε τις νέες πληροφορίες για την τοποθεσία <loc> http://www.website.com/products.html </loc> στο προηγούμενο χάρτη ιστοτόπου XML. Όμως, προτείνεται η δήλωση των βίντεο να γίνεται σε ξεχωριστό χάρτη, για λόγους εύκολης εποπτείας και διαχείρισης.

Οδηγίες σύνταξης

- Κάθε χάρτης οφείλει να περιέχει το πολύ 50,000 βίντεο και να μην ξεπερνάει τα 10MB σε μέγεθος.
- Όλες οι διευθύνσεις URL οφείλουν να καταχωρούνται με την ίδια σύνταξη (π.χ. όλες να καταχωρούνται είτε με είτε χωρίς τη χρήση του 'www')
- Η κωδικοποίηση των χαρακτήρων οφείλει να είναι η UTF-8, ενώ πρέπει οπωσδήποτε να δηλώνονται οι πρότυποι χώροι ονομάτων που αναφέρθηκαν στη σύνταξη.
- Οι τύποι αρχείων που υποστηρίζονται από τις μηχανές αναζήτησης είναι συγκεκριμένοι (mpg, mpeg, mp4, m4v, mov, wmv, asf, avi, ra, ram, rm, flv, swf), ενώ τα αρχεία οφείλουν να είναι προσβάσιμα μέσω πρωτοκόλλου HTTP και όχι να απαιτούν μεταφόρτωση με άλλα πρωτόκολλα.
- Το περιεχόμενο των ετικετών <video:title> και <video:description> οφείλει να είναι το ίδιο με τον τίτλο της σελίδας και τη meta ετικέτα περιγραφής της σελίδας που βρίσκεται στην τοποθεσία <loc>.

4.6.3 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).

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.

Πίνακας 8 Ετικέτες σύνταξης XML χάρτη εικόνων

Ο πρότυπος χώρος ονομάτων που πρέπει να δηλώνεται είναι ο εξής:

xmlns:image=http://www.google.com/schemas/sitemap-image/1.1

Έστω ο ίδιος ιστότοπος www.website.com, ο οποίος, στη σελίδα με τα προϊόντα, περιλαμβάνει μία εικόνα για το προϊόν σε προσφορά, καθώς και μία φωτογραφία του καταστήματος, στη σελίδα της ιστορικής πορείας της επιχείρησης. Ένας νέος, ξεχωριστός χάρτης ιστοτόπου μπορεί να δημιουργηθεί για την καταχώρηση των εικόνων δύο περιζήτητων προϊόντων, ως εξής:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:image="http://www.google.com/schemas/sitemap-image/1.1">

<url>
      <loc>http://www.website.com/products.html</loc>
      <image:image>
            <image:loc> http://www.website.com/images/S3O.jpg
            </image:loc>
            <image:caption> The new revolutionary vacuum cleaner from
            Website, S3O</image:caption>
            <image:title> Website S3O cleaner </image:title>
      </image:image>
</url>
<url>
      <loc>http://www.website.com/history.html</loc>
      <image:image>

      <image:loc>http://www.website.com/images/store.jpg</image:loc>
            <image:caption> Our store is located in the heart of
            Athens, currently expanded to support a 300-slot parking
            </image:caption>
            <image:title>Website - Electrical appliances</image:title>
            <image:geo_location> Athens, Greece </image:geo_location>
      </image:image>
</url>
</urlset>
```

Όπως και στην περίπτωση του χάρτη βίντεο, μπορούμε να ενσωματώσουμε τις νέες πληροφορίες για τις παραπάνω διευθύνσεις URL στον πρώτο χάρτη ιστοτόπου που δημιουργήθηκε, όμως συνίσταται η δημιουργία ξεχωριστών χαρτών ιστοτόπου για λόγους εποπτείας.

Οδηγίες σύνταξης

- Κάθε χάρτης οφείλει να περιέχει το πολύ 50,000 διευθύνσεις URL, το πολύ 50,000 εικόνες και να μην ξεπερνάει τα 10MB σε μέγεθος.
- Όλες οι διευθύνσεις URL οφείλουν να καταχωρούνται με την ίδια σύνταξη (π.χ. όλες να καταχωρούνται είτε με είτε χωρίς τη χρήση του ‘www’)
- Η κωδικοποίηση των χαρακτήρων οφείλει να είναι η UTF-8, ενώ πρέπει οπωσδήποτε να δηλώνονται οι πρότυποι χώροι ονομάτων που αναφέρθηκαν στη σύνταξη και το παράδειγμα.
- Οι διευθύνσεις URL οφείλουν να μη περιλαμβάνουν session IDs.

4.6.4 Χάρτες ιστοτόπων συμβατών με κινητά τηλέφωνα

Προς το παρόν, τα αποτελέσματα αναζήτησης στον Παγκόσμιο Ιστό ενδέχεται να 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/>
```

Ο πρότυπος χώρος ονομάτων που πρέπει να δηλωθεί, παράλληλα με τον βασικό, είναι ο εξής:

xmlns:mobile="http://www.google.com/schemas/sitemap-mobile/1.0"

Έστω, για παράδειγμα, ο ίδιος ιστότοπος www.website.com, που περιλαμβάνει τις τέσσερις προαναφερθείσες σελίδες (αρχική, πορεία, προϊόντα, επικοινωνία) οι οποίες είναι συμβατές και με φυλλομετρητές κινητών τηλεφώνων. Ο επιπλέον χάρτης ιστοτόπου που μπορούμε να δημιουργήσουμε είναι ο εξής:

```
<?xml version="1.0" encoding="UTF-8" ?>
 <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:mobile="http://www.google.com/schemas/sitemap-mobile/1.0">

      <url>
            <loc> http://www.website.com </loc>
            </mobile:mobile/>
      </url>
      <url>
            <loc> http://www.website.com/history.html </loc>
            </mobile:mobile/>
      </url>
      <url>
            <loc> http://www.website.com/products.html </loc>
            </mobile:mobile/>
      </url>
      <url>
            <loc> http://www.website.com/contact.html </loc>
            </mobile:mobile/>
      </url>

</urlset>
```

Τονίζεται, όμως, από τις μηχανές αναζήτησης ότι ο χάρτης ιστοτόπου προορισμένου (και) για προβολή από κινητά τηλέφωνα οφείλει να διαχωρίζεται από τον βασικό χάρτη, ενώ πρέπει να περιλαμβάνει μόνο διευθύνσεις ιστοσελίδων οι οποίες είναι συμβατές με κινητά τηλέφωνα. Ενδέχεται οι σελίδες αυτές να είναι συμβατές τόσο με κινητά τηλέφωνα όσο και με φυλλομετρητές υπολογιστών, όμως σε καμία περίπτωση δεν πρέπει να συμπεριλαμβάνονται διευθύνσεις URL που δεν έχουν συμβατότητα με κινητά τηλέφωνα.

4.6.5 Πολλαπλοί χάρτες

Δίνεται η δυνατότητα στους διαχειριστές μίας σελίδας που έχουν δημιουργήσει πολλαπλούς χάρτες για τον ιστότοπό τους 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. Δηλαδή, πρόκειται για έναν απλό χάρτη, στον οποίο τη θέση των διευθύνσεων URL των ιστοσελίδων του ιστοτόπου έχουν πάρει οι διευθύνσεις URL των υπόλοιπων χαρτών, οι οποίοι, με τη σειρά τους, περιλαμβάνουν τις URL διευθύνσεις των σελίδων ή, ομοίως, URL διευθύνσεις άλλων χαρτών. Ο αρχικός χάρτης στον οποίο καταχωρούνται οι διευθύνσεις άλλων χαρτών ιστοτόπων ονομάζεται και ευρετήριο χαρτών.

Με τον τρόπο αυτό, ενθαρρύνονται οι διαχειριστές να 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, βελτιώνοντας σημαντικά το ποσοστό των σελίδων και στοιχείων ενός ιστοχώρου που ευρετηριάζονται και διατίθενται προς προβολή στα αποτελέσματα αναζητήσεων.

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.

Πίνακας 9 Ετικέτες σύνταξης πολλαπλών χαρτών XML

Έστω ο ίδιος ιστότοπος του παραδείγματος, καθώς και όλοι οι διαφορετικοί χάρτες που έχουν δημιουργηθεί για την καταχώρηση. Ένα παράδειγμα ευρετηρίου των διαφόρων χαρτών αποτελεί το παρακάτω τμήμα κώδικα:

```
<?xml version="1.0" encoding="UTF-8"?>

<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

      <sitemap>
            <loc> http://www.website.com/XMLsitemap.xml </loc>
            <lastmod> 2011-05-25T18:00:00+00:00 </lastmod>
      </sitemap>
      <sitemap>
            <loc> http://www.website.com/maps/VideoSitemap.xml </loc>
            <lastmod> 2011-04-21T18:00:00+00:00 </lastmod>
      </sitemap>
      <sitemap>
            <loc> http://www.website.com/maps/ImageSitemap.xml </loc>
            <lastmod> 2011-05-25T18:00:00+00:00 </lastmod>
      </sitemap>
      <sitemap>
            <loc>
            http://www.website.com/usr/alex/seo/MobileSitemap.xml
            </loc>
            <lastmod> 2011-05-25T18:00:00+00:00 </lastmod>
      </sitemap>

</sitemapindex>
```

4.6.6 Δήλωση των χαρτών

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, όπως αυτή έχει αναλυθεί στο αντίστοιχο κεφάλαιο. Ενδεικτικά, για το παραπάνω παράδειγμα της εταιρείας "Website", μπορεί να συμπεριληφθεί το παρακάτω τμήμα εντός του αρχείου robots.txt.

```
#Δήλωση όλων των χαρτών:
Sitemap: http://www.website.com/XMLsitemap.xml
Sitemap: http://www.website.com/maps/VideoSitemap.xml
Sitemap: http://www.website.com/maps/ImageSitemap.xml
Sitemap: http://www.website.com/usr/alex/seo/MobileSitemap.xml

#Εναλλακτικά, μπορούμε να δηλώσουμε μόνο το ευρετήριο χαρτών:
Sitemap: http://www.website.com/maps/SitemapIndex.xml
```

Στη συνέχεια, ενδείκνυται η ειδικευμένη υποβολή του χάρτη ή πολλαπλών χαρτών ιστοτόπου σε καθεμία εκ των μεγάλων μηχανών αναζήτησης Google, Yahoo και Bing, με τη χρήση του αντίστοιχου εργαλείου Google Webmasters Central, Yahoo SiteExplorer και Bing Webmaster, αντίστοιχα.

Για την υποβολή του χάρτη στη μηχανή Ask.com, μπορεί να χρησιμοποιηθεί απευθείας η διεύθυνση URL http://submissions.ask.com/ping?sitemap=, ακολουθούμενη από τη διεύθυνση URL του χάρτη ιστοτόπου.

4.6.7 Καλύτερες πρακτικές

α) Μία πολύ καλή τακτική αντιμετώπισης ενδεχόμενης κακής σχεδίασης της δομής και της αρχιτεκτονικής ενός ιστοτόπου είναι η καθυστέρηση δημιουργίας και υποβολής του XML χάρτη. Με τον τρόπο αυτό, μπορούμε να εντοπίσουμε ποιες σελίδες εντοπίζονται, υπό κανονικές συνθήκες, από τους ανιχνευτές και ευρετηριάζονται και ποιες όχι. Στη συνέχεια, δημιουργούμε τον χάρτη του ιστοτόπου και τον υποβάλλουμε, φροντίζοντας να συμπεριλάβουμε οπωσδήποτε τις σημαντικότερες σελίδες εξ αυτών που δεν ανιχνεύθηκαν, καθώς επίσης και να διορθώσουμε τυχόν λάθη διασύνδεσης μεταξύ των σελίδων και προς τις ελαττωματικές αυτές σελίδες.

β) Όπως αναφέρθηκε ήδη, ο χάρτης ιστοτόπου μπορεί να εντοπισθεί από τις μηχανές αναζήτησης επειδή είναι εκτεθειμένος σε οποιονδήποτε επισκέπτη, είτε πρόκειται για φυσικό πρόσωπο είτε για ανιχνευτή. Αυτό σημαίνει ότι οι σημαντικές πληροφορίες που συνοδεύουν το XML αρχείο, όπως η σχετική προτεραιότητα κάθε διεύθυνσης URL, γίνονται άμεσα γνωστές σε οποιονδήποτε ανταγωνιστή ανακαλύψει την τοποθεσία του χάρτη. Στο σημείο

αυτό, τονίζεται ότι το sitemap μπορεί να φορτωθεί σε οποιονδήποτε φάκελο ή υποφάκελο εντός του διακομιστή, στο ίδιο όνομα τομέα με τον ιστότοπο, ή και εκτός του διακομιστή (όπως πραγματοποιείται με εταιρείες διαμοιρασμού χαρτών βίντεο ή εικόνων ιστοτόπων), αρκεί να υποβληθεί στις κατάλληλες πλατφόρμες (αντίστοιχα για κάθε διαφορετική μηχανή αναζήτησης) και, κατά προτίμηση, να δηλώνεται επίσης στο αρχείο robots.txt, όπως αναλύθηκε παραπάνω.

γ) Η χρήση ενός ενιαίου χάρτη ιστοτόπου, στον οποίο θα δηλώνονται οι πληροφορίες για κάθε διεύθυνση URL, οι πληροφορίες για κάθε βίντεο ή εικόνα καθώς και η συμβατότητα με τα κινητά τηλέφωνα, είναι δυνατή αλλά δε συνίσταται. Ο βασικός λόγος γι' αυτό είναι η μειωμένη δυνατότητα διαχείρισης, διόρθωσης και ενημέρωσης των χαρτών ιστοτόπων, καθώς στην περίπτωση του ενιαίου sitemap το αρχείο XML είναι αχανές. Δίνεται, μάλιστα, η δυνατότητα υποβολής ενός χάρτη δήλωσης όλων των διαφορετικών χαρτών, αντί της υποβολής όλων των χαρτών ξεχωριστά. Οι μηχανές αναζήτησης συνιστούν τη χρήση διαφορετικών χαρτών ιστοτόπων ακόμη και ίδιου τύπου, όταν πρόκειται για αρκετά μεγάλους ιστοχώρους με πολλαπλούς υποτομείς ή υποκατηγορίες (για παράδειγμα, τρεις διαφορετικοί XML χάρτες για το blog, τα προϊόντα και τις υπόλοιπες συνολικά κατηγορίες ενός μεγάλου ιστοχώρου).

δ) Τέλος, πολύ σημαντική τακτική αποτελεί η παράλειψη της δήλωσης ορισμένων λιγότερο σημαντικών διευθύνσεων URL από το χάρτη XML, με σκοπό τη μείωση του αριθμού των URLs που συμπεριλαμβάνονται στο χάρτη και την επιλεκτική ενίσχυση της ευρετηριάσης και απόδοσης στα αποτελέσματα ορισμένων ασθενέστερων, από άποψη αρχιτεκτονικής, αλλά εξίσου ή περισσότερο σημαντικών σελίδων.

4.7 Στρατηγική domain

4.7.1 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 (global top level domain names – gTLDs) 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 «ECE NTUA». Ο οργανισμός μπορεί να κατοχυρώσει μόνο το domain name «ecentua.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:

- Ece-ntua.gr
- Ece-ntua.org
- Ece-ntua.eu
- Ece-ntua.info
- Ecentua.gr
- Ecentua.org
- Ecentua.eu
- Ecentua.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 τον βαθμό PageRank που αποτελεί έναν από τους πιο κρίσιμους παράγοντες που αφορούν τη βελτιστοποίηση των ιστοσελίδων και που θα αναλυθεί στο επόμενο κεφάλαιο διεξοδικά. 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.

Επίσης, όταν το επιτρέπουν οι συνθήκες και, κυρίως, το αντικείμενο του οργανισμού ή της επιχείρησης, καλό είναι να συμπεριλαμβάνονται φράσεις ή λέξεις – κλειδιά που αφορούν το ευρύτερο αντικείμενο εντός του ονόματος domain, όπως ακριβώς συνίσταται και για τα ονόματα των εγγράφων ενός ιστοχώρου, καθώς, όπως έχει ήδη αναφερθεί, η ύπαρξη των

λέξεων ενδιαφέροντος στις διευθύνσεις URL αποτελεί καθοριστική παράμετρο στον αλγόριθμο κατάταξης των αποτελεσμάτων αναζήτησης για τις λέξεις αυτές.

Έπειτα, έχει παρατηρηθεί ότι ιστοσελίδες που αφορούν ένα απλό, περιορισμένο θέμα τείνουν να εμφανίζονται υψηλότερα στα αποτελέσματα αναζήτησης, σε σχέση με αυτές που αφορούν ποικίλα θέματα, σχετικά ή άσχετα μεταξύ τους. Έτσι, μία καλή πρακτική αποτελεί ο διαχωρισμός του περιεχομένου ενός ιστοχώρου σε περισσότερους από έναν τομείς ή υποτομείς (domains ή sub-domains, αντίστοιχα), με κριτήριο το θέμα που, με αυτό τον τρόπο, δίνει τη δυνατότητα περαιτέρω εξειδίκευσης. Έτσι, για παράδειγμα, ένα βρετανικό ειδησεογραφικό portal, με όνομα «into», θα μπορούσε να περιλαμβάνει τα εξής:

- Intopolitics.co.uk
- Intoeconomy.co.uk
- IntoBritain.co.uk

Ή, αντίστοιχα:

- Politics.into.co.uk
- Economy.into.co.uk
- Britain.into.co.uk

Φυσικά, παρότι η καταχώρηση διαφορετικών domains ή sub-domains για έναν ιστοχώρο μπορεί να έχει καλύτερα αποτελέσματα, όσον αφορά το SEO, είναι δεδομένο ότι επιβαρύνει περισσότερο σε όρους πόρων, κόστους, εμπειρίας του χρήστη και προώθησης brand name. Επίσης, στην περίπτωση αυτή, επανέρχεται το πρόβλημα του διαιρούμενου βαθμού PageRank, μόνο που τώρα δεν είναι εφικτή η ανακατεύθυνση καθώς πρόκειται για ιστοσελίδες αυτόνομες, από πλευράς τόσο ονόματος domain όσο και περιεχομένου.

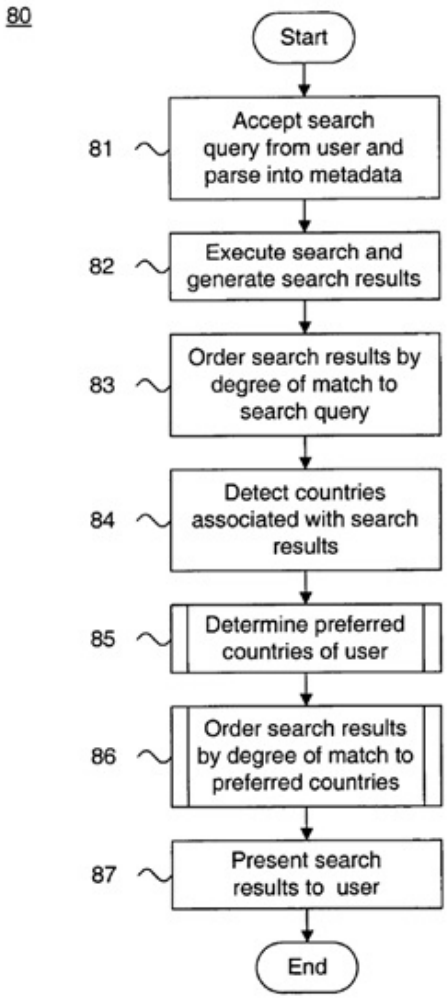
4.7.2 Γεωγραφική τοποθέτηση

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.

Ο παραπάνω ελληνικός εκπαιδευτικός οργανισμός, για παράδειγμα, δεν έχει ιδιαίτερο ενδιαφέρον να εμφανίζεται στη Χιλή, ή το Εκουαδόρ. Αντίστοιχα, η αποδοτικότητα της μηχανής αναζήτησης δεν περιορίζεται μόνο στην παροχή σχετικών αποτελεσμάτων στους χρήστες αλλά και την ταξινόμηση αυτών. Έτσι, εάν ο ιστότοπος του συγκεκριμένου

οργανισμού υστερεί έναντι ενός αντίστοιχου βρετανικού, σε όρους βελτιστοποίησης (συνολικά, σε ορισμένους ή όλους τους παράγοντες), τότε 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):

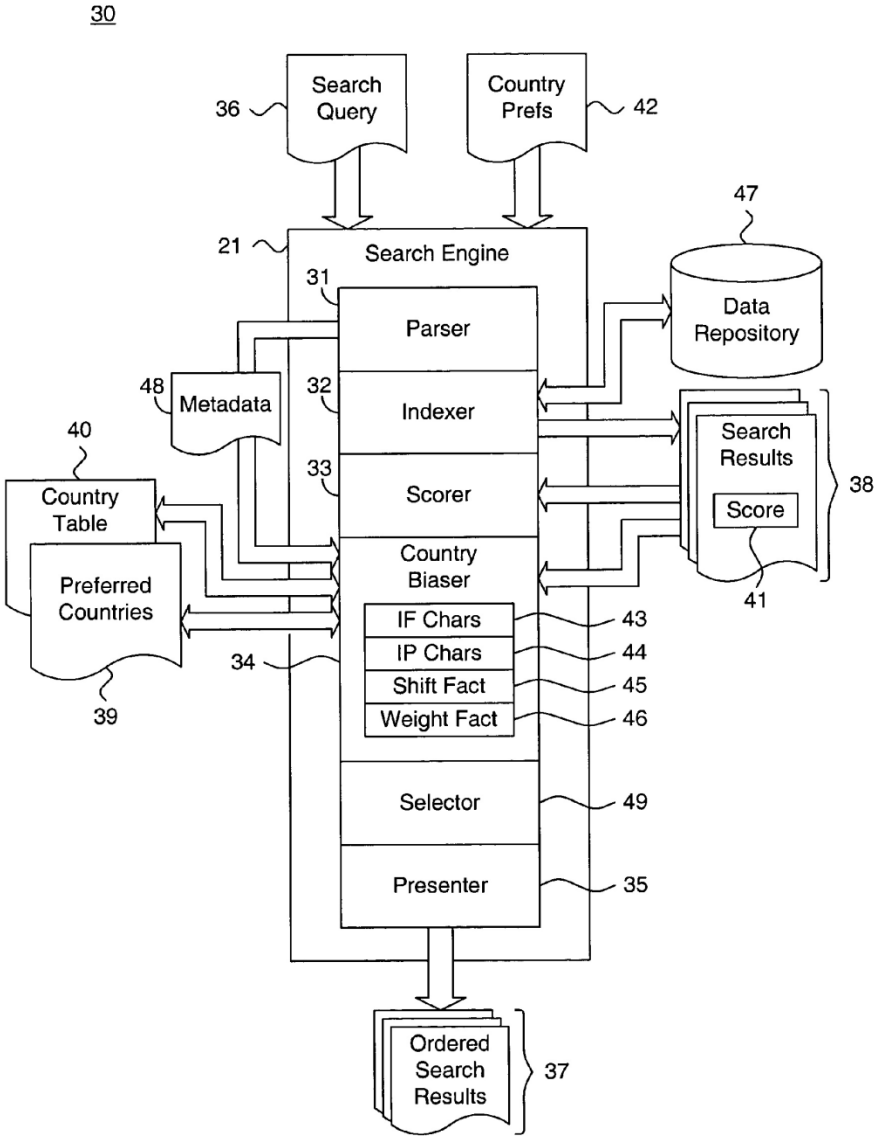


Εικόνα 19 Διάγραμμα ροής γεωγραφικού φιλτραρίσματος αποτελεσμάτων της Google

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, «ecentua.gr» για Ελλάδα, ή «politics.into.co.uk» για Ηνωμένο Βασίλειο). Οι διάφοροι κωδικοί των χωρών μπορούν να βρεθούν.
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)

Συγκεκριμένα, σύμφωνα με τη Google, κατά την περιγραφή της διαδικασίας εντοπισμού της χώρας προέλευσης που προηγείται του φιλτραρίσματος βάσει γεωγραφικών κριτηρίων των αποτελεσμάτων αναζήτησης, «ο indexer (32)», κατά τη διαδικασία, δηλαδή, της ευρετηρίασης, όπως αυτή έχει αναλυθεί σε προηγούμενο κεφάλαιο, «εντοπίζει τις χώρες που συνδέονται με κάθε αποτέλεσμα αναζήτησης (38). Η χώρα μπορεί να προσδιοριστεί με ποικίλους τρόπους. Πρώτα, η χώρα μπορεί να προσδιοριστεί από την επέκταση της διεύθυνσης URL, κάθε αποτελέσματος αναζήτησης (38). Για παράδειγμα, η διεύθυνση URL www.whsmith.co.uk φέρεται είτε να βρίσκεται είτε να συνδέεται με το Ηνωμένο Βασίλειο. Δεύτερον, η διεύθυνση της υπηρεσίας κατοχύρωσης ονομάτων domain μπορεί να εξετασθεί και η χώρα προέλευσης της επιχείρησης (ή του οργανισμού) να εξαχθεί ως συμπέρασμα. Ομοίως, η χώρα μπορεί να εξαχθεί ως συμπέρασμα από τη διεύθυνση IP του διακομιστή από τον οποίον αποκτήθηκε το αποτέλεσμα αναζήτησης (38), από το έγγραφο του αποτελέσματος αναζήτησης, ή από άλλες ιστοσελίδες στον ίδιο ιστότοπο. Τέλος, το anchor text των υπερσυνδέσμων στο αποτέλεσμα αναζήτησης, το κείμενο πλησίον του υπερσυνδέσμου, ή οι χώρες των ιστοσελίδων με υπερσυνδέσμους προς το αποτέλεσμα αναζήτησης μπορεί να εξετασθεί. Κατά προτίμηση, ένας συνδυασμός των παραπάνω μεθόδων μπορεί να πραγματοποιηθεί για την παροχή αποτελεσμάτων μεγαλύτερης ακρίβειας. Επιπροσθέτως, άλλες τεχνικές για τον εντοπισμό των χωρών που σχετίζονται με τα αποτελέσματα αναζήτησης είναι πιθανές.»



Εικόνα 20 Γεωγραφικός εντοπισμός των ιστοσελίδων από τη Google

Όσον αφορά τις άλλες μηχανές αναζήτησης, αυτές φέρονται να αξιοποιούν μόνο το όνομα domain και την επέκτασή του καθώς και την ανάλυση των υπερσυνδέσμων (παράγοντες 1 και 4 της αξιολόγησης της Google), με εξαίρεση τη μηχανή της GigaBlast, όπως αναλύεται παρακάτω.

Τέλος, μία πρακτική τακτική που ακολουθείται, όσον αφορά διεθνείς επιχειρήσεις ή οργανισμούς, είναι η δημιουργία τοπικών εκδοχών των ιστοσελίδων, καθώς έτσι πολλαπλασιάζεται ο αριθμός των σελίδων που επιστρέφονται στα αποτελέσματα αναζήτησης, χωρίς να διατρέχεται ο κίνδυνος διπλότυπου περιεχομένου, ενώ, εάν λαμβάνονται υπόψη οι παραπάνω 4 παράμετροι της γεωγραφικής τοποθέτησης, οι τοπικές ιστοσελίδες ευνοούνται

περισσότερο, σε σύγκριση με τις διεθνείς εκδοχές των ίδιων ιστοτόπων, ειδικά όταν οι χρήστες προτιμούν τις αντίστοιχες τοπικές εκδοχές των μηχανών αναζήτησης (π.χ. την «google.gr» έναντι της «google.com»). Στο σημείο αυτό τονίζεται ότι έχει υπολογισθεί πως η πλειοψηφία (περίπου 70–90%) των χρηστών προτιμούν την τοπική εκδοχή της μηχανής αναζήτησης που χρησιμοποιούν ή, εφόσον αυτή η επιλογή δεν προσφέρεται, φιλτράρουν τα αποτελέσματα βάσει της χώρας τους.

Γεωγραφικές Meta ετικέτες

Η μηχανή αναζήτησης GigaBlast διαφέρει από τις υπόλοιπες μηχανές, στο επίπεδο της geolocation, καθώς αναγνωρίζει και λαμβάνει υπόψη τις meta ετικέτες περιεχομένου γεωγραφικής τοποθεσίας, οι οποίες είναι οι εξής:

```
<head>

      <meta name="zipcode" content="15780" />

      <meta name="city" content="Αθήνα, Ζωγράφου, Athens,
      Zografoy" />

      <meta name="country" content="Greece, Hellas,
      Ελλάδα" />

<!-- Στην περίπτωση των ΗΠΑ, δηλώνεται και η πολιτεία -->
      <meta name="state" content="Massachusetts" />


</head>
```

4.7.3 Κανονικοποίηση

Όπως έχει ήδη αναλυθεί, οι μηχανές αναζήτησης αποδίδουν ποινές, σε όρους θέσεων στα αποτελέσματα αναζήτησης, στους ιστοτόπους που περιλαμβάνουν πολλαπλές σελίδες με ακριβώς το ίδιο περιεχόμενο. Στη βιομηχανία του Διαδικτύου και του Search Engine Optimization, η ποινή αυτή ονομάζεται ποινή διπλότυπου περιεχομένου (duplicate content penalty). Ένα από τα προβλήματα που συναντώνται συχνά και σχετίζονται με την ανάγκη ύπαρξης διπλότυπου περιεχομένου είναι, όπως αναφέρθηκε, η αναγκαιότητα κατοχύρωσης πολλών διαφορετικών domains για έναν ιστότοπο.

Με τη μέθοδο της κανονικοποίησης, επιτυγχάνεται η συνύπαρξη των διαφορετικών domain names, έναντι της επικράτησης ενός επιθυμητού τομέα στον οποίο οι υπόλοιποι θα ανακατευθύνουν, χωρίς να διατρέχεται ο κίνδυνος της ποινής διπλότυπου περιεχομένου.

Έστω, για παράδειγμα, η ιστοσελίδα του οργανισμού «ecentua» και μία σελίδα για τις τεχνολογίες ανανεώσιμων πηγών ενέργειας, η οποία ενδέχεται να είναι προσβάσιμη από τους χρήστες (αλλά και τους ανιχνευτές των μηχανών αναζήτησης), μέσω ποικίλλων διευθύνσεων URL. Έτσι, έστω η διεύθυνση URL με την επιθυμητή δομή URL:

- http://www.ecentua.org/item=renewable-energy

Η σελίδα αυτής της URL μπορεί να προβληθεί από τον επισκέπτη, σίγουρα μέσω πλοήγησης εντός του ιστοτόπου κι ενδεχομένως, μέχρι να εφαρμοσθεί η τεχνική κανονικοποίησης, μέσω των αποτελεσμάτων αναζήτησης, σε άλλες διευθύνσεις URL, λόγω αναγνωριστικού συνεδρίας (session ID), προεπιλεγμένης (ή μη) γλώσσας, παράλληλης πρόσβασης από δύο κατηγορίες, και άλλων παραγόντων, ως εξής:

- http://www.ecentua.org/item=renewable-energy&sessionid=39820
- http://www.ecentua.org/item=renewable-energy&language=EN
- http://www.ecentua.org/item=renewable-energy&category=technologies

Αντίστοιχα, το ίδιο πρόβλημα, σε επίπεδο ονόματος και τύπου τομέα, υπάρχει και μεταξύ των σελίδων:

- http://www.ecentua.org/item=renewable-energy
- http://www.ecentua.gr/item=renewable-energy

Έτσι, προτάθηκε τον Φεβρουάριο του 2009, αρχικά από την Google και στη συνέχεια από τις Ask, Microsoft Live και Yahoo, η χρήση μίας ετικέτας κανονικοποίησης, στην οποία ανατίθεται ως τιμή η προτιμώμενη διεύθυνση URL που ο διαχειριστής επιθυμεί να ευρετηριάζεται ως μοναδική, για το αντίστοιχο περιεχόμενο, σελίδα στα αποτελέσματα αναζήτησης.

Σύνταξη

Η ετικέτα της κανονικοποίησης συνδέσμων εισάγεται εντός του <HEAD> τομέα του αρχείου HTML, ως εξής:

```
<head>

      <link rel="canonical" href="επιθυμητή προτιμώμενη
      διεύθυνση URL" />

</head>
```

Η ετικέτα rel=”canonical” εισάγεται στις σελίδες εκείνες με το διπλότυπο περιεχόμενο, ενώ «δείχνει» προς την επιθυμητή διεύθυνση URL. Έτσι, στο παραπάνω παράδειγμα, για τη μεγαλύτερη δυνατή βελτιστοποίηση της ιστοσελίδας, οφείλουμε να προσθέσουμε στον <HEAD> τομέα των σελίδων

- http://www.ecentua.org/item=renewable-energy&sessionid=39820,
- http://www.ecentua.org/item=renewable-energy&language=EN και
- http://www.ecentua.org/item=renewable-energy&category=technologies

την παρακάτω γραμμή κώδικα:

```
<head>

      <link rel="canonical"
      href="http://www.ecentua.org/item=renewable-energy" />

</head>
```

Σύμφωνα με τους μηχανικούς των μηχανών αυτών αναζήτησης, δύο η περισσότερες διευθύνσεις URL δεν απαιτείται να είναι ταυτόσημες, σε ποσοστό 100%, προτού κανονικοποιηθούν, αλλά επιτρέπονται ελάχιστες διαφορές μεταξύ τους, όπως, για παράδειγμα, η σειρά ταξινόμησης των προϊόντων ενός πίνακα, τα ενδεχόμενα προτεινόμενα προϊόντα, κ.λπ., ενώ δεν απαγορεύεται, ούτε τιμωρείται, η κανονικοποίηση διευθύνσεων προς σελίδα σφάλματος 404, δηλαδή σελίδα που δεν υπάρχει. Σε αυτή την περίπτωση, οι μηχανές θα προχωρήσουν στην ευρετηρίαση του περιεχομένου και θα χρησιμοποιήσουν κάποια ευριστική για να βρουν μία κανονική διεύθυνση. Μάλιστα, σε περίπτωση που δεν έχει ακόμη ευρετηριαστεί η κανονική διεύθυνση αλλά μόνο κάποια σελίδα διπλότυπου περιεχομένου, οι ανιχνευτές θα αγνοήσουν την ετικέτα κανονικοποίησης μέχρις ότου εντοπίσουν κι ευρετηριάσουν την κανονική διεύθυνση.

Επίσης, ενδέχεται η κανονική διεύθυνση να συνδέει σε επόμενη διεύθυνση URL, μέσω ανακατεύθυνσης (όπως αναλύεται στην επόμενη παράγραφο), δηλαδή μία σελίδα να κανονικοποιείται σε μία διαφορετική διεύθυνση και, στη συνέχεια, αυτή να επανακατευθύνει σε μία άλλη σελίδα, με τη μέθοδο της ανακατεύθυνσης, ή να οδηγεί σε μία σειρά κανονικοποιήσεων. Οι ανιχνευτές αναζήτησης έχουν τη δυνατότητα να ακολουθούν αλυσίδες κανονικοποιήσεων (και ανακατευθύνσεων) και να ευρετηριάζουν ως βασική την διεύθυνση

στην οποία θα καταλήξουν, παρότι προτείνεται η εξ αρχής κανονικοποίηση όλων των συνδέσμων στην τελευταία.

Τέλος, από τις αρχές του 2010, οι μηχανές αναζήτησης επιτρέπουν την κανονικοποίηση διευθύνσεων προς διεύθυνση ενός τελείως διαφορετικού ονόματος τομέα (cross-domain canonicalization). Η τακτική αυτή είναι ιδιαιτέρως χρήσιμη όταν κάποιος χρήστης διατηρούσε μία προσωπική σελίδα, ή υποτομέα, σε έναν ξένο διακομιστή και domain που τον φιλοξενούσε (π.χ. websites.com/~user1245) και επιθυμεί να κανονικοποιήσει το υπάρχον περιεχόμενο στη νέα του προσωπική ιστοσελίδα (π.χ. user1245.com), ή όταν μία επιχείρηση (ή οργανισμός) διατηρεί και ένα blog στον ιστοχώρο της στον οποίο αναδημοσιεύει δελτία τύπου νέων υπηρεσιών τα οποία έχει ήδη παρουσιάσει επακριβώς σε κάποια άλλη σελίδα (π.χ. αυτή με τα προϊόντα και τις υπηρεσίες της).

4.7.4 Ανακατεύθυνση

Ως ανακατεύθυνση ορίζεται η διαδικασία προώθησης μίας URL σε μία διαφορετική.

Οι περιπτώσεις στις οποίες η διαδικασία αυτή εμφανίζεται χρήσιμη, έως καθοριστική για την πορεία μίας ιστοσελίδας στα αποτελέσματα αναζητήσεων, 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 το βαθμό PageRank του ιστοτόπου και να διατρέχουν τον κίνδυνο οι μηχανές αναζήτησης να αποδώσουν κάποια ποινή για διπλότυπο περιεχόμενο. Για παράδειγμα, οι διευθύνσεις http://www.ecentua.org/index.php, http://www.ecentua.org και http://ecentua.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%, σύμφωνα με την SEOmoz, της αξίας των συνδέσμων (βαθμός PageRank) 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 Search Engine Optimization, 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 δε φροντίζει να μεταφερθεί η αξία των συνδέσμων και ο βαθμός PageRank 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 έγκειται στην έκδοση του Πρωτοκόλλου HTTP (Hyper Text Transfer Protocol) με την οποία είναι συμβατός ο διακομιστής φιλοξενίας. Το πρωτόκολλο αυτό Διαδικτύου αφορά τον τρόπο λειτουργίας των συνδέσμων. Έτσι, 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. Παρ' όλα αυτά, η χρήση της 307 ανακατεύθυνσης επιβάλλει την εκ των προτέρων επίγνωση αφενός της συμβατότητας του διακομιστή με την HTTP 1.1, και αφετέρου της επίγνωσης αυτής της συμβατότητας από τις μηχανές αναζήτησης. Επειδή το τελευταίο δεν είναι πιθανό να γνωρίζεται από το διαχειριστή, ή οποιονδήποτε χρήστη, συνίσταται η αποκλειστική χρήση της 302 ανακατεύθυνσης, που στην HTTP 1.1 μεταφράζεται πλέον ως «Found» από τις μηχανές αναζήτησης, γι' αυτό και στη βιβλιογραφία αναφέρεται μόνο ως 302 ανακατεύθυνση.

- Η Meta ανανέωση (Meta refresh). Πρόκειται για μία ανακατεύθυνση που δε 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. Δεδομένου, μάλιστα, ότι η 302 προτιμάται εκούσια και στοχευμένα δεν αποδίδει μέρος του βαθμού PageRank, θεωρείται και η λιγότερο αποδοτική μέθοδος ανακατεύθυνσης.

Επομένως, η μέθοδος της Meta ανανέωσης δε συνηθίζεται να χρησιμοποιείται, ως τεχνική βελτιστοποίησης της ιστοσελίδας, καθώς πραγματοποιεί την ίδια ακριβώς λειτουργία με την 301 ανακατεύθυνση, χωρίς, όμως, να διαβιβάζει την αξία συνδέσμων της διεύθυνσης URL στη νέα σελίδα – προορισμό. Η χρήση της συνίσταται μόνο σε περίπτωση που ο διαχειριστής εκούσια επιθυμεί τη μη μεταβίβαση οποιασδήποτε SEO 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, το βαθμό PageRank και τη γενικότερη αξία που η σελίδα καταλάμβανε πριν την αλλαγή. Εάν η ανακατεύθυνση οδηγεί τους ανιχνευτές αναζήτησης, ιδιαίτερα, σε διαφορετικό domain (δηλαδή, η αλλαγή δεν υφίσταται εντός του ίδιου τομέα), τότε η διαδικασία επιβραδύνει ακόμη περισσότερο, εξαιτίας της κατάχρησης που παρατηρείται και της αναμενόμενης καχυποψίας των μηχανών αναζήτησης απέναντι σε τέτοιες πρακτικές.

Ανακατεύθυνση 301 σε Apache διακομιστή

Για την αξιοποίηση αυτής της τεχνικής, ο εξυπηρετητής οφείλει να περιέχει εγκατεστημένη την επέκταση (module) «mod_rewrite», η οποία συνήθως είναι εργοστασιακά εγκατεστημένη.

Έστω ο ιστότοπος www.website.com.

Καταρχάς, πρέπει να πραγματοποιηθεί η δημιουργία (εάν δεν υπάρχει ήδη) της σελίδας phpinfo. Αρκεί να δημιουργηθεί ένα αρχείο κειμένου (.txt) στο οποίο να αντιγραφεί η παρακάτω γραμμή κώδικα:

```
<?php phpinfo(); ?>
```

Στη συνέχεια, αποθηκεύεται το αρχείο ως “phpinfo.php” και τοποθετείται στο διακομιστή, στην αρχική κατηγορία (root category). Ανοίγοντας την αντίστοιχη σελίδα στο φυλλομετρητή www.website.com/phpinfo.php, παρατηρούμε όλες τις σχετικές με την PHP πληροφορίες του διακομιστή και, αναζητώντας τη μονάδα “mod_rewrite”, βρίσκουμε ότι αυτή έχει φορτωθεί στο διακομιστή:

apache2handler

Apache Version	Apache/2
Apache API Version	20051115
Server Administrator	webmaster@website.com
Hostname:Port	
User/Group	apache(1000)/105
Max Requests	Per Child: 0 - Keep Alive: on - Max Per Connection: 100
Timeouts	Connection: 60 - Keep-Alive: 1
Virtual Server	Yes
Server Root	/etc/httpd
Loaded Modules	core mod_authn_file mod_authn_default mod_authz_host mod_authz_groupfile mod_authz_user mod_authz_default mod_auth_basic mod_include mod_filter mod_deflate mod_log_config mod_logio mod_env mod_headers mod_unique_id mod_setenvif mod_proxy mod_proxy_connect mod_proxy_ftp mod_proxy_http mod_proxy_ajp mod_proxy_balancer mod_ssl prefork http_core mod_mime mod_dav mod_status mod_autoindex mod_asis mod_suexec mod_cgi mod_dav_fs mod_dav_lock mod_negotiation mod_dir mod_actions mod_userdir mod_alias mod_rewrite mod_so mod_php5

Εικόνα 21 Προβολή phpinfo.php για την εποπτεία των φορτωμένων modules

Σε αντίθετη περίπτωση ευθύνεται η εταιρεία φιλοξενίας που κατέχει το διακομιστή.

Δημιουργούμε ένα αρχείο .htaccess, το οποίο περιέχει εντολές Apache που ισχύουν στο φάκελο που περιλαμβάνει το .htaccess, καθώς και τους υποφακέλους που περιέχονται σε αυτόν. Για τη δημιουργία του, αρκεί πάλι η δημιουργία ενός .txt αρχείου και η αποθήκευσή του ως .htaccess (κενό όνομα αρχείου και επέκταση htaccess). Ακολουθεί η σύνταξη που απαιτείται για την επιθυμητή λειτουργία mod_rewrite, όπως αυτή υποδεικνύεται από την Apache.org (Apache, 2004).

Για την ενεργοποίηση της λειτουργίας, αρκεί η προσθήκη της παρακάτω γραμμής εντολών στο αρχείο .htaccess:

```
RewriteEngine On
```

Για την ανακατεύθυνση διευθύνσεων URL, μέσω του αρχείου αυτού, απαιτείται καλή γνώση ορισμένων βασικών κανονικών εκφράσεων (Regular Expressions ή Regexes), δηλαδή συμβολοσειρών που περιγράφουν ή αντιστοιχούν σε σύνολα συμβολοσειρών, σύμφωνα με

αυστηρώς καθορισμένους συντακτικούς κανόνες. Συγκεκριμένα, για τις ανάγκες της 301, μας απασχολούν τα εξής σύμβολα:

- . (τελεία) – αντιστοιχεί σε οτιδήποτε.
- (αστερίσκος) – υπάρχει καμία ή περισσότερες φορές ο προηγούμενος χαρακτήρας
- + (συν) – υπάρχει τουλάχιστον μία φορά ο προηγούμενος χαρακτήρας
- () (παρένθεση) – διαχωρισμός πράξεων, προς τα πίσω αναφορά
- | (κάθετη γραμμή) – διάζευξη (το λογικό Ή – OR)
- \ (ανάστροφη κάθετος) – αφαίρεση συντακτικής έννοιας του ακόλουθου χαρακτήρα
- ^ (περισπωμένη) – άρνηση (το λογικό ΟΧΙ – NOT)

Έτσι, ορίζονται οι ανακατευθύνσεις τύπου 301 για τα πλέον συνήθη σενάρια που σχεδόν αποκλειστικά απασχολούν τους διαχειριστές ιστοσελίδων.

α. Ανακατεύθυνση ορισμένων αρχείων και φακέλων από έναν τομέα σε έναν άλλο

Έστω ότι ο χρήστης user1245, σε προηγούμενο παράδειγμα, επιθυμεί να μεταφερθεί από τον ιστοχώρο φιλοξενίας websites.com, στον δικό του τομέα, ανακατευθύνοντας τα αρχεία και τους φακέλους του από τη διεύθυνση http://www.websites.com/~user1245/, στη νέα διεύθυνση URL http://www.user1245.com. Αρκεί να προστεθεί στο .htaccess η ακόλουθη εντολή:

```
RewriteEngine On
RedirectMatch 301 /~user1245/(.*) http://www.user1245.org/$1
```

Η κανονική έκφραση /~user1245/(.*) δηλώνει το φάκελο /~user1245/ ακολουθούμενο από οποιονδήποτε (ή κανένα) χαρακτήρα, δηλαδή κάθε αρχείο, φάκελο και υποφάκελο εντός του /~user1245/. Η χρήση των παρενθέσεων αποθηκεύει τη συμβολοσειρά ως προς τα πίσω αναφορά σε μία μεταβλητή, η οποία τοποθετείται στο τέλος της URL διεύθυνσης στην οποία ανατέθηκε και φέρει όνομα, εδώ, $1.

β. Ανακατεύθυνση κανονικοποιημένων ονομάτων τομέα

Ο ίδιος χρήστης επιθυμεί την ανακατεύθυνση όλων των σελίδων που δεν περιλαμβάνουν το πρόθεμα «www.» σε αυτές που το κάνουν, για λόγους αισθητικής, για την αποφυγή συνήθων λαθών κανονικοποίησης αλλά και στα πλαίσια της βελτιστοποίησης του ιστοτόπου του, καθώς, παρότι ο υποτομέας «www» δεν αποτελεί επιλογή του χρήστη αλλά ταυτίζεται με τον

τομέα ως προεπιλογή, η αξία του ιστοτόπου σε εισερχόμενους συνδέσμους διαιρείται όταν οι δύο διευθύνσεις είναι εξίσου προσβάσιμες και διαφορετικές. Έτσι το πρόβλημα ανάγεται στην ανακατεύθυνση των URL διευθύνσεων της μορφής http://user1245.com/ στις διευθύνσεις http://www.user1245.com/. Για την επίλυση αυτού του προβλήματος προστίθενται οι εξής εντολές:

```
RewriteEngine On
RewriteCond %{HTTP_HOST} ^user1245\.com [NC]
RewriteRule (.*) http://www.user1245.com/$1 [L,R=301]
```

Η εντολή ενημερώνει το διακομιστή να εξετάσει το όνομα τομέα, δηλαδή το user1245.com, και εάν αυτό δεν (^) είναι ίδιο με το www.user1245.com τότε να το ανακατευθύνει στο www.user1245.com. Το σύμβολο [NC] δηλώνει να μη ληφθεί υπόψη εάν οι χαρακτήρες είναι γραμμένοι πεζά ή κεφαλαία (No Case), ενώ το [L] σημαίνει «Τελευταίος Κανόνας» (Last Rule), δηλαδή μόλις διαβαστεί αυτός ο κανόνας όλοι οι επόμενοι να αγνοηθούν, και το [R=301] υποδηλώνει τον τύπο της ανακατεύθυνσης (301).

γ. Ανακατεύθυνση χωρίς τη διατήρηση του ονόματος αρχείου

Στο ίδιο παράδειγμα, υπάρχουν αρχεία που ο χρήστης δεν επιθυμεί να μεταφέρει στο νέο του ιστότοπο. Δηλαδή, το αρχείο host-privacy.php, π.χ., που αφορά στο ιδιωτικό απόρρητο του ιστοχώρου φιλοξενίας websites.com, δε χρειάζεται να αντιγραφεί στο νέο διακομιστή και να υπάρχει στη νέα ιστοσελίδα. Με την παραπάνω ανακατεύθυνση 301, όμως, κάτι τέτοιο θα οδηγούσε στην ανύπαρκτη σελίδα http://www.user1245.com/host-privacy.php που θα επέστρεφε σελίδα σφάλματος 404. Για να αποφευχθεί αυτό και οποιαδήποτε αίτηση για τη σελίδα αυτή να ανακατευθύνει στην αρχική σελίδα του χρήστη, χρησιμοποιείται η παρακάτω εντολή:

```
RewriteEngine On
RedirectMatch 301 /~user1245/host-privacy.php http://www.user1245.com
```

δ. Ανακατεύθυνση με παράλληλη μετατροπή της επέκτασης αρχείου

Έστω, τώρα, ότι ο χρήστης έχει προσθέσει νέες λειτουργίες στον ιστότοπό του δυναμικά, αξιοποιώντας τη γλώσσα PHP, ενώ αρχικά οι σελίδες του ήταν σχεδόν όλες σε HTML. Για

την ανακατεύθυνση όλων των αρχείων του σε νέες διευθύνσεις σελίδων PHP, ανεξαρτήτως προέλευσης (PHP ή HTML), χρησιμοποιείται η εξής εντολή:

```
RewriteEngine On
RedirectMatch 301 /~user1245/(.*)\.(php|html)
http://www.user1245.com/$1.php
```

4.8 Βελτιστοποίηση Flash περιεχομένου

Η τεχνολογία Flash αποτελεί ένα αμφιλεγόμενο κομμάτι, από άποψη βελτιστοποίησης, καθώς 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.

Στις 30 Ιουνίου 2008, η Google γίνεται η πρώτη μηχανή αναζήτησης που αναπτύσσει αλγόριθμο, σε συνεργασία με την Adobe, για την ευρετηρίαση κειμένου εντός των αρχείων Flash (Google Blog, 2008). Έκτοτε, ακολουθούν προσπάθειες από όλες τις μεγάλες μηχανές αναζήτησης και φημολογείται ότι η Yahoo δύναται να ευρετηριάσει κείμενο από Flash ιστοσελίδες ομοίως με την πρωτοπόρο Google, όμως μόνο η τελευταία έχει δημοσιεύσει επίσημη ανακοίνωση. 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.

Είναι προφανές ότι με την τήρηση των προτύπων που θέτει η μηχανή αναζήτησης, ο κάτοχος της Flash ιστοσελίδας μπορεί μόνο να ελπίζει ότι η μηχανή θα καταφέρει να ευρετηριάσει μέρος του περιεχομένου, χωρίς, όμως, να επιτυγχάνει την ευρετηρίαση ολόκληρου του Flash περιεχομένου και χωρίς να αποκτά οποιαδήποτε δυνατότητα βελτιστοποίησης της σελίδας αυτής.

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».**

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, καθιστά βέβαιη την ευρετηρίαση του Flash περιεχομένου από τις μηχανές αναζήτησης, καθώς προτυποποιεί την εμφάνιση του στοιχείου για τις μεγαλύτερες μηχανές αναζήτησης, ανεξαρτήτως κωδικοποίησης.

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. Προφανώς, για τη βελτιστοποίηση του εναλλακτικού κώδικα HTML του περιεχομένου, χρησιμοποιούνται όλες οι τεχνικές που αναπτύχθηκαν λεπτομερώς στο παρόν κεφάλαιο. 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, από άποψη Search Engine Optimization, κατά το οποίο ο διαχειριστής δε 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, δε 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.

Σύνταξη

Ο αλγόριθμος που χρησιμοποιεί η δέσμη ενεργειών «swfobject» επισυνάπτεται μαζί με τη διπλωματική εργασία. Για την αξιοποίηση του «swfobject», αρκεί να συμπεριληφθεί το αρχείο swfobject.js στο διακομιστή φιλοξενίας του ιστοτόπου καθώς και ο ελάχιστος δυνατός κώδικας στην ιστοσελίδα με το flash στοιχείο, όπως φαίνεται παρακάτω:

```
<script type="text/javascript" src="swfobject.js"></script>

<div id="flashcontent">
      Σε αυτό το div γράφεται ο κώδικας του εναλλακτικού
      περιεχομένου, κάνοντας χρήση κάθε εντολής, ετικέτας και
      στοιχείου της HTML.
</div>

<script type="text/javascript">
   var so = new SWFObject("file.swf", "id", "width", "height",
"version", "#colour");
   so.write("flashcontent");
</script>
```

α) Εξετάζοντας τον κώδικα αποσπασματικά, το περιεχόμενο του Flash παρουσιάζεται εναλλακτικά με HTML εντός του div:

```
<div id="flashcontent">
      …
</div>
```

Οι χρήστες που έχουν την πιο πρόσφατη έκδοση Flash, ή αυτή που απαιτείται από το Flash στοιχείο, δε θα χρειαστεί να δουν το περιεχόμενο αυτού του στοιχείου και είναι το ζητούμενο, από άποψη βελτιστοποίησης, κομμάτι της τεχνικής «swfobject», καθώς ό,τι υπάρχει εντός του στοιχείου θα είναι ορατό στις μηχανές αναζήτησης και θα υπόκειται σε θεμιτές τεχνικές βελτιστοποίησης.

β) Η ακόλουθη γραμμή κώδικα καλεί το αρχείο Flash να αναπαραχθεί, εφόσον έχει πραγματοποιηθεί ο έλεγχος από τη δέσμη ενεργειών κι έχει επιβεβαιωθεί δυνατότητα αναπαραγωγής από το φυλλομετρητή του χρήστη:

```
var so = new SWFObject("file.swf", "id", "width", "height",
"version", "#bgcolour");
```

Τα στοιχεία που αναγράφονται εντός των παρενθέσεων αφορούν το αρχείο Flash. Συγκεκριμένα, με τη σειρά που αναγράφονται, αφορούν στο μονοπάτι του αρχείου Flash (π.χ. "myflash.swf"), το αναγνωριστικό του αντικειμένου, το μήκος και πλάτος του αρχείου, όπως θα προβάλλεται στο φυλλομετρητή, σε pixels, η έκδοση Flash που απαιτείται από το αρχείο για αναπαραγωγή (π.χ. "8" ή "8.0.2.15"), καθώς και το χρώμα του παρασκηνίου του Flash στοιχείου στη δεκαεξαδική του τιμή (π.χ. "#4D6479").

γ) Τέλος, η ακόλουθη γραμμή κώδικα ενημερώνει τη δέσμη ενεργειών SWFObject να γράψει το περιεχόμενο Flash στη σελίδα (εφόσον έχει ολοκληρωθεί ο έλεγχος), αντικαθιστώντας το περιεχόμενο εντός του HTML στοιχείου (div):

```
so.write("flashcontent");
```

Η δέσμη ενεργειών δίνει, επίσης, τη δυνατότητα στο χειριστή της να ενσωματώσει επιπρόσθετες παραμέτρους για το Flash περιεχόμενο.

Υπάρχουν δύο επιλογές για την κωδικοποίηση του αλγορίθμου SWFObject, η στατική και η δυναμική.

- Η στατική επιλογή ενσωματώνει, με ετικέτες σήμανσης, τόσο περιεχόμενο SWF (Flash) όσο κι εναλλακτικό περιεχόμενο, ενώ χρησιμοποιεί την JavaScript σε πολύ περιορισμένο βαθμό (διακριτικά, όπως επισημαίνεται στην ιστοσελίδα της Adobe) για να επιλύσει ορισμένα θέματα που δεν επιλύονται με τη σήμανση. Το κυριότερο πλεονέκτημα της στατικής κωδικοποίησης του «swfobject» είναι ότι ο μηχανισμός της ενσωμάτωσης SWF εξαρτάται αποκλειστικά από το στοιχείο «object», έτσι το περιεχόμενο μπορεί να φθάσει σε σημαντικά μεγαλύτερο αριθμό επισκεπτών απ' ό,τι

εάν παρουσιαζόταν σε γλώσσα σεναρίων. Από την άλλη, φαίνεται πως εμφανίζει προβλήματα με ορισμένους φυλλομετρητές (Internet Explorer, Opera), όσον αφορά μηχανισμούς που ενεργοποιούνται με κλικ, παρότι η Microsoft έχει κάνει σχετικές προόδους.

- Η δυναμική επιλογή χρησιμοποιεί σήμανση για τον προσδιορισμό του εναλλακτικού περιεχομένου μόνο, ενώ χρησιμοποιεί την JavaScript για την αντικατάσταση του εναλλακτικού περιεχομένου με το Flash περιεχόμενο, εφόσον αυτό κριθεί δυνατό από τον έλεγχο συμβατότητας που πραγματοποιεί η δέσμη «swfobject». Η μέθοδος αυτή δεν εξαρτάται από μηχανισμούς που ενεργοποιούνται με κλικ και είναι πιο περιεκτική από τη στατική. Το βασικό της μειονέκτημα, όμως, είναι ότι η ενσωμάτωση SWF αρχείων εξαρτάται και από την JavaScript, με αποτέλεσμα μικρότερη κάλυψη σε σχέση με την στατική κωδικοποίηση.

Από άποψη Search Engine Optimization, η στατική επιλογή θεωρείται καλύτερη, καθώς δεν εξαρτάται από JavaScript και παρέχει ένα πολύ πιο άμεσο μονοπάτι για το αντικείμενο Flash, σε σχέση με τη δυναμική, χωρίς αυτό να σημαίνει ότι η χρήση της δυναμικής κωδικοποίησης θα επιφέρει αρνητικά αποτελέσματα.

Το παράδειγμα που δόθηκε παραπάνω, ως σύνταξη, αποτελεί δείγμα δυναμικής κωδικοποίησης.

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.
- Οι μηχανές αναζήτησης απαιτούν, με αρκετά τακτικές ανακοινώσεις και άρθρα, την πλήρη συμβατότητα και αντιστοιχία του περιεχομένου Flash με το εναλλακτικό περιεχόμενο, για να αποφεύγονται τεχνικές Black – hat βελτιστοποίησης. Επομένως, δεν επιλύει το πρόβλημα στις σελίδες Flash που περιέχουν όλο το SWF υλικό σε μία, αρχική σελίδα, εκτός εάν πρόκειται για SWF αρχείο με ένα μοναδικό περιεχόμενο, χωρίς στοιχεία μενού.

Άλλες πρακτικές

α) Σε περίπτωση που όλο το αρχείο Flash βρίσκεται σε μία σελίδα, η τεχνική «swfobject» μπορεί να αξιοποιηθεί, χωρίζοντας το αρχείο σε τμήματα, αντί να διαχωριστεί σε μικρότερα αρχεία. Με την παράμετρο FlashVars της <object> ετικέτας της HTML, μπορούν να ανατεθούν μεταβλητές στο αρχείο SWF, ώστε να εμφανίζεται το ίδιο αρχείο σε κάθε διαφορετική σελίδα του ιστοτόπου, αλλά με διαφορετική εκκίνηση κάθε φορά, προβάλλοντας το περιεχόμενο που αντιστοιχεί στο σωστό σημείο κάθε φορά. Μάλιστα, μπορεί να προστεθεί η ετικέτα (tag) του κάθε σημείου, μετά τη διεύθυνση URL της αρχικής σελίδας, μετά από μία "#". Έτσι, θα μπορεί ο χρήστης και να πλοηγηθεί σωστά μπροστά ή πίσω, κάνοντας χρήση των λειτουργιών του φυλλομετρητή. Μετά το χωρισμό του αρχείου SWF σε μικρότερα τμήματα (που αντιστοιχούν σε διαφορετικό κουμπί του μενού του, για παράδειγμα), μπορεί να εφαρμοστεί η τεχνική «swfobject» όπως έχει αναλυθεί.

β) Για Flash αρχεία που περιέχουν πολύ μεγάλα κείμενα, μπορεί να χρησιμοποιηθεί η μέθοδος SifR (Scalable Inman Flash Replacement), με την οποία το Flash στοιχείο που περιέχει κείμενο αντικαθίσταται από ένα πλαίσιο με HTML κείμενο, πλήρως προσβάσιμο και ανιχνεύσιμο από τις μηχανές αναζήτησης. Έτσι, εάν δε χρησιμοποιηθεί η μέθοδος «swfobject», το κείμενο θα μπορεί να ευρετηριασθεί από τις μηχανές αναζήτησης και, παράλληλα, θα υπόκειται σε τεχνικές βελτιστοποίησης (κυρίως με ετικέτες μορφοποίησης), ενώ, σε αντίθετη περίπτωση, η σελίδα θα περιλαμβάνει το κείμενο τόσο ως εναλλακτικό κείμενο όσο και ως κυρίως κείμενο ενσωματωμένο στο SWF αρχείο.

γ) Ένα βασικό μειονέκτημα της τεχνολογίας Flash είναι και η έλλειψη εσωτερικών συνδέσμων εκτός του στοιχείου Flash. Έτσι, συνίσταται η δημιουργία ενός μενού (ψηλά ή χαμηλά) στη σελίδα, τόσο για την εύκολη πλοήγηση των χρηστών, ανεξάρτητα από το εάν έχουν πρόσβαση στο Flash ή το εναλλακτικό περιεχόμενο, όσο και για την ενίσχυση των σελίδων για τις μηχανές αναζήτησης.

4.9 Θέματα χρόνου και συχνότητας

4.9.1 Το φαινόμενο «sandbox»

Ένα ιδιαίτερα αμφιλεγόμενο ζήτημα, τόσο στη βιομηχανία όσο και την ακαδημαϊκή βιβλιογραφία του Search Engine Optimization, αποτελεί το φαινόμενο «sandbox» που αφορά αποκλειστικά στη μηχανή αναζήτησης της Google και είναι γνωστό ως το «Google Sandbox Effect». Πρόκειται για ένα φαινόμενο που ενδέχεται να 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, 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 τη, 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. Αξίζει να σημειωθεί ότι άλλες μετρήσεις από τα εργαλεία της Google (όπως, για παράδειγμα, ο βαθμός PageRank) έδειχναν ιδιαίτερα θετικά.

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.

Όσον αφορά τη διάρκεια του φαινομένου «sandbox», αυτή ποικίλλει από ιστοσελίδα σε ιστοσελίδα, αλλά και από περίοδο σε μία άλλη. Ενδιαφέροντα σχετικά συμπεράσματα που έχουν εξαχθεί, κατά καιρούς, από τη συμπεριφορά της μηχανής της Google απέναντι στις νέες ιστοσελίδες είναι τα εξής:

- Η απόκτηση, όπως αναλύεται στο επόμενο κεφάλαιο, περισσότερων και ποιοτικών εισερχόμενων συνδέσμων συμβάλλει στην επιτάχυνση της διαδικασίας και τη συντομότερη έξοδο από το «sandbox».
- Αντίθετα, η απόκτηση συνδέσμων «χαμηλής ποιότητας» προς τον ιστότοπο ενδέχεται να επιφέρει τα αντίθετα αποτελέσματα. Τέτοιοι σύνδεσμοι θεωρούνται εκείνοι που δημιουργούνται χειροκίνητα από το διαχειριστή, κυρίως σε καταλόγους καταχώρησης ιστοσελίδων.
- Αιτήματα αναθεώρησης ιστοσελίδων, όπως αυτά προσφέρονται από τη Google, ενδέχεται να απελευθερώσουν τον ιστοχώρο από το φαινόμενο, παρότι είναι αρκετά δύσκολο να γνωρίζουμε εάν τα αιτήματα αυτά, στις σχετικές έρευνες, είχαν καθοριστικό ρόλο ή, απλώς, είχε παρέλθει ο χρόνος ισχύος του φαινομένου.
- Τέλος, έχει παρατηρηθεί ότι οι ιστότοποι δεν εξέρχονται μόνοι από το «sandbox» αλλά κατά ομάδες. Φαίνεται πως η Google ενεργοποιεί ορισμένες διαδικασίες, εσωτερικά στη μηχανή, κατά τις οποίες ένας συγκεκριμένο σύνολο ιστοσελίδων που έχουν υποπέσει στο φαινόμενο ανακαταλαμβάνουν τις αναμενόμενες θέσεις στα αποτελέσματα αναζήτησης την ίδια ακριβώς στιγμή.

Από τα δύο τελευταία, γίνεται εμφανές ότι το φαινόμενο «sandbox» αποτελεί μεν παράμετρο που καθορίζει τη θέση μίας καινούριας ιστοσελίδας στα αποτελέσματα αναζητήσεων της συγκεκριμένης μηχανής, αλλά δεν αποτελεί μεταβλητή στους αλγορίθμους κατάταξης. Αντίθετα, ενεργοποιείται και απενεργοποιείται ξεχωριστά από τη δυναμική διαδικασία της κατάταξης.

Καλύτερες πρακτικές

α) Από τα παραπάνω, γίνεται σαφές πως ο διαχειριστής ενός νέου ιστοτόπου που επιθυμεί να επιταχύνει την απελευθέρωση αυτού από το φαινόμενο «sandbox» οφείλει να επιδιώξει, δημιουργήσει ή ενθαρρύνει ποιοτικούς συνδέσμους από ιστοσελίδες που σχετίζονται με το περιεχόμενό του και να αποφύγει, κατ' αρχάς, την καταχώρηση σε πολλούς διαδικτυακούς

καταλόγους ιστοσελίδων, καθώς και τις «φάρμες» συνδέσμων, που θα αναλυθούν στο επόμενο κεφάλαιο.

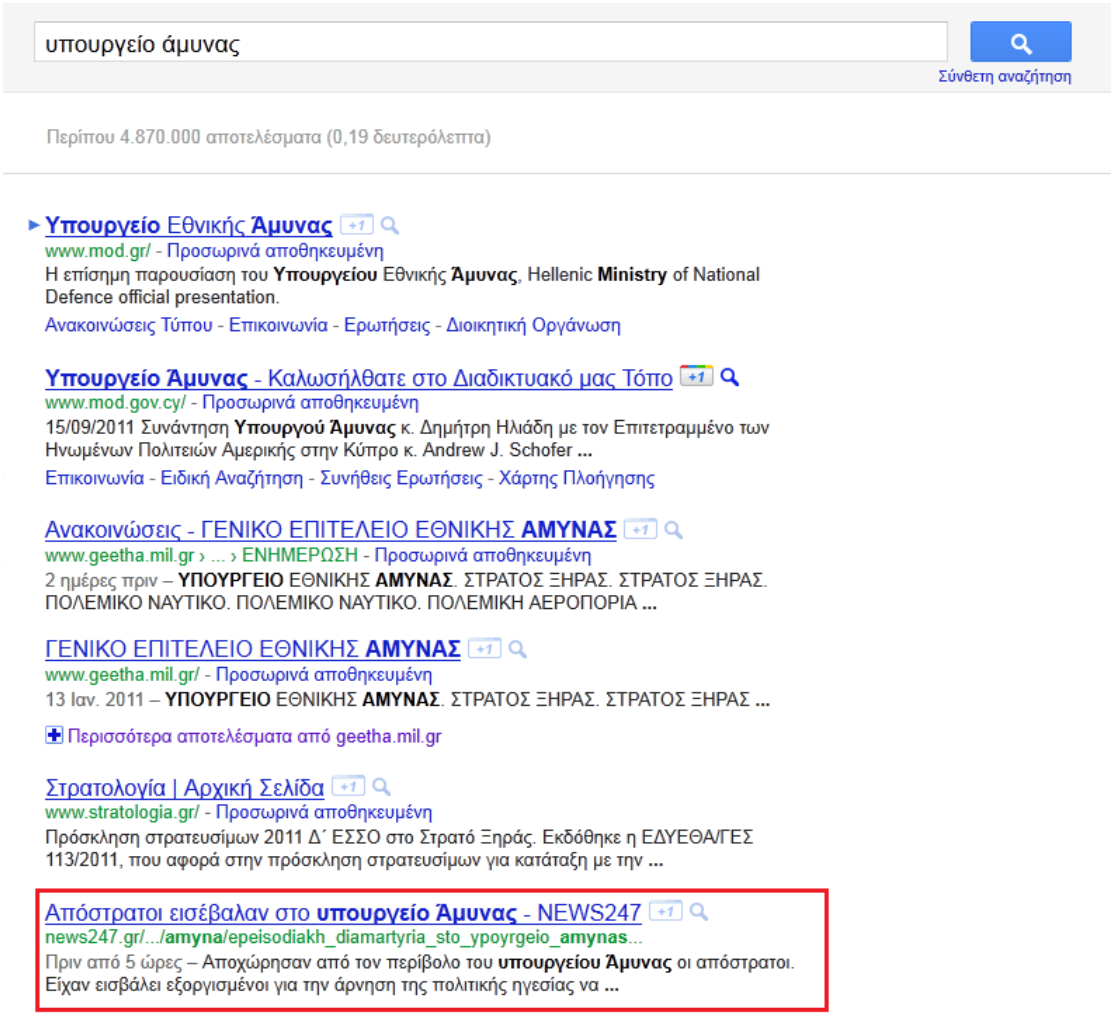
β) Μία αποτελεσματική πρακτική που ακολουθείται είναι η δημιουργία νέων ιστοσελίδων σε ήδη υπάρχοντα ονόματα τομέα, μέσω της μεταφοράς τους σε ανενεργούς παλαιότερους ιστοτόπους. Παράλληλα, η δημιουργία νέων υπηρεσιών (blog, portal, σελίδα καμπάνιας ή προσφορών) είναι καλύτερο να πραγματοποιείται εντός του ίδιου τομέα, ή σε υποτομείς ενός ήδη υπάρχοντος τομέα, όπως, για παράδειγμα, η δημιουργία ενός blog για τον ιστότοπο into.gr θα μπορούσε να γίνει στον υποτομέα blog.into.gr και όχι σε διαφορετικό domain.

γ) Τέλος, συνίσταται η δημιουργία Google AdSense λογαριασμού και η σύνδεσή του με το νέο ιστότοπο, καθώς και η παροχή ενός χώρου σε αυτόν αξιοποίησης για την προβολή διαφημίσεων του δικτύου της Google. Με τον τρόπο αυτό, έχει παρατηρηθεί ότι επιταχύνεται ιδιαίτερα η διαδικασία επίδρασης του φαινομένου. Η χρήση του τεχνάσματος αυτού, προφανώς, δε χρειάζεται να συνεχιστεί αφότου επανέλθουν οι θέσεις του ιστοτόπου στα αποτελέσματα αναζήτησης.

4.9.2 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, η λειτουργία του οποίου αφορά στη διαρκή αναζήτηση νέου περιεχομένου. 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 22 Παράδειγμα ευνοϊκής κατάταξης φρέσκων σελίδων

Επομένως, συνίσταται 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.

4.9.3 Μακροβιότητα ιστοτόπου

Στον αντίποδα του παραπάνω παράγοντα βελτιστοποίησης, εκτιμάται ότι 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.

4.9.4 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. Είναι προφανές ότι, καθώς 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 μελετάται διαρκώς και επιδρά, παράλληλα με τον αριθμό και την ποιότητα αυτών των συνδέσμων καθαυτό (όπως θα αναλυθεί στο επόμενο κεφάλαιο), στον αλγόριθμο κατάταξης των αποτελεσμάτων αναζήτησης κάθε μηχανής.

Μεγάλη προσοχή πρέπει να δοθεί σε αυτό τον παράγοντα, καθώς οι αλλαγές αυτές που μελετώνται από τις μηχανές αναζήτησης δεν πρέπει να εμφανίζονται δραματικές, εν συγκρίσει με τη νόρμα. 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).

5 *Τεχνικές βελτιστοποίησης εκτός της ιστοσελίδας*

Η διαδικασία της βελτιστοποίησης της κατάταξης ενός ιστοτόπου στα αποτελέσματα των μηχανών αναζήτησης δεν πραγματοποιείται μόνο σε επίπεδο του κώδικα και του διακομιστή φιλοξενίας αυτού. Αντιθέτως, επεκτείνεται και εκτός αυτού, καθώς οι μηχανές αναζήτησης υπολογίζουν την αξία ενός εγγράφου στο Διαδίκτυο βάσει των συνδέσμων όλων των υπόλοιπων εγγράφων προς αυτό.

Έτσι, αναπτύσσονται και τεχνικές βελτιστοποίησης που εφαρμόζονται όχι εντός της ιστοσελίδας, όπως αναλύθηκε στο προηγούμενο κεφάλαιο, αλλά και εκτός αυτής, οι οποίες περιορίζονται αποκλειστικά στη δημιουργία συνδέσμων. Η διαδικασία αυτή ονομάζεται κατασκευή ή δημιουργία συνδέσμων (ή Link Building, στη διεθνή βιβλιογραφία) και αφορά στη δημιουργία ή φυσική απόκτηση εισερχόμενων συνδέσμων από εξωτερικές ιστοσελίδες.

Στο παρόν κεφάλαιο αναλύεται πώς και γιατί οι σύνδεσμοι μεταξύ σελίδων είναι τουλάχιστον εξίσου σημαντικοί με τη βελτιστοποίηση εντός της ιστοσελίδας, όσον αφορά τις μηχανές αναζήτησης, καθώς και τρόποι με τους οποίους βελτιστοποιείται ο παράγοντας αυτός.

5.1 Κατασκευή συνδέσμων και η σημασία της στο SEO

Ένας από τους βασικούς λόγους για τη δραματική ανάπτυξη του Διαδικτύου είναι η ευκολία με την οποία μπορούν να ανταλλάσσονται σύνδεσμοι μεταξύ σχετικών αλλά και άσχετων μεταξύ τους ιστοτόπων, καθώς και μεταξύ των σελίδων του ίδιου ιστοτόπου, δηλαδή με την οποία μία σελίδα μπορεί να συνδεθεί με μία άλλη. Το φαινόμενο αυτό υπήρξε κι εξακολουθεί να αποτελεί, ως επί τω πλείστω, μία φυσική διαδικασία, στα πλαίσια της οποίας οι σελίδες με

αυθεντικό, καλής ποιότητας περιεχόμενο λαμβάνουν συνδέσμους από άλλες ιστοσελίδες, οι διαχειριστές των οποίων αντιλαμβάνονται το περιεχόμενο στο οποίο συνδέουν άξια πηγή πληροφοριών για τους επισκέπτες τους.

Η παραπάνω διαδικασία, με την οποία οι σελίδες συνδέονται με τρόπο φυσικό και με σκοπό την παραπομπή των επισκεπτών σε ενδιαφέρον ή σχετικό περιεχόμενο, είχε ως αποτέλεσμα οι μηχανές αναζήτησης, με πρωτοπόρο τη Google και τους εμπνευστές της, να συνειδητοποιήσουν ότι η δημοτικότητα των ιστοσελίδων, που πρέπει να αποτελεί σημαντικό παράγοντα για την κατάταξη των αποτελεσμάτων αναζήτησης, μπορεί να εκφρασθεί σε όρους αριθμού και ποιότητας συνδέσμων και να προσδιορισθεί.

Ένα ενδιαφέρον πείραμα, το οποίο αποδείκνυε τη σημασία της συγκεκριμένης διαδικασίας και τη δύναμη των συνδέσμων, υπήρξε στο παρελθόν η τακτική του «Google Bombing» (βομβαρδισμός της Google), κατά την οποία εκατομμύρια χρήστες και διαχειριστές ιστοσελίδων χαρακτήριζαν, μέσω των συνδέσμων (με τεχνικές βελτιστοποίησης που αναφέρθηκαν στο προηγούμενο κεφάλαιο, και συγκεκριμένα που αφορούν την ετικέτα <a href>), αρνητικά κάποιο πρόσωπο, παραθέτοντας στο σύνδεσμο την προσωπική του ιστοσελίδα. Το αποτέλεσμα ήταν να εμφανίζεται στα αποτελέσματα των μηχανών, με όρο αναζήτησης τον συγκεκριμένο χαρακτηρισμό, η ιστοσελίδα αυτή σε αρκετά υψηλή θέση (συνήθως την πρώτη), παρότι η ιστοσελίδα προφανώς δεν ανέφερε πουθενά τον όρο αναζήτησης στο περιεχόμενο ή το σύνολο των εγγράφων του διακομιστή στον οποίο φιλοξενούνταν. Έκτοτε, οι μηχανές «διόρθωσαν» το εν λόγω πρόβλημα, χωρίς αυτό να σημαίνει ότι αφαίρεσαν τη βαρύτητα των εισερχόμενων συνδέσμων μίας σελίδας από τους αλγορίθμους κατάταξης, εξαλείφοντας μεμονωμένα περιστατικά και βελτιώνοντας τους αλγορίθμους ελέγχου συσχέτισης του όρου αναζήτησης με τις σελίδες.

Για τις ανάγκες μέτρησης αυτής της δημοτικότητας των σελίδων που απορρέει από τους συνδέσμους προς αυτές, ο ένας εκ των δημιουργών της μηχανής Google, Larry Page, εμπνεύστηκε έναν αλγόριθμο ανάλυσης συνδέσμων και τον αντίστοιχο βαθμό, που ονομάστηκε PageRank (από το όνομα του Larry Page).

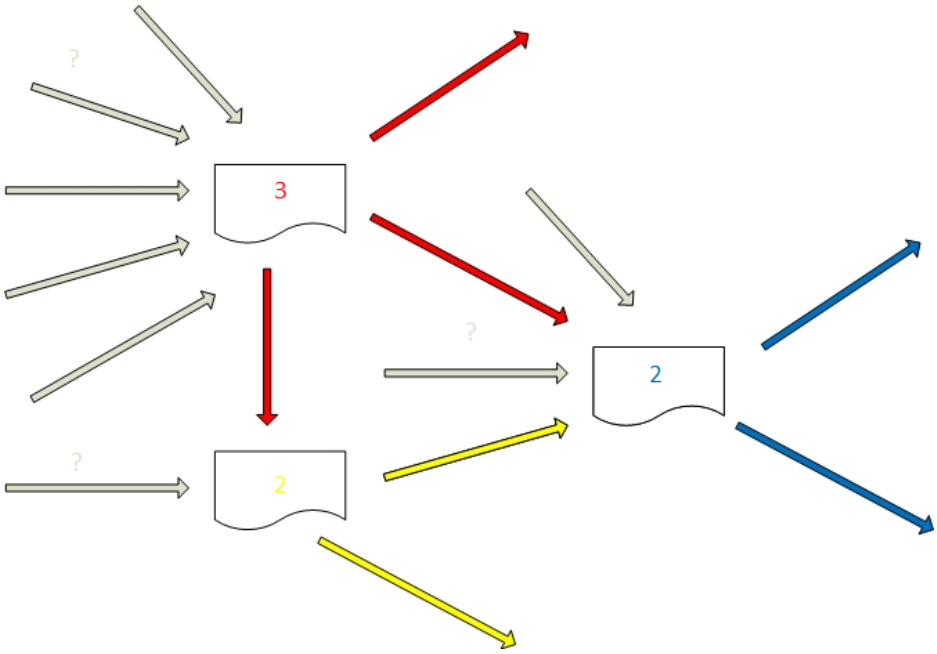
5.2 Ο βαθμός PageRank

Όπως χαρακτηριστικά αναφέρουν οι Page et al. (1999), για να μετρηθεί η σχετική σημασία των ιστοσελίδων, αναπτύχθηκε μία μέθοδος υπολογισμού της κατάταξης κάθε σελίδας στο Διαδικτυακό γράφο, ο αλγόριθμος PageRank. Πρόκειται για μία συνάρτηση πυκνότητας της πιθανότητας ένα φυσικό πρόσωπο που κάνει κλικ τυχαία σε συνδέσμους να φθάσει σε μία συγκεκριμένη σελίδα.

5.2.1 Η δομή του διαδικτυακού γράφου

Ο γράφος του Διαδικτύου αποτελείται από δισεκατομμύρια σελίδες (κόμβους) και πολύ περισσότερους συνδέσμους (ακμές) που συνδέουν τους κόμβους μεταξύ τους. Κάθε κόμβος του γράφου έχει ορισμένους εξερχόμενους και εισερχόμενους συνδέσμους, ενώ είναι προφανές, λόγω διαστάσεων και εποπτείας του Διαδικτύου, ότι είναι αδύνατο να γνωρίζουμε, ανά πάσα στιγμή, εάν έχουμε ακριβή αριθμό των εισερχόμενων συνδέσμων προς μία σελίδα, όμως μπορούμε να γνωρίζουμε όλους τους εξερχόμενους συνδέσμους από αυτήν, οποιαδήποτε στιγμή. Αυτό συμβαίνει διότι ένας σύνδεσμος εντοπίζεται όταν εξέρχεται από μία σελίδα, εφόσον μεταφορτωθεί και αναγνωσθεί η σελίδα αυτή, μέσω των <a href> ετικετών κι όχι όταν εισέρχεται σε μία σελίδα.

Έτσι, οι σελίδες διαφοροποιούνται ιδιαίτερα, όσον αφορά τους εισερχόμενους συνδέσμους που λαμβάνουν. Η δημιουργία εξερχόμενων συνδέσμων σε μία σελίδα δεν υποδεικνύει υψηλή ποιότητα για την ίδια τη σελίδα κι ελέγχεται αποκλειστικά από τους διαχειριστές της. Αντίθετα, η δημιουργία εισερχόμενων συνδέσμων πραγματοποιείται φυσικά, με το αντικειμενικό κριτήριο του ενδιαφέροντος των εξωτερικών ως προς τη σελίδα παραγόντων κι εκφράζει τη δημοτικότητά της. Έτσι, αποκτά ιδιαίτερο ενδιαφέρον ο φαινομενικά άγνωστος αριθμός των ακμών - συνδέσμων που εισέρχονται σε κάθε κόμβο – σελίδα:



Εικόνα 23 Η a priori κι εύκολη εποπτεία των εξερχόμενων συνδέσμων των σελίδων

Γενικά, σελίδες με μεγάλο αριθμό τέτοιων συνδέσμων είναι πιο σημαντικές για το χρήστη, άρα και τις μηχανές αναζήτησης, από άλλες που έχουν λιγότερους.

Παράλληλα, όμως, ο αλγόριθμος PageRank προχωράει ένα βήμα παραπάνω στην αξιολόγηση της δημοτικότητας των ιστοσελίδων, αξιολογώντας την ποιότητα των ίδιων των συνδέσμων. Υπάρχουν περιπτώσεις όπου κάθε ένας σύνδεσμος δεν ανταποκρίνεται σε ισόποση «σημασία» για μία σελίδα, όπως, για παράδειγμα, όταν μία σελίδα σχετική με την αεροναυτική λαμβάνει έναν σύνδεσμο από την ιστοσελίδα της NASA και μία διαφορετική σελίδα λαμβάνει περισσότερους συνδέσμους από διάφορα blogs. Είναι δεδομένο πως, σε αυτή την περίπτωση, ο ένας σύνδεσμος της NASA έχει μεγαλύτερη βαρύτητα από τους υπόλοιπους κι, ενδεχομένως, από το σύνολο όλων των υπόλοιπων συνδέσμων. Αυτό το πρόβλημα επιλύει η μέθοδος PageRank, αναζητώντας πώς μία επαρκής εκτίμηση της «σημασίας» μίας σελίδας μπορεί να εξαχθεί μόνο από τη δομή των συνδέσμων στο Διαδίκτυο.

Έτσι, αποδίδεται μία πιο σφαιρική περιγραφή στον όρο PageRank και τη λειτουργία του αλγορίθμου: μία σελίδα κατέχει υψηλό δείκτη θέσης στην ιεραρχία του διαδικτυακού γράφου εάν το άθροισμα των βαθμών των εισερχόμενων συνδέσμων προς αυτήν είναι υψηλό, ή, αλλιώς, μία σελίδα κατέχει υψηλό βαθμό PageRank εάν το άθροισμα των βαθμών PageRank των σελίδων που συνδέουν προς αυτήν είναι υψηλό.

5.2.2 Ο ορισμός του βαθμού PageRank

Έστω u μία ιστοσελίδα, F_u το σύνολο των σελίδων στις οποίες η u δείχνει και B_u το σύνολο των σελίδων που δείχνουν προς τη σελίδα u. Επίσης, έστω $N_u = |F_u|$ ο αριθμός των συνδέσμων από την u και c ένας παράγοντας κανονικοποίησης (έτσι ώστε ο συνολικός βαθμός όλων των ιστοσελίδων να είναι σταθερός).

Μία ελαφρώς απλουστευμένη μορφή του τύπου PageRank είναι η εξής:

$$Rank(u) = c\sum_{v\in B_u}\frac{Rank(v)}{N_v}$$

Με τον τρόπο αυτό, το Rank μίας οποιασδήποτε σελίδας διαιρείται εξίσου στους εξερχόμενους συνδέσμους της (εσωτερικούς και εξωτερικούς, δηλαδή όλους τους συνδέσμους που εξέρχονται από τον κόμβο, είτε αυτοί δείχνουν σε εξωτερικές σελίδες είτε δείχνουν σε σελίδες που βρίσκονται στον ίδιο ιστότοπο) για να συνεισφέρουν στο συνολικό Rank των σελίδων στις οποίες δείχνουν.

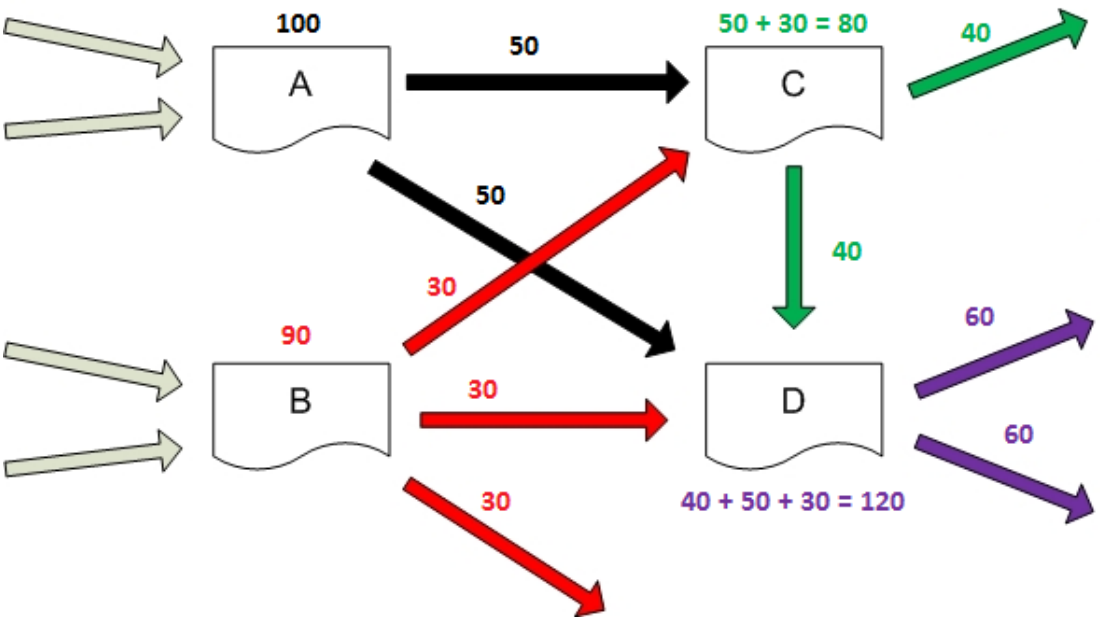
Ο παράγοντας κανονικοποίησης c διορθώνει το πρόβλημα που προκύπτει όταν μία σελίδα λαμβάνει συνδέσμους, δηλαδή υπάρχουν ακμές οι οποίες δείχνουν σε αυτήν, αλλά δεν υπάρχουν σελίδες στις οποίες να δείχνει η ίδια και το βάρος των συνδέσμων που συνεισφέρουν στο Rank της σελίδας αυτής χάνεται.

Έτσι, έχουμε:

$$c < 1$$

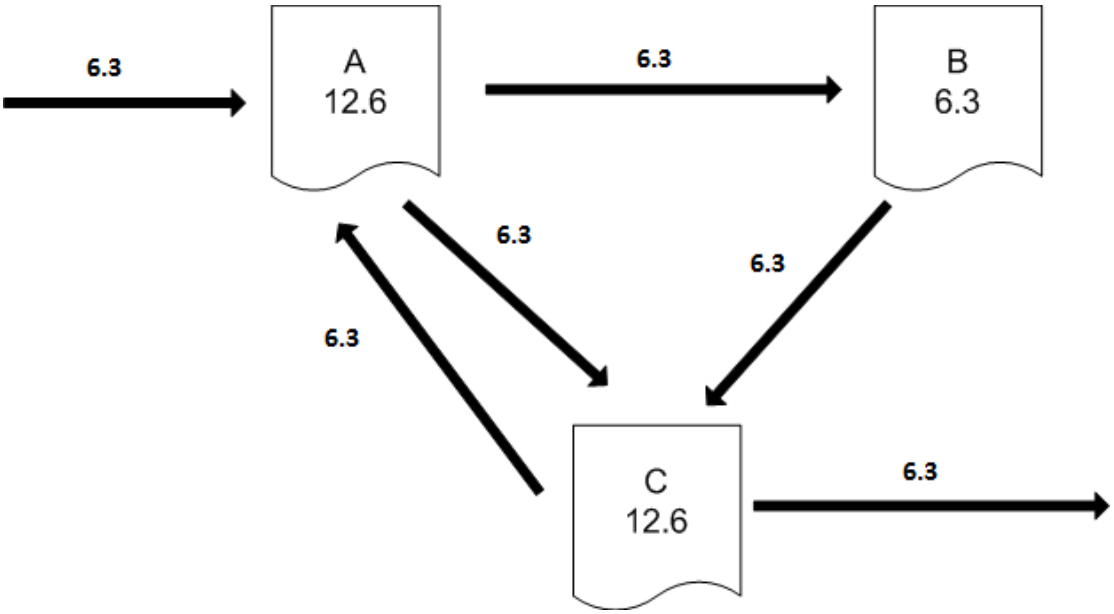
Ο παραπάνω τύπος είναι αναδρομικός, όμως μπορεί να χρησιμοποιηθεί για τον υπολογισμό των Ranks, εάν ξεκινήσουμε με μία αρχική μήτρα των Ranks των σελίδων κι επαναλαμβάνοντας τους υπολογισμούς μέχρι να επιτευχθεί σύγκλιση.

Έτσι, ξεκινώντας από μία αρχική κατάσταση για τις δύο σελίδες A, B, π.χ. Rank (A) = 100 και Rank (B) = 90, μπορούμε να υπολογίσουμε, στο ίδιο βήμα, το Rank των σελίδων C και D:



Εικόνα 24 Παράδειγμα υπολογισμού του Rank, δεδομένης αρχικής κατάστασης

Η σύγκλιση επιτυγχάνεται όταν βρεθεί λύση σταθερής ισορροπίας για ένα σύνολο σελίδων. Για παράδειγμα, στο ν-οστό βήμα της αναδρομής, υπολογίζεται το Rank των τριών σελίδων A, B και C:

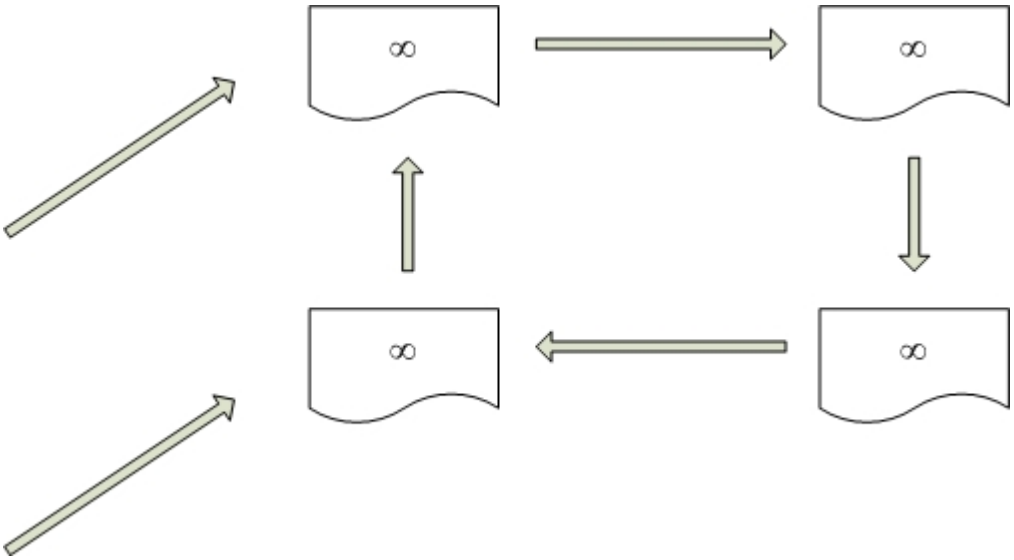


Εικόνα 25 Ισορροπία του Rank των σελίδων και σύγκλιση του αλγορίθμου

Παρατηρούμε ότι στο παραπάνω παράδειγμα, όλοι οι σύνδεσμοι φέρουν ισόποση αξία, καθώς, στο συγκεκριμένο σύνολο των τριών σελίδων, κάθε σελίδα έχει ίσο αριθμό εισερχόμενων και εξερχόμενων συνδέσμων – ακμών.

Εναλλακτικά, η παραπάνω απλουστευμένη εκδοχή του βαθμού PageRank ορίζεται ως εξής: Έστω $\boldsymbol{A}$ ένας τετραγωνικός πίνακας με γραμμές και στήλες τις ιστοσελίδες. Έστω $A_{u,v} = 1/N_u$ εάν υπάρχει μία ακμή από τη σελίδα u προς τη σελίδα v και $A_{u,v} = 0$ διαφορετικά. Εάν $\boldsymbol{R} = \boldsymbol{Rank}$ είναι ένα διάνυσμα ιστοσελίδων, τότε ορίζεται $\boldsymbol{R} = c\boldsymbol{AR}$, και το $\boldsymbol{R}$ είναι ένα ιδιοδιάνυσμα του πίνακα $\boldsymbol{A}$, με ιδιοτιμή τον παράγοντα κανονικοποίησης c.

Οι δύο ερευνητές του Stanford παρατήρησαν ότι ο παραπάνω τύπος παρουσίαζε προβληματική συμπεριφορά στην περίπτωση που δύο ή περισσότερες σελίδες αντάλλαζαν συνδέσμους αναμεταξύ τους και αποκλειστικά, ενώ μία άλλη σελίδα συνέδεε σε μία εκ των δύο αυτών ιστοσελίδων, με αποτέλεσμα να συγκεντρώνεται βαθμός (Rank) στις δύο σελίδες, ανά τις επαναλήψεις, χωρίς να κατανέμεται στη συνέχεια αυτός ο βαθμός. Το φαινόμενο αυτό ονομάστηκε «νεροχύτης βαθμού» (rank sink), όπως φαίνεται στο παρακάτω παράδειγμα, μετά από έναν σημαντικό αριθμό επαναλήψεων του βρόχου:



Εικόνα 26 Απλουστευμένο παράδειγμα ενός rank sink

Για την αντιμετώπισή του φαινομένου αυτού, επινοήθηκε μία πηγή βαθμού *E(u)*, με τη χρήση της οποίας ορίστηκε ο τύπος υπολογισμού PageRank.

Έστω *E(u)* κάποιο διάνυσμα ιστοσελίδων που αντιστοιχεί σε μία πηγή βαθμού. Τότε, ο βαθμός PageRank ενός συνόλου ιστοσελίδων είναι η τιμή που ανατίθεται στις ιστοσελίδες αυτές η οποία ικανοποιεί την αναδρομική σχέση

$$PageRank(u) = c\sum_{v\in B_u}\frac{PageRank(v)}{N_v} + cE(u)$$

έτσι ώστε ο παράγοντας κανονικοποίησης *c* να μεγιστοποιείται και η L_1 νόρμα του PageRank να ισούται με τη μονάδα, δηλαδή:

$$\left\|PageRank\right\|_1 = 1$$

Αποδεικνύεται ότι ο βαθμός PageRank είναι ένα ιδιοδιάνυσμα του (***A*** + ***E*** *x* ***1***), όπου **1** το μοναδιαίο διάνυσμα.

5.2.3 Ο υπολογισμός του βαθμού PageRank

Έστω S οποιοδήποτε διάνυσμα ιστοσελίδων (όπως, για παράδειγμα, είναι και το διάνυσμα πηγής βαθμού *E*). Τότε, το PageRank (*R'*) μπορεί να ευρεθεί, για τη βέλτιστη δυνατή σύγκλιση, ως εξής:

$$R'_0 \leftarrow S$$

while $\delta > 0$:

$$R'_{i+1} \leftarrow AR'_i$$

$$d \leftarrow \left\| R'_i \right\|_1 - \left\| R'_{i+1} \right\|_1$$

$$R'_{i+1} \leftarrow R'_{i+1} + dE$$

$$d \leftarrow \left\| R'_{i+1} - R'_i \right\|_1$$

Σύμφωνα με τους Brin και Page (1998), στη δημοσίευσή τους «The Anatomy of a Large-Scale Hypertextual Web Search Engine» που πραγματοποιήθηκε στα πλαίσια της προσπάθειάς τους να αναπτύξουν μία νέα βελτιωμένη μηχανή αναζήτησης που αργότερα έγινε γνωστή ως μηχανή αναζήτησης Google, ο βαθμός PageRank δεν περιορίζεται στον υπολογισμό του αριθμού των εισερχόμενων συνδέσμων μίας σελίδας, με σκοπό την εκτίμηση της δημοτικότητας της ιστοσελίδας αυτής, αλλά επεκτείνει την ιδέα της οργάνωσης του Διαδικτύου, αποδίδοντας διαφορετική βαρύτητα (ποιότητα) σε κάθε σύνδεσμο που δείχνει στη σελίδα και κανονικοποιώντας την αξία αυτής με κριτήριο τον αριθμό των συνδέσμων που ξεκινούν από μία ιστοσελίδα.

Υποθέτουμε, λοιπόν, ότι υπάρχουν $T_1, T_2, \ldots, T_v$ σελίδες που συνδέουν στη σελίδα Α. Έστω η παράμετρος *d* ένας συντελεστής απόσβεσης που μπορεί να πάρει τιμές στο διάστημα [0,1] και που, συνήθως, λαμβάνεται η τιμή $d = 0,85$ και ως *C(A)* ορίζεται ως ο αριθμός των εξερχόμενων συνδέσμων της σελίδας Α. Τότε, ο βαθμός *PageRank* της σελίδας Α, *PR(A)*, υπολογίζεται ως εξής:

$$PR(A) = (1-d) + d[PR(T_1)/C(T_1) + \ldots + PR(T_v)/C(T_v)]$$

ή αλλιώς:

$$PageRank(A) = (1-d) + d\sum_{i=1}^{n} \frac{PageRank(T_i)}{C(T_i)}$$

Έτσι, η συγκεκριμένη τιμή του βαθμού PageRank μπορεί να υπολογιστεί με τον αναδρομικό αλγόριθμο που ορίστηκε παραπάνω και αντιστοιχεί στο θεμελιώδες ιδιοδιάνυσμα του κανονικοποιημένου πίνακα των συνδέσμων του Διαδικτύου, ενώ εκτιμάται ότι ο βαθμός PageRank για τα δισεκατομμύρια των σελίδων που υπάρχουν στο Διαδίκτυο μπορεί να υπολογιστεί σε διάστημα λίγων ωρών, δεδομένων των σύγχρονων τεχνολογικών πόρων.

Ένα εκ νέου πρόβλημα που παρουσιάζεται σε αυτή την τελική μορφή υπολογισμού του βαθμού PageRank είναι οι «αδιέξοδοι σύνδεσμοι» (dangling links). Πρόκειται για συνδέσμους που δείχνουν προς μία σελίδα, η οποία στη συνέχεια δεν περιέχει εξερχόμενους συνδέσμους, δηλαδή είναι κόμβος με αποκλειστικά εισερχόμενες ακμές. Σε αυτή την

κατηγορία συγκαταλέγονται και οι σύνδεσμοι που δείχνουν σε σελίδες που δεν έχουν μεταφορτωθεί και αναλυθεί ακόμη για τον προσδιορισμό των εξερχόμενων συνδέσμων τους. Για την αντιμετώπιση του προβλήματος αυτού, παρατηρήθηκε ότι η αφαίρεση των συνδέσμων αυτών από τον γράφο επιτρέπει τον υπολογισμό του βαθμού PageRank των ιστοσελίδων, καθότι δεν επηρεάζουν το βαθμό των υπόλοιπων σελίδων. Στη συνέχεια και μόλις οι δείκτες PageRank των σελίδων έχουν υπολογισθεί, οι αδιέξοδοι σύνδεσμοι επανατοποθετούνται στο διαδικτυακό γράφο, επιφέροντας έτσι ορισμένες ασήμαντες αλλαγές στην παράμετρο κανονικοποίησης των άλλων συνδέσμων που βρίσκονται στις ίδιες σελίδες με αυτούς. Ο αντίκτυπος αυτών των αλλαγών θεωρείται αμελητέος.

5.2.4 Το μοντέλο του τυχαίου χρήστη

Ο βαθμός PageRank μπορεί να θεωρηθεί ως ένα μοντέλο συμπεριφοράς ενός χρήστη. Υποθέτοντας ότι υπάρχει ένας «τυχαίος χρήστης», ο οποίος βρίσκεται σε μία τυχαία αρχική σελίδα και στοχαστικά συνεχίζει να κάνει κλικ σε συνδέσμους που συναντάει, χωρίς να πατάει «πίσω» αλλά, εν τέλει, βαριέται και ξεκινάει από μία νέα, διαφορετική, τυχαία σελίδα. Ο βαθμός PageRank μίας συγκεκριμένης σελίδας είναι η συνάρτηση πυκνότητας της πιθανότητας ο χρήστης να επισκεφθεί τη σελίδα αυτή, ενώ ο συντελεστής απόσβεσης είναι η πιθανότητα, για κάθε σελίδα, ο χρήστης να βαρεθεί αυτή τη σελίδα και να αιτηθεί την επανέναρξη από μία νέα τυχαία σελίδα.

Συγκεκριμένα, ένας πραγματικός χρήστης, όταν εισαχθεί σε έναν ατέρμονα βρόχο σελίδων, είναι απίθανο να συνεχίσει να κάνει κλικ στους συνδέσμους, οπότε θα επιλέξει να επισκεφθεί μία εκ νέου τυχαία αφετηρία για την επανέναρξη της πλοήγησής του στο Διαδίκτυο. Ο παράγοντας ***E***, όπως ορίσθηκε παραπάνω, μπορεί να θεωρηθεί ως ένας τρόπος μοντελοποίησης αυτής της συμπεριφοράς, κατά την οποία ο χρήστης περιοδικά «βαριέται» και παύει να κάνει κλικ διαδοχικά στους συνδέσμους, επιλέγοντας νέα αφετηρία, η οποία επιλέγεται βάσει της κατανομής του ***E***. Έτσι, ο εν λόγω παράγοντας θεωρείται μία παράμετρος που μπορεί να οριστεί από το χρήστη, ενώ διαφορετικές τιμές του E μπορούν να έχουν ως αποτέλεσμα εξατομικευμένα συστήματα PageRank.

Παράλληλα με την επιλογή του παράγοντα ***E***, εξατομίκευση του PageRank μπορεί να επιτευχθεί και με την επιλεκτική πρόσθεση του παράγοντα απόσβεσης *d* σε μεμονωμένες σελίδες ή ομάδες σελίδων, όπως θα αναλυθεί παρακάτω.

Μία εναλλακτική διαισθητική επεξήγηση του όρου είναι ότι μία σελίδα μπορεί να έχει υψηλό PageRank εάν υπάρχουν πολλές σελίδες που δείχνουν σε αυτήν, ή εάν υπάρχουν λίγες σελίδες που δείχνουν σε αυτήν αλλά και που οι ίδιες έχουν υψηλό βαθμό PageRank. Αναλυτικότερα, σελίδες που ενδεχομένως έχουν μόνο έναν εισερχόμενο σύνδεσμο από μία

πολύ σημαντική ιστοσελίδα (όπως, για παράδειγμα, την αρχική σελίδα μίας μηχανής αναζήτησης ή της Wikipedia), συνήθως αξίζουν την επίσκεψη των χρηστών διότι είναι υψηλής ποιότητας. Εάν δεν ήταν υψηλής ποιότητας, τότε η σημαντική αυτή ιστοσελίδα δε θα περιελάμβανε σύνδεσμο προς αυτήν. Ο βαθμός PageRank διαχειρίζεται αποτελεσματικά αυτές τις περιπτώσεις, κατανέμοντας διαρκώς και αναδρομικά βάρη σε ολόκληρη τη δομή των διαδικτυακών συνδέσμων.

5.2.5 Εφαρμογή του αλγορίθμου

Για την εφαρμογή του αλγορίθμου, οι μηχανικοί της Google μετέτρεψαν κάθε διεύθυνση URL σε ένα μοναδικό ακέραιο και αποθήκευσαν κάθε σύνδεσμο σε μία βάση δεδομένων, χρησιμοποιώντας το σύστημα των ακεραίων για την ταυτοποίηση των σελίδων. Στη συνέχεια, αφαιρέθηκαν, όπως αναλύθηκε παραπάνω, οι αδιέξοδοι σύνδεσμοι. Το σημαντικό βήμα για τον υπολογισμό των βαθμών PageRank ήταν η αρχικοποίηση αυτών, δηλαδή η κατάλληλη επιλογή της αρχικής κατάστασης. Εφόσον η διαδικασία, όμως, επαναλαμβάνεται μέχρις ότου επιτευχθεί η σύγκλιση της μεθόδου, είναι προφανές ότι η επιλογή των αρχικών τιμών δεν επηρεάζει τις τελικές τιμές καθαυτές, παρά μόνο τον αριθμό των επαναλήψεων, δηλαδή την ταχύτητα της σύγκλισης. Επομένως, η βελτιστοποίηση του αλγορίθμου εξαρτάται σε μεγάλο βαθμό από την επιλογή της αρχικής τιμής PageRank των ιστοσελίδων, όχι, όμως, ως προς το τελικό αποτέλεσμα.

Μετά τη σύγκλιση των βαρών των συνδέσμων, επανεισάγονται οι αδιέξοδοι σύνδεσμοι στο γράφο, εφαρμόζοντας τον αλγόριθμο τόσες φορές όσες ήταν και οι επαναλήψεις που απαιτήθηκαν για την απαλοιφή των αδιέξοδων συνδέσμων, για να αποφευχθεί η ανάθεση μηδενικών τιμών σε κάποιους εξ αυτών.

Σύμφωνα με τους Page et al. (1999), το κόστος υπολογισμού του βαθμού PageRank όλων των ιστοσελίδων είναι αμελητέο, εν συγκρίσει με το κόστος που συνοδεύει την κατασκευή ενός ευρετηρίου πλήρους κειμένου για τις ιστοσελίδες.

Τέλος, σύμφωνα με τη δοκιμή του αλγορίθμου σε βάση δεδομένων 322 εκατομμυρίων συνδέσμων, αριθμός που ανταποκρινόταν πλήρως στα δεδομένα του 1998, η ταχύτητα σύγκλισης ήταν περίπου 52 επαναλήψεις, ενώ η συνάρτηση της διαφοράς μεταξύ των επαναλήψεων και του αριθμού των επαναλήψεων, σε ημιλογαριθμική κλίμακα, είναι σχεδόν γραμμική.

5.2.6 Εξατομίκευση του βαθμού PageRank

Όπως αναφέρθηκε κατά τον ορισμό του PageRank, το διάνυσμα ***E*** των ιστοσελίδων χρησιμοποιείται ως η πηγή βαθμών για να αποφευχθεί το προβληματικό φαινόμενο των rank sinks. Πέραν αυτής της χρησιμότητάς του, όμως, ο παράγοντας ***E*** απεδείχθη μία σημαντική παράμετρος με μεγάλες δυνατότητες «ρύθμισης» του βαθμού των σελίδων.

Συγκεκριμένα, μπορούμε να αποδώσουμε την ίδια ακριβώς τιμή σε όλες τις σελίδες, με $\|E\|_1 = a$, που εκφράζει ότι ένας τυχαίος χρήστης «βαριέται» και επανακινεί τη διαδικασία της τυχαίας πλοήγησης (σύμφωνα με το μοντέλο του τυχαίου χρήστη) περιοδικά, κάτι που θεωρείται δίκαιο και αντικειμενικό, καθώς όλες οι ιστοσελίδες αξιολογούνται αρχικά απλά και μόνο επειδή υπάρχουν. Το μοναδικό πρόβλημα που παρατηρείται σε αυτή την, κατά τα άλλα, επιτυχή τακτική είναι αυτό κατά το οποίο ορισμένες σελίδες αποκτούν υπερβολικά μεγάλο βαθμό.

Μία άλλη ακραία περίπτωση περιλαμβάνει την αντιστοίχιση του ***E*** αποκλειστικά σε μία ιστοσελίδα, που ταυτίζεται με την προεπιλογή της ιστοσελίδας του τυχαίου χρήστη, όπως για παράδειγμα η ιστοσελίδα της Google, ένα ειδησεογραφικό μπλογκ, ή η αρχική σελίδα της σχολής των Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Είναι προφανές ότι η διαδοχή των συνδέσμων που ακολουθεί τυχαία ο χρήστης του μοντέλου δε θα μπορούσε να ταυτιστεί.

Αφενός, λοιπόν, η ανάγκη διαφοροποίησης της πιθανότητας ο χρήστης να «βαρεθεί» ανά τις διαφορετικές ιστοσελίδες, αφετέρου τα διαφορετικά ενδιαφέροντα, όπως εκφράζονται από τις διευθύνσεις που επιλέγει να εκκινήσει την πλοήγησή του, του κάθε χρήστη, θέτουν τις ανάγκες αλλά και τις **μελλοντικές προοπτικές εξατομίκευσης του βαθμού PageRank**, με πιθανές εφαρμογές σε προσωπικές μηχανές αναζήτησης, προχωρώντας ένα βήμα πιο κοντά στον **Σημασιολογικό Ιστό.**

Μία πιθανή χρυσή τομή μεταξύ των δύο περιπτώσεων, της ανάθεσης της ίδιας τιμής $\|E\|_1 = a$ και της αντιστοίχισης του E με μία συγκεκριμένη ιστοσελίδα, αποτελεί η αντιστοίχιση του παράγοντα E με όλες τις αρχικές σελίδες κάθε ιστοχώρου. Η περίπτωση αυτή, προφανώς, δίνει μικρά περιθώρια χειραγώγησης του βαθμού PageRank, καθώς κάποιος που θα επιθυμούσε να «ξεγελάσει» το σύστημα του προσαρμοσμένου βαθμού PageRank, θα μπορούσε να κατασκευάσει πολλούς ιστοτόπους, η αρχική σελίδα όλων των οποίων θα έδειχνε προς μία συγκεκριμένη επιθυμητή σελίδα. Την ίδια, όμως, δυνατότητα θα του έδινε ακριβώς η ίδια μέθοδος, εάν επρόκειτο και για τον αρχικό, μη προσαρμοσμένο βαθμό PageRank, καθώς η συγκεκριμένη σελίδα θα συγκέντρωνε μεγάλο βαθμό από τους πολλούς εισερχόμενους συνδέσμους.

Αρκετά μετά την κατασκευή και χρήση του αλγορίθμου PageRank, οι ερευνητές Haveliwala et al. (2003) προσέγγισαν μαθηματικά μοντέλα προσαρμογής κι εξατομίκευσης του αλγορίθμου.

Για τον σκοπό αυτό, εξέφρασαν τον ορισμό του PageRank εναλλακτικά, ως εξής: Συμβολίζουμε με $u \to v$ το γεγονός η σελίδα u του διαδικτυακού γράφου *G* να δείχνει στη σελίδα *v*, δηλαδή την ύπαρξη ακμής από την σελίδα *u* προς την σελίδα *v*, ενώ με $\deg(u)$ το σύνολο των εξερχόμενων ακμών από την σελίδα *u*. Έστω ότι ο γνωστός «τυχαίος χρήστης» του Διαδικτύου επισκέπτεται στην σελίδα *u* την αρχική στιγμή $t_0 = k$, επομένως, βάσει του μοντέλου, την επόμενη χρονική στιγμή $k+1$, ο χρήστης φθάνει στον κόμβο $v_i \in \{v \mid u \to v\}$, με πιθανότητα $1/\deg(u)$. Ο βαθμός PageRank της σελίδας *i* θα ορίζεται ως η πιθανότητα ότι σε κάποια συγκεκριμένη χρονική στιγμή $k > K$ ο χρήστης θα βρίσκεται στη σελίδα *i*. Με ορισμένες μικρές τροποποιήσεις στο μοντέλο και K επαρκώς υψηλό, η πιθανότητα αυτή είναι μοναδική.

Εάν εκφραστεί ο «τυχαίος» αυτός περίπατος στον γράφο *G* του Διαδικτύου ως αλυσίδα Markov, τότε οι διάφορες καταστάσεις της θα δίνονται από τους κόμβους του *G* και η πιθανοτική (στοχαστική) μετάβαση από τη σελίδα *i* στη σελίδα *j* θα δίνεται από την πιθανότητα $P_{ij} = \frac{1}{\deg(i)}$, η οποία θα καθορίζει και τα στοιχεία του αντίστοιχου πίνακα μεταβάσεων.

Για απεριοδική και αμείωτη πιθανότητα *P* που διασφαλίζει τη μοναδικότητα της κατανομής, ορίζεται μία νέα, αμείωτη αλυσίδα Markov (Motwani & Paghavan, 1995) ως εξής:

$$A = [cP + (1-c)E]^T$$

Στην αλυσίδα αυτή, ο *E* είναι πίνακας $n \times n$ που ορίζεται ως $E = ev^T$ (όπου *e* είναι το *n* – διάστατο μοναδιαίο διάνυσμα και *v* είναι ένα *n* – διάστατο με θετικά στοιχεία, το άθροισμα των οποίων είναι ίσο με τη μονάδα), ενώ ως $(1-c)$ ορίζεται η πιθανότητα ο τυχαίος χρήστης να «βαρεθεί», σε κάθε χρονική στιγμή, και να προχωρήσει σε επανεκκίνηση της στοχαστικής πλοήγησής του.

Ο προορισμός του τυχαίου άλματος, που προκύπτει από την ύπαρξη του *E* και ονομάζεται τηλεκατεύθυνση, καθορίζεται από την κατανομή της πιθανότητας που αφορά τη σελίδα *v*. Στην περίπτωση που το διάνυσμα *v* είναι ανομοιόμορφο, ο παράγοντας *E* προσθέτει τεχνητές μεταβάσεις με μη ομοιόμορφη πιθανοτική κατανομή και ο βαθμός PageRank που προκύπτει παύει να είναι αμερόληπτο στο σύνολο των ιστοσελίδων. Έτσι το διάνυσμα *v* αναφέρεται ως **διάνυσμα εξατομίκευσης**.

Έτσι, έστω n ο αριθμός των σελίδων στο Διαδίκτυο, *x(ν)* το διάνυσμα *n* – διαστάσεων του εξατομικευμένου PageRank που αντιστοιχεί στο *n* – διαστάσεων διάνυσμα εξατομίκευσης *ν*. Επιλύοντας το πρόβλημα ιδιοτιμής:

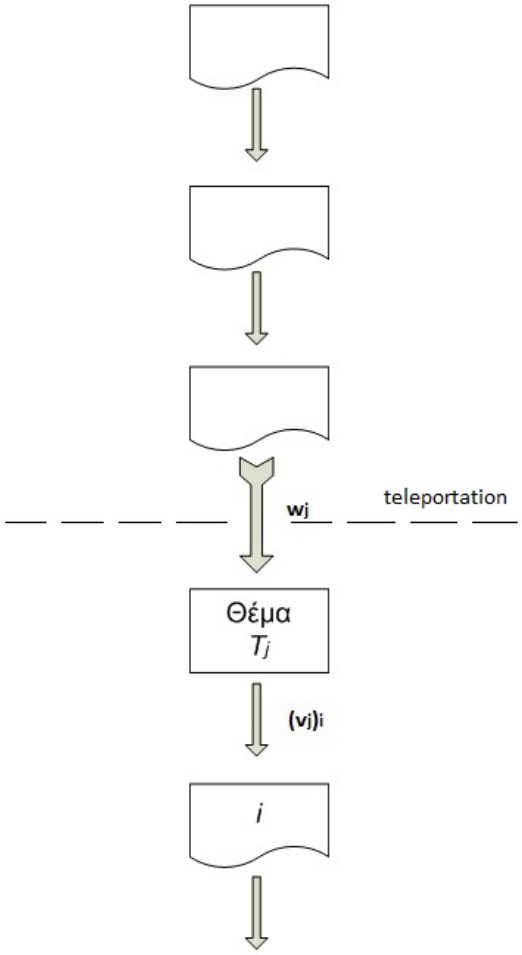
$$\begin{aligned} x &= Ax \\ x &= [cP^T + (1-c)E^T]x \\ x - cP^T x &= (1-c)v \\ (I - cP^T)x &= (1-c)v \\ x &= (1-c)(I - cP^T)^{-1}v \end{aligned}$$

Θεωρώντας $Q = (1-c)(I - cP^T)^{-1}$, που προκύπτει από την παραπάνω εξίσωση (δεδομένου ότι ο $(I - cP)$ είναι διαγωνίως δεσπόζων πίνακας άρα και αντιστρέψιμος, άρα και ο πίνακας $(I - cP)^T = (I - cP^T)$ είναι επίσης αντιστρέψιμος), και θέτοντας $v = e_i$, παρατηρούμε ότι η *i* – οστή στήλη του πίνακα *Q* είναι $x(e_i)$, δηλαδή το διάνυσμα του εξατομικευμένου PageRank που αντιστοιχεί στο διάνυσμα εξατομίκευσης e_i. Επομένως, οι στήλες του πίνακα *Q* περιλαμβάνουν μία πλήρη βάση για διανύσματα εξατομικευμένου βαθμού PageRank, έτσι όπως προσεγγίστηκε από τους Brin και Page (1998), στην αρχική τους δημοσίευση περί του αλγορίθμου. Παρότι δεν καθίσταται εφικτός ο ακριβής υπολογισμός του *Q*, έχουμε τη δυνατότητα να πραγματοποιήσουμε προσεγγίσεις χαμηλού βαθμού του *Q*, που συμβολίζουμε ως $\hat{Q}$.

Οι διάφορες μαθηματικές προσεγγίσεις που ακολούθησαν οι τρεις ερευνητές (Haveliwala et al., 2003) οδήγησαν σε τρεις διαφορετικούς εξατομικευμένους βαθμούς PageRank των σελίδων, ο Topic – Sensitive PageRank (βαθμός με διάκριση θεματολογίας), ο Modular PageRank (βαθμός με διάκριση των υψηλόβαθμων σελίδων) και ο BlockRank (βαθμός με διάκριση μπλοκ σελίδων).

5.2.6.1 Topic – Sensitive PageRank

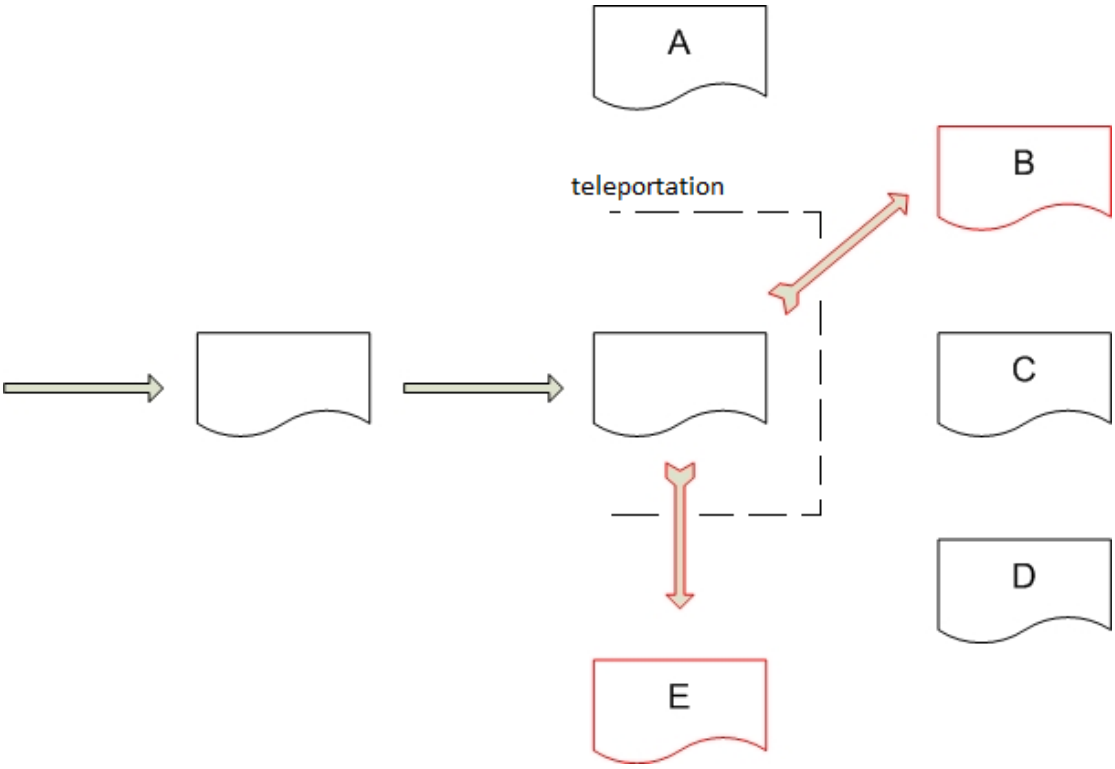
Στη συγκεκριμένη εκδοχή του βαθμού PageRank, η προσέγγιση $\hat{Q}$ υπολογίζεται εκτός σύνδεσης. Σε όρους του μοντέλου του τυχαίου χρήστη, ο Topic – Sensitive PageRank περιορίζει την επιλογή των μεταβάσεων τηλεκατεύθυνσης (δηλαδή, όχι αυτών που συμβαίνουν διαδοχικά, αλλά λόγω «βαρεμάρας» του χρήστη κι επανεκκίνησης της πλοήγησης), έτσι ώστε ο χρήστης να μπορεί να τηλεμεταφερθεί σε ένα συγκεκριμένο θέμα T_j, με πιθανότητα w_j, κι ακολούθως σε μία συγκεκριμένη σελίδα *i*, με πιθανότητα $(v_j)_i$.



Εικόνα 27 Ο τυχαίος χρήστης στον αλγόριθμο Topic-Sensitive PageRank

5.2.6.2 Modular PageRank

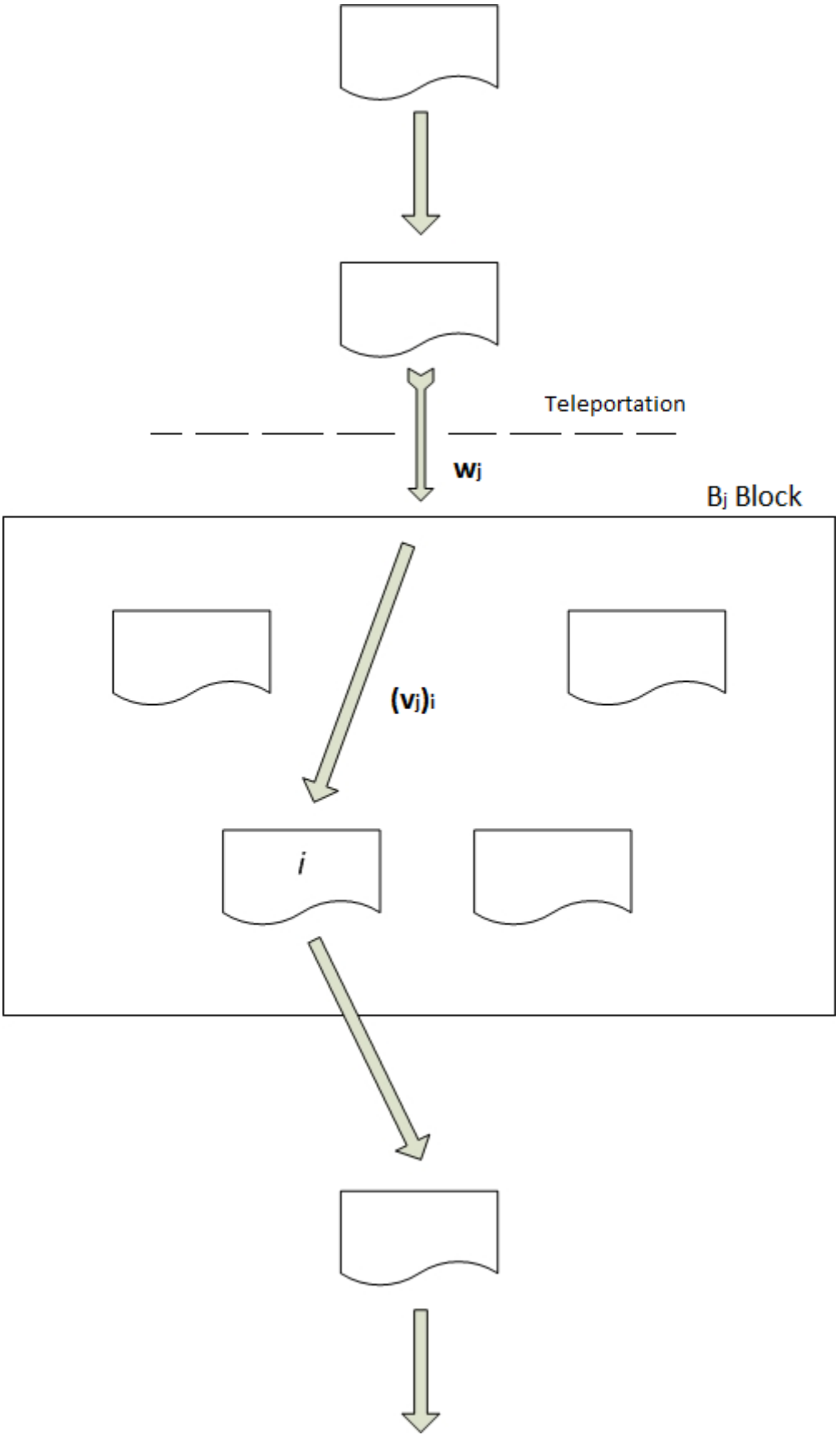
Η παραλλαγή Modular PageRank υπολογίζει έναν $n \times k$ πίνακα, χρησιμοποιώντας τις k στήλες του Q που αντιστοιχούν σε σελίδες με υψηλό βαθμό. Σε όρους του μοντέλου του τυχαίου χρήστη του PageRank, ο εν λόγω εξατομικευμένος βαθμός PageRank περιορίζει την επιλογή των μεταβάσεων τηλεκατεύθυνσης, έτσι ώστε ο τυχαίος χρήστης να μπορεί να τηλεμεταφερθεί σε συγκεκριμένες υψηλόβαθμες σελίδες, αντί των αυθαίρετα επιλεγμένων τυχαίων σελίδων. Στο παρακάτω διάγραμμα, τη στιγμή που «βαριέται», ο τυχαίος χρήστης επιλέγει μία εκ των σελίδων B και E, λόγω του υψηλότερου βαθμού που αυτές διαθέτουν έναντι των A, C και D:



Εικόνα 28 Το μοντέλο του τυχαίου χρήστη, κατά τον αλγόριθμο Modular PageRank

5.2.6.3 BlockRank

Ομοίως με τον προηγούμενο τύπο PageRank, ο BlockRank υπολογίζει έναν πίνακα $n \times k$ που αντιστοιχεί σε k «μπλοκ». Κάθε μπλοκ j αντιστοιχεί σε μπλοκ σελίδων που ομαδοποιούνται με κριτήριο ένα χαρακτηριστικό, π.χ. ο τομέας, ή ο εξυπηρετητής και περιλαμβάνει έναν τοπικό βαθμό PageRank, έστω v_j. Το διάνυσμα $x(v_j)$ αντιστοιχίζεται στην j στήλη του πίνακα $\hat{Q}$ της προσέγγισης. Σε όρους του μοντέλου μας, ο βαθμός αυτός περιορίζει την επιλογή των μεταβάσεων τηλεκατεύθυνσης, έτσι ώστε ο τυχαίος χρήστης να μπορεί να τηλεμεταφερθεί στο μπλοκ B_j, με πιθανότητα w_j, κι ακολούθως σε μία συγκεκριμένη σελίδα i του μπλοκ B_j, με πιθανότητα $(v_j)_i$.



Εικόνα 29 Η συμπεριφορά του τυχαίου χρήστη, στο μοντέλο του BlockRank

5.2.7 Άλλες χρήσεις του αλγορίθμου PageRank

Μία σπουδαία ιδιότητα του PageRank είναι ότι αποτελεί ένα εξαιρετικό εργαλείο πρόγνωσης των εισερχόμενων συνδέσμων που υπάρχουν στο διαδικτυακό γράφο (Cho et al., 1998). Αποδεικνύεται ότι ο αλγόριθμος PageRank είναι πιο αποδοτικός στην πρόγνωση των μελλοντικών εισερχόμενων συνδέσμων από ό,τι είναι τα στατιστικά στοιχεία των συνδέσμων καθαυτά. Το πείραμα που πραγματοποιήθηκε στο πανεπιστήμιο του Stanford υποθέτει ότι το σύστημα ξεκινάει με μόνο μία διεύθυνση URL και καμία άλλη πληροφορία, και ο στόχος είναι η ανίχνευση των σελίδων με την καλύτερη δυνατή σειρά, δηλαδή η **βελτιστοποίηση της ανίχνευσης** των ιστοσελίδων του Διαδικτύου, που σημαίνει ανίχνευση των σελίδων με κριτήριο το βαθμό τους, βάσει μίας συνάρτησης αξιολόγησης. Έτσι, θεωρούμε ως συνάρτηση αξιολόγησης τον αριθμό των εισερχόμενων συνδέσμων, δεδομένων όλων των πληροφοριών, καθώς όλες οι πληροφορίες είναι διαθέσιμες αφότου όλα τα έγγραφα έχουν ανιχνευθεί. Απεδείχθη ότι, έχοντας στη διάθεσή μας ατελείς πληροφορίες για τις σελίδες, ο αλγόριθμος PageRank αποτελεί έναν πιο αποδοτικό τρόπο για την οργάνωση και τον προσανατολισμό της ανίχνευσης από ό,τι ο αριθμός των εισερχόμενων συνδέσμων (backlinks). Με άλλα λόγια, ο PageRank είναι μία καλύτερη προγνωστική εφαρμογή από την καταμέτρηση των εισερχόμενων συνδέσμων, ακόμη κι αν ο στόχος είναι ο υπολογισμός του αριθμού των εισερχόμενων συνδέσμων. Η αιτιολόγηση βρίσκεται στο γεγονός ότι ο αλγόριθμος PageRank αγνοεί τα τοπικά μέγιστα στα οποία η ευθεία καταμέτρηση των backlinks εγκλωβίζεται, προσπαθώντας να ανιχνεύσει τοπικές συλλογές εγγράφων και σπαταλώντας πολύ χρόνο για να απεγκλωβιστεί και να εντοπίσει άλλες σελίδες με υψηλό αριθμό εισερχόμενων συνδέσμων. Έτσι, ο PageRank αποδεικνύεται πιο αποδοτικός στην ταξινομημένη, βάσει βαθμού – δημοτικότητας, χαρτογράφηση του Διαδικτύου.

Μία άλλη ιδιότητα του αλγορίθμου PageRank είναι η εκτίμηση της διαδικτυακής κίνησης (web traffic). Φαίνεται πως σελίδες με αρκετά υψηλά ποσοστά επισκεψιμότητας (όπως σελίδες πορνογραφικού περιεχομένου, για παράδειγμα) δε συγκεντρώνουν αρκετά υψηλό βαθμό PageRank, κάτι που, εν μέρει, είναι λογικό δεδομένου ότι κανένας δε θα ήθελε να δείχνει ότι επισκέπτεται τέτοιου είδους σελίδες, πόσο μάλλον να τοποθετήσει συνδέσμους προς αυτές στην προσωπική του ιστοσελίδα. Έτσι, συγκρίνοντας το βαθμό PageRank με τα στατιστικά στοιχεία χρήσης του Διαδικτύου, μπορούμε να εξαγάγουμε συμπεράσματα σχετικά με ιστοτόπους που οι άνθρωποι βρίσκουν ενδιαφέροντες αλλά αποφεύγουν να αναφέρουν στις προσωπικές σελίδες τους, σε μέσα κοινωνικής δικτύωσης, φόρουμ ή blogs.

Υπάρχουν κι άλλες πιθανές χρήσεις του βαθμού PageRank, όπως η εκτίμηση της ποιότητας των σελίδων που ο χρήστης έχει στη διάθεσή του προς πλοήγηση στην αναζήτηση μίας πληροφορίας, ο συστηματικός έλεγχος του ανταγωνισμού που μπορεί μία επιχείρηση ή ένας

οργανισμός να πραγματοποιεί επιβλέποντας τους σημαντικούς εισερχόμενους συνδέσμους που οι ανταγωνιστικοί ιστότοποι λαμβάνουν κατά καιρούς.

Η σημαντικότερη, όμως, εφαρμογή του αλγορίθμου PageRank είναι αυτή για την οποία επινοήθηκε, δηλαδή η «τακτοποίηση» του Διαδικτυακού Γράφου, και η οποία τον κατέστησε μία σημαντική παράμετρο για την εκτίμηση της κατάταξης που θα έπρεπε να έχει μία ιστοσελίδα στα αποτελέσματα των μηχανών αναζήτησης, για σχετικά με αυτήν ερωτήματα.

5.2.8 PageRank και μηχανές αναζήτησης

Ο αλγόριθμος PageRank ξεκίνησε ως ακαδημαϊκή έρευνα κι αποτέλεσε μία καινοτομία στις τεχνολογίες ανίχνευσης εγγράφων και κατάταξης των αποτελεσμάτων αναζήτησης, ενώ εξ αρχής συνδέθηκε άρρηκτα με τη μηχανή αναζήτησης της Google. Οι υπόλοιπες μηχανές προσπάθησαν να αντιγράψουν το μηχανισμό αυτό που έδειχνε να δίνει σημαντικό προβάδισμα στην Google, τουλάχιστον ως προς την ποιότητα των παρεχόμενων αποτελεσμάτων αναζήτησης.

Έκτοτε, οι σπουδαιότερες έως και σήμερα μηχανές αναζήτησης προσάρμοσαν τους αλγορίθμους τους, ενσωματώνοντας μία παράμετρο αξιολόγησης της ποιότητας των σελίδων. Δεν υπάρχουν επίσημα στοιχεία, στο Διαδίκτυο ή τη βιβλιογραφία, για τέτοιους παρόμοιους αλγορίθμους, όπως δεν υπάρχουν γενικότερα πληροφορίες για τους αλγορίθμους κατάταξης που χρησιμοποιούν οι εταιρείες, όμως μπορούμε με βεβαιότητα να υποθέσουμε την ύπαρξή τους στη λειτουργία των πλέον ανταγωνιστικών μηχανών αναζήτησης. Είναι δε προφανές ότι δε θα μπορούσε να γίνει η χρήση του ακριβούς μηχανισμού που η Google χρησιμοποιεί, καθώς η τελευταία δεν έχει αποκαλύψει πληροφορίες σχετικά με την προσαρμογή του βαθμού PageRank στις σύγχρονες απαιτήσεις του Διαδικτύου, ή τη βαρύτητα της παραμέτρου στο γενικότερο αλγόριθμο κατάταξης των αποτελεσμάτων.

Συγκεκριμένα, τόσο η Yahoo όσο και η Bing παρέχει εργαλείο εντοπισμού των εισερχόμενων συνδέσμων (εντός του Yahoo Site Explorer και του Bing Webmaster, αντίστοιχα).

5.2.9 Toolbar PageRank

Πολλές γραμμές (μπάρες) εργαλειών, όπως κι αυτή της Google (Google Toolbar), δίνουν τη δυνατότητα στο χρήστη να παρακολουθεί το βαθμό PageRank κάθε ιστοσελίδας. Ο βαθμός που αναγράφεται στις συγκεκριμένες γραμμές εργαλειών ονομάζεται Toolbar PageRank (βαθμός PageRank γραμμής εργαλειών).

Παράλληλα, ο βαθμός PageRank όπως αναλύθηκε στο παρόν κεφάλαιο ονομάζεται Actual ή Internal PageRank (Πραγματικός ή Εσωτερικός βαθμός PageRank). Η πλειοψηφία των χρηστών του Διαδικτύου δεν αντιλαμβάνεται ότι οι τιμές PageRank που φαίνονται στις διάφορες μπάρες εργαλειών δεν είναι οι πραγματικές τιμές PageRank που ο αλγόριθμος της Google χρησιμοποιεί για την κατάταξη των ιστοσελίδων (Sisson, 2006).

Συγκεκριμένα, η γραμμή εργαλειών διαθέτει μία μπάρα που διαιρείται σε 10 ίσα και γραμμικά μεταξύ τους διαστήματα, με τιμές από το 0 έως το 10, σε αντίθεση με την πραγματική πιθανοτική τιμή που κυμαίνεται από το 0 έως το 1. Οι γραμμικές αυτές διαιρέσεις αντιστοιχούν σε μία λογαριθμική κλίμακα, η βάση της οποίας υπολογίζεται κάπου μεταξύ του 5 και του 10 και η συσχέτιση με την οποία φαίνεται στον παρακάτω πίνακα:

Βαθμός PageRank γραμμής εργαλειών (**γραμμικός**)	Πραγματικός βαθμός PageRank (**$\log_5$**)	Πραγματικός βαθμός PageRank (**$\log_{10}$**)
0 - 1	0.15 - 1	0.15 - 1
1 - 2	1 - 5	1 - 10
2 - 3	5 - 25	10 - 100
3 - 4	25 - 125	100 - 1,000
4 - 5	125 - 625	1,000 - 10,000
5 - 6	625 - 3,125	10,000 - 100,000
6 - 7	3,125 - 15,625	100,000 - 1,000,000
7 - 8	15,625 - 78,125	1,000,000 - 10,000,000
8 - 9	78,125 - 390,625	10,000,000 - 100,000,000
9 - 10	> 390,625	> 100,000,000

Πίνακας 10 Συσχέτιση Toolbar PageRank και πραγματικού βαθμού PageRank

Οι τιμές του παραπάνω πίνακα υποδεικνύουν πόσο πιο δύσκολη είναι, για παράδειγμα, η μετάβαση από την τιμή 6 του Toolbar PageRank στην τιμή 7, εν συγκρίσει με τη μετάβαση από την τιμή 1 στην τιμή 2, και δεν αντιστοιχούν στον πραγματικό βαθμό PageRank, καθώς, όπως αναλύθηκε προηγουμένως, το άθροισμα των βαθμών PageRank όλων των κόμβων του διαδικτυακού γράφου(δηλαδή των ιστοσελίδων) ισούται με τη μονάδα.

Τον Αύγουστο του 2011, οι μηχανικοί της SEOMoz δημοσίευσαν ότι εκτιμούν, μετά από τη διεξαγωγή πολλών μετρήσεων και την κατασκευή ενός παρόμοιου αλγορίθμου, του *MOZRank*, ότι η λογαριθμική κλίμακα έχει μία δεκαδική βάση μεταξύ του 8 και του 9.

Επίσης, ο βαθμός PageRank της γραμμής εργαλειών ανανεώνεται κάθε ένα ορισμένο και μεγάλο χρονικό διάστημα σε αντίθεση με τον πραγματικό βαθμό που ανανεώνεται αυτόματα αρκετές φορές μέσα σε μία ημέρα, ανάλογα με την κατάσταση «ισορροπίας» του διαδικτυακού γράφου. Συγκεκριμένα και παρότι η Google είχε αρχικά ανακοινώσει ότι ο

βαθμός PageRank που είναι διαθέσιμος προς προβολή στις μπάρες εργαλειών ανανεώνεται κάθε μήνα, υπάρχει αισθητή ασυνέπεια καθώς τα διαστήματα των ανανεώσεων ποικίλλουν από 1 έως και 11 μήνες, μέχρι σήμερα.

Εξαιτίας της μεγάλης απόστασης μεταξύ της γραμμικής και της λογαριθμικής κλίμακας του βαθμού PageRank, ο βαθμός που αναγράφεται στις γραμμές εργαλειών δεν αποτελεί κάποια πραγματική ένδειξη του τρόπου με τον οποίο οι μηχανές αντιμετωπίζουν μία ιστοσελίδα, ενώ μπορεί μόνο να χρησιμοποιηθεί ως κριτήριο σύγκρισης μεταξύ των βαθμών PageRank της ιστοσελίδας, στον άξονα του χρόνου, καθώς και μεταξύ των βαθμών PageRank διαφορετικών (ανταγωνιστικών) ιστοσελίδων, παρότι και σε αυτές τις περιπτώσεις ακόμη και μία ίδια τιμή του Toolbar PageRank, μετά από μία ανανέωση, μπορεί να κρύβει μεγάλες αποκλίσεις.

Τέλος, ξαφνικές πτώσεις στο βαθμό του Toolbar PageRank συνήθως σημαίνουν ποινικοποίηση της ιστοσελίδας για κάποιο λόγο, όπως, για παράδειγμα, για την αγοραπωλησία εισερχόμενων κι εξερχόμενων συνδέσμων που ενδεχομένως οι μηχανικοί της Google εντόπισαν, ή επειδή εντοπίσθηκαν σύνδεσμοι προς την ιστοσελίδα από φάρμες συνδέσμων που προσφάτως ποινικοποιήθηκαν.

5.2.10 Εξέλιξη του βαθμού PageRank

5.2.10.1 Οι βασικές αδυναμίες

Όπως εξηγήθηκε παραπάνω, ο βαθμός PageRank επινοήθηκε για την εκτίμηση της δημοτικότητας μίας ιστοσελίδας, που, με τη σειρά της, είναι εκτίμηση της ποσότητας και της ποιότητας των συνδέσμων που εισέρχονται σε αυτήν.

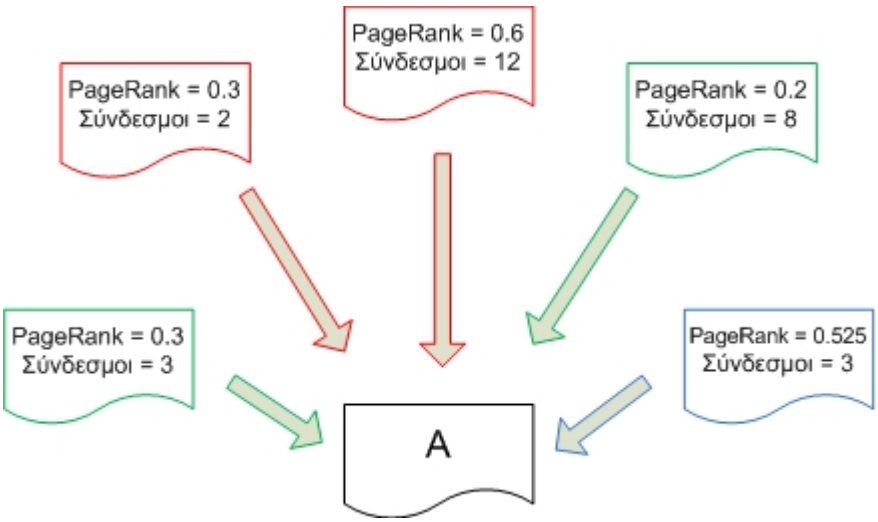
Κατά την ανάλυση του αλγορίθμου, έγινε εμφανές ότι η εκτίμηση της ποιότητας είναι στην ουσία «πλασματική», ενώ, στην ουσία, υπολογίζεται μονάχα ποσότητα. Για την ακρίβεια, υπολογίζεται ο αριθμός όχι των ακμών που δείχνουν προς μία σελίδα, αλλά των διαφορετικών «μονοπατιών» που οδηγούν στη σελίδα αυτή, και ο αναδρομικός αλγόριθμος τελικά αποδίδει την πιθανότητα να επιλεχθεί από έναν τυχαίο χρήστη κάποιο μονοπάτι που οδηγεί στη συγκεκριμένη σελίδα. Άλλωστε, αυτό που καταλήγει να εκλαμβάνεται ως συνάρτηση ποσότητας και ποιότητας πηγάζει, στην πρώτη επανάληψη, αποκλειστικά ως ποσότητα, με ανάθεση μίας αρχικής τιμής στους βαθμούς όλων των ιστοσελίδων (αρχικό διάνυσμα).

Παράλληλα, ο βαθμός PageRank αφορά τη σελίδα καθαυτή και είναι ένα απόλυτο μέτρο. Αυτό σημαίνει ότι αποτελεί έναν παράγοντα κατάταξης των αποτελεσμάτων αναζήτησης ο οποίος είναι ανεξάρτητος από το ερώτημα – όρο αναζήτησης (query – independent factor).

Έτσι, ο βαθμός PageRank, όπως διατυπώθηκε το 1998 από τον Larry Page, κατατάσσει τις σελίδες του διαδικτυακού γράφου και όχι τα αποτελέσματα των διαφορετικών αναζητήσεων, επομένως η συμβολή του στον γενικό αλγόριθμο κατάταξης των αποτελεσμάτων είναι συγκεκριμένη για κάθε σελίδα, ανεξαρτήτως του όρου αναζήτησης. Παράλληλα, αδυνατεί από μόνος του να διακρίνει μεταξύ σελίδων αυθεντιών (authoritative pages) γενικού ενδιαφέροντος και αυθεντιών πάνω σε συγκεκριμένη θεματολογία.

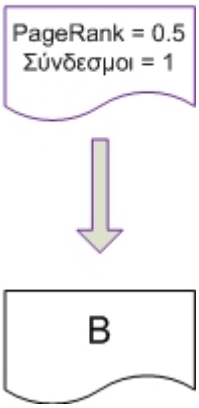
Έστω, λοιπόν, δύο σελίδες, A και B, με βαθμό PageRank $PageRank(A) = PageRank(B) = 0.575$, αλλά με αισθητές διαφορές στη δομή με την οποία συνδέονται με τον υπόλοιπο γράφο, όπως φαίνεται παρακάτω. Η ροή του βαθμού PageRank έχει υπολογιστεί από τον τύπο $PageRank(A) = (1-d) + d\sum_{i=1}^{n} \frac{PageRank(T_i)}{C(T_i)}$ με συντελεστή $d = 0.85$, ενώ σε κάθε σελίδα αναγράφονται ο βαθμός PageRank και οι εξερχόμενοι από αυτήν Σύνδεσμοι.

Έτσι, για την σελίδα A έχουμε:



$$PageRank(A) = (1-0.85) + 0.85\left(\frac{0.3}{3} + \frac{0.3}{2} + \frac{0.6}{12} + \frac{0.2}{8} + \frac{0.525}{3}\right) = 0.575$$

ενώ, για την σελίδα B:



$$PageRank(B) = (1-0.85) + 0.85\left(\frac{0.5}{1}\right) = 0.575$$

Κατ' επέκτασιν των παραπάνω, η αρχική διατύπωση του PageRank αντιμετωπίζει τις δύο σελίδες εξίσου και αποδίδει την ίδια ακριβώς συμβολή στο γενικό αλγόριθμο κατάταξης των αποτελεσμάτων. Όμως, μία φαινομενικά ποιοτικότερη σελίδα με βαθμό 0.6 συμβάλλει λιγότερο, λόγω αριθμού εξερχόμενων συνδέσμων, στο βαθμό της σελίδας Α από μία λιγότερο ποιοτική σελίδα με βαθμό 0.3 (κόκκινο ζεύγος), σε αντίθεση με το ζεύγος των σελίδων με PageRank 0.3 και 0.2, 3 και 8 εξωτερικούς συνδέσμους, αντίστοιχα (πράσινο ζεύγος), ενώ δε μπορεί να γνωρίζει εάν η σελίδα με βαθμό 0.525 αποτελεί διανομέα πληροφοριών ή αυθεντία, με ειδίκευση σε συγκεκριμένη θεματολογία, σχετική με τον εκάστοτε όρο αναζήτησης ή με το θέμα που πραγματεύεται η σελίδα Α (μπλε σελίδα). Παράλληλα, ο αλγόριθμος αδυνατεί να κρίνει εάν ο σύνδεσμος από την ποιοτική σελίδα βαθμού 0.5 προς την ιστοσελίδα Β είναι προϊόν χρηματικής συναλλαγής, άρα εάν υπάρχει θεματική σχέση μεταξύ των δύο σελίδων.

Κρίθηκε, λοιπόν, απαραίτητο από τις μηχανές αναζήτησης να διαχωρίσουν την ποιότητα από τη δημοτικότητα, και να εκτιμήσουν την πρώτη προσεγγίζοντας το βαθμό εξειδίκευσης των σελίδων σε θεματολογία, σε μία προσπάθεια εξέλιξης του αρχικού αλγορίθμου που εισήγαγε η Google.

5.2.10.2 Ο αλγόριθμος Hilltop

Μία ιδέα που εστίασε στην επίλυση των παραπάνω προβλημάτων είναι ο αλγόριθμος Hilltop των Bharat και Mihaila (1999).

Σύμφωνα με τους συγγραφείς της σχετικής δημοσίευσης, η προσέγγιση του αλγορίθμου PageRank περιλαμβάνει την ανάλυση των υπερσυνδέσμων μεταξύ σελίδων στο διαδίκτυο, με την υπόθεση ότι σελίδες πάνω σε θεματολογία συνδέουν η μία προς την άλλη και σελίδες αυθεντίας τείνουν να δείχνουν σε άλλες σελίδες αυθεντίας. Δηλαδή, όπως εξηγήθηκε και παραπάνω, υπολογίζει έναν βαθμό για κάθε σελίδα ανεξάρτητο από ερωτήματα αναζήτησης και τον χρησιμοποιεί για να κατατάξει το σύνολο των αποτελεσμάτων συγκεκριμένων αναζητήσεων. Με τον τρόπο αυτό, ο αλγόριθμος αδυνατεί να διαχωρίσει τις σελίδες – αυθεντίες σε κάποιο συγκεκριμένο θέμα από τις σελίδες που αποτελούν αυθεντίες γενικότερα. Συγκεκριμένα, οι τελευταίες ενδέχεται να ταιριάζουν τέλεια με κάποιο ερώτημα αναζήτησης αλλά δεν αποτελούν αυθεντίες στο συγκεκριμένο θέμα που πραγματεύεται το ερώτημα αναζήτησης και τέτοιες σελίδες δε θα έπρεπε να κατηγοριοποιούνται και να αντιμετωπίζονται ως τέτοιες.

Οι έννοιες και ιδιότητες των αυθεντιών και διανομέων πληροφοριών (ή συνδέσμων) περιγράφονται από τους Bharat και Mihaila, ενώ η κεντρική ιδέα της διάκρισης βασίζεται στην πιθανότητα ότι σελίδες που ανήκουν σε μία κοινότητα θα συνδέουν στους πλέον σχετικούς πόρους, ενώ άλλες σελίδες της κοινότητας θα αποτελούν αυθεντίες σε συγκεκριμένα θέματα, καθ' υπόδειξη της δημοσιότητάς τους σε όρους συνδέσμων εντός της κοινότητας αυτής. Έτσι, διακρίνονται δύο νέα είδη ιστοσελίδων, τα οποία αναλύονται παρακάτω και τα οποία συναποτελούν ένα αδιάσπαστο δίπολο, καθότι δε νοείται η ύπαρξη του διανομέα εξειδικευμένων συνδέσμων χωρίς την ύπαρξη αυθεντιών στις οποίες να συνδέει και, αντίστροφα, δε νοείται η ύπαρξη μίας αυθεντίας χωρίς την ύπαρξη ενός διανομέα εξειδικευμένων πληροφοριών που να συνδέει προς αυτήν.

Σημειώνεται εδώ ότι μία σελίδα ενδέχεται να προκριθεί τόσο ως διανομέας εξειδικευμένων συνδέσμων όσο και ως αυθεντία, όπως, για παράδειγμα, η σελίδα της Wikipedia σχετικά με μία πόλη, καθώς περιλαμβάνει εξειδικευμένες πληροφορίες πάνω στην πόλη αυτή αλλά και εξερχόμενους συνδέσμους προς σελίδες που είναι εξίσου σχετικές με την πόλη ή από τις οποίες αντλεί το περιεχόμενό της.

5.2.10.3 Διανομείς εξειδικευμένων πληροφοριών

Ως διανομέας εξειδικευμένων πληροφοριών (hub), ή ειδική σελίδα (expert page) όπως αναφέρεται στη δημοσίευση για τον αλγόριθμο Hilltop, ονομάζεται μία σελίδα που πραγματεύεται ένα ορισμένο θέμα και περιλαμβάνει εξερχόμενους συνδέσμους, οι οποίοι δείχνουν σε άσχετες μεταξύ τους σελίδες πάνω στο θέμα αυτό. Κατά τη σχετική έρευνα του 1999, μόλις 2,5 εκατομμύρια σελίδων του Διαδικτύου αναγνωρίσθηκαν ως ειδικές σελίδες.

Παράδειγμα τέτοιων σελίδων μπορεί να είναι η σελίδα ενός διαδικτυακού καταλόγου, π.χ. κάποια συγκεκριμένη κατηγορία του Open Directory Project (DMOZ).

Ορισμένες τεχνικές βελτιστοποίησης των ιστοσελίδων που σχετίζονται με τις σελίδες διανομής εξειδικευμένων πληροφοριών είναι οι εξής:

α) Συνίσταται η κατασκευή σελίδων που μπορούν να χαρακτηρισθούν ως ειδικές, καθώς αυτές αποκτούν ιδιαίτερη βαρύτητα στις μηχανές αναζήτησης και αποδίδουν ιδιαίτερη βαρύτητα στις σελίδες στις οποίες δείχνει, βάσει του διπόλου. Τέτοιες ιστοσελίδες εντός του ιστοτόπου μπορεί να είναι σελίδες που περιλαμβάνουν «σχετικούς συνδέσμους» ή ένας χάρτης ιστοτόπου HTML (και όχι XML για τις μηχανές αναζήτησης), δηλαδή ένας χάρτης που περιλαμβάνει σελίδες του ιστοτόπου και είναι διαθέσιμος προς προβολή από τους χρήστες. Δηλαδή, η ιδέα μπορεί να εφαρμοστεί και σε επίπεδο σελίδων εντός του ίδιου ιστοτόπου, συντάσσοντας διαφορετικές ειδικές σελίδες που να δείχνουν στα διαφορετικά σύνολα των σχετικών σελίδων (για παράδειγμα, σε έναν ηλεκτρονικό κατάλογο, η κατηγορία «Υγεία» αποτελεί ειδική στο θέμα της υγείας σελίδα που συνδέει προς αυθεντίες – υποκατηγορίες, όπως είναι οι «φαρμακεία», «ομοιοπαθητικοί», «καρδιοχειρούργοι», κ.ο.κ.).

β) Σελίδες που περιλαμβάνουν μία συλλογή συνδέσμων σε «σχετικές σελίδες» θα είναι πιο αποδοτικές, σε όρους PageRank, όσο πιο σχετικές μεταξύ τους είναι οι σελίδες. Από αυτή την οπτική γωνία, είναι πιο συμφέρουσα η κατασκευή διαφορετικών χαρτών ιστοτόπων ή σελίδων με «σχετικούς συνδέσμους», με κριτήριο το αντικείμενο των σελίδων στις οποίες συνδέουν.

γ) Για την κατασκευή μίας ειδικής σελίδας, είναι ιδιαίτερα χρήσιμη η εφαρμογή των τεχνικών βελτιστοποίησης εντός της ιστοσελίδας, όπως αυτή αναλύθηκε στο προηγούμενο κεφάλαιο. Η αξιοποίηση, για παράδειγμα, των τίτλων της σελίδας, των επικεφαλίδων, των μέτα δεδομένων, της μορφοποίησης των λέξεων και, κυρίως, του anchor text των ειδικών συνδέσμων μπορεί να συμβάλλει ιδιαίτερα στον σκοπό αυτό.

δ) Τέλος, είναι προφανές ότι, κατά την εφαρμογή τεχνικών κατασκευής εισερχόμενων συνδέσμων, όπως αυτές θα αναλυθούν παρακάτω, ενδέχεται να ωφελήσει ιδιαίτερα τον ιστότοπο να αποκτήσει συνδέσμους από expert pages, καθώς αυτοί τον καθιστούν αυτόματα αυθεντία.

5.2.10.4 Αυθεντίες

Ως αυθεντία (authority), ή στόχος (target) όπως αναφέρεται στην αρχική δημοσίευση, ορίζεται μία σελίδα που πραγματεύεται ένα ορισμένο θέμα και περιλαμβάνει πολλαπλούς εισερχόμενους συνδέσμους σχετικούς το θέμα αυτό.

Παραδείγματα τέτοιων σελίδων που αποτελούν αυθεντία σε ένα θέμα είναι οι ακαδημαϊκές δημοσιεύσεις που χρησιμοποιούνται ως βιβλιογραφία στην παρούσα διπλωματική εργασία, η σελίδα προγνώσεων της Εθνικής Μετεωρολογικής Υπηρεσίας και οι διάφορες σελίδες της Wikipedia.

Ορισμένες τεχνικές βελτιστοποίησης των ιστοσελίδων που σχετίζονται με τις σελίδες – αυθεντίες είναι οι εξής:

α) Η δημιουργία ποιοτικού, αυθεντικού και χρήσιμου περιεχομένου θα προσελκύσει συνδέσμους με τον φυσικό τρόπο, στον οποίο βασίζεται η φιλοσοφία του βαθμού PageRank, με αποτέλεσμα, σταδιακά, κάποια σελίδα ενός ιστοτόπου να προκριθεί ως αυθεντία.

β) Πέρα από την αναζήτηση συνδέσμων από ειδικές σελίδες, όπως αναφέρθηκε παραπάνω, για τον πρόκριση μίας ιστοσελίδας ως αυθεντίας, είναι παράλληλα χρήσιμη και η απόκτηση συνδέσμων από αυθεντίες, καθώς αυτές είναι αρκετά πιθανό να συσχετίσουν, στα αποτελέσματα αναζήτησης, τη σελίδα άμεσα με το θέμα που πραγματεύονται.

5.2.10.5 Το νοηματικό πλαίσιο των υπερσυνδέσμων

Σύμφωνα με τη δημοσίευση των Bharat και Mihaila (1999), πέραν της καθιέρωσης των όρων «ειδική σελίδα» και «στόχος», δίνεται μία νέα κατεύθυνση στη φιλοσοφία και τον τρόπο με τον οποίο μεταφράζονται οι σύνδεσμοι που βρίσκονται σε μία σελίδα. Έτσι, για κάθε ειδική σελίδα E που δείχνει σε έναν στόχο T, σχεδιάζουμε μία κατευθυνόμενη ακμή (E,T), και ορίζουμε την ακόλουθη σχέση μεταξύ φράσεων – κλειδιών και ακμών, δεδομένων του τίτλου της σελίδας (page title), των επικεφαλίδων (headings) και του anchor text, όπως αυτά αναλύθηκαν στο προηγούμενο κεφάλαιο:

- Ο τίτλος της σελίδας χαρακτηρίζει το σύνολο των ακμών που εξέρχονται από την ειδική σελίδα.
- Μία επικεφαλίδα ορισμένης σπουδαιότητας (h1, h2, h3) χαρακτηρίζει τις ακμές εκείνες των οποίων οι αντίστοιχοι υπερσύνδεσμοι εμφανίζονται, εντός του εγγράφου, μετά από την επικεφαλίδα αυτή και πριν την επόμενη επικεφαλίδα ίσης ή μεγαλύτερης σπουδαιότητας.
- Το anchor text του υπερσυνδέσμου χαρακτηρίζει την ακμή που αντιστοιχεί στον υπερσύνδεσμο αυτό.

```
<head>
<title>Εθνικό Μετσόβιο Πολυτεχνείο - Εργαστήρια</title>
</head>

<body>
<h1>Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών</h1>

<h2>Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων</h2>

<a href="http://imu.ntua.gr/">Μονάδα Διοίκησης Πληροφοριακών Συστημάτων</a>
<a href="http://academics.epu.ntua.gr/">Συστήματα Αποφάσεων και Διοίκησης</a>
<a href="http://www.fsu.gr">Μονάδα Προβλέψεων και Στρατηγικής</a>

<h2>Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής</h2>
..

<h2>Τομέας Ηλεκτρικής Ισχύος</h2>

<a href="http://www.e3mlab.ntua.gr/"> Οικονομίας-Ενέργειας-Περιβάλλοντος</a>
<a href="http://lighting.ece.ntua.gr/">Εργαστήριο Φωτοτεχνίας</a>
<a href="highvoltages.ece.ntua.gr">Εργαστήριο Υψηλών Τάσεων</a>

<h1>Σχολή Μηχανολόγων Μηχανικών</h1>
..

</body>
```

Στο παραπάνω παράδειγμα, για δεδομένη ειδική σελίδα που πραγματεύεται τα εργαστήρια των σχολών του Εθνικού Μετσόβιου Πολυτεχνείου, ο τίτλος της σελίδας χαρακτηρίζει το σύνολο των συνδέσμων που υπάρχουν στη σελίδα αυτή, ενώ οι επικεφαλίδες <h1> χαρακτηρίζουν τις ακμές που ξεκινούν μετά από αυτές, μέχρι την επόμενη επικεφαλίδα ίσης ή μεγαλύτερης σημασίας (h1), δηλαδή όλοι οι εμφανιζόμενοι σύνδεσμοι αφορούν τη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών. Στη συνέχεια, οι επικεφαλίδες <h2> χαρακτηρίζουν τις ακμές που ξεκινούν μετά από αυτές, μέχρι την επόμενη επικεφαλίδα ίσης ή σπουδαιότερης σημασίας (h1 ή h2), όπως για παράδειγμα οι σύνδεσμοι του Τομέα Ηλεκτρικής Ισχύος που πλαισιώνονται από την αντίστοιχη <h2> επικεφαλίδα και την <h1> επικεφαλίδα της Σχολής Μηχανολόγων Μηχανικών. Τελευταίο σε προτεραιότητα, ακολουθεί το anchor text κάθε ενός υπερσυνδέσμου.

Έτσι, η σελίδα http://imu.ntua.gr σχετίζεται με το εργαστήριο της Μονάδας Διοίκησης Πληροφοριακών Συστημάτων, του Τομέα Ηλεκτρικών Βιομηχανικών Διατάξεων & Συστημάτων Αποφάσεων, της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου.

Η παραπάνω σχέση μεταξύ των στοιχείων της σελίδας και του υπερσυνδέσμου αφορά το δίπολο «ειδικής σελίδας» - «στόχου». Οι μηχανές αναζήτησης, όμως, λαμβάνουν υπόψη το

ευρύτερο νοηματικό πλαίσιο των συνδέσμων, ανεξάρτητα από τον τύπο των σελίδων που συνδέονται, το οποίο μπορεί να αναφέρεται και ως πλαίσιο συνδέσμων.

Έτσι, στοιχεία όπως το anchor text, οι επικεφαλίδες, ο τίτλος της σελίδας που πλαισιώνουν έναν σύνδεσμο τον χαρακτηρίζουν και αποδίδουν μεγαλύτερη βαρύτητα στη σελίδα στην οποία δείχνουν, σε περίπτωση αναζήτησης λέξεων που σχετίζονται με ένα από αυτά τα στοιχεία. Υπενθυμίζεται, στο σημείο αυτό, πως παράγοντας βελτιστοποίησης των συνδέσμων αποτελεί και ο τίτλος του συνδέσμου, που αποδίδεται με την προσθήκη του γνωρίσματος **title=" "** στην ετικέτα του συνδέσμου <a href>.

Ένας παρόμοιος με τον PageRank και σχετικός με τον Hilltop αλγόριθμος είναι ο αλγόριθμος HITS (Hypertext – Induced Topic Search) του καθηγητή Jon Kleinberg του Cornell University, ο οποίος αναλύεται εκτενώς σε σχετική με τον PageRank δημοσίευση των Langville και Meyer, παράλληλα με τη σχέση και τη συνεισφορά του στο σύστημα κατανομής βαθμών PageRank (Langville, Meyer & FernÁndez, 2006).Η βασική ιδέα του αλγορίθμου HITS είναι ότι, κατά την αξιολόγηση της σημασίας μίας ιστοσελίδας, δεν υπολογίζονται αποκλειστικά και μόνο οι σελίδες που δείχνουν προς αυτή την ιστοσελίδα αλλά και οι σελίδες στις οποίες δείχνει η ίδια η ιστοσελίδα, καθώς «οι ποιοτικοί διανομείς πληροφοριών (hubs) συνδέουν σε ποιοτικές αυθεντίες (authorities) αλλά και οι ποιοτικές αυθεντίες συνδέονται από ποιοτικούς διανομείς πληροφοριών» (Shi, 2007).

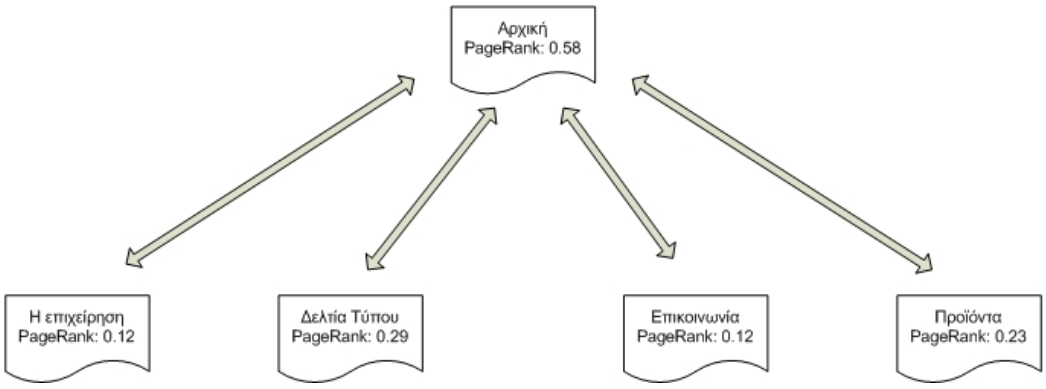
5.2.11 PageRank κι εσωτερική δομή ενός ιστοτόπου

Σύμφωνα και με την αρχική διατύπωση, η ροή του βαθμού PageRank πραγματοποιείται κατά τη σύνδεση μεταξύ σελίδων, μέσω ενός ή περισσότερων εξερχόμενων συνδέσμων, είτε πρόκειται για εσωτερικούς (δείχνουν σε σελίδες εντός του ιστοτόπου) είτε για εξωτερικούς (δείχνουν σε σελίδες διαφορετικών ιστοτόπων). Από αυτή την πρόταση εξάγονται δύο πολύ βασικά και χρήσιμα, από πλευράς τεχνικής βελτιστοποίησης ενός ιστοτόπου, συμπεράσματα:

- Ο βαθμός PageRank ποικίλλει ανά τις σελίδες του ίδιου ιστοτόπου, βάσει της εσωτερικής δομής του.
- Ο βαθμός PageRank των σελίδων ενός ιστοτόπου μπορεί να χειραγωγηθεί, επιλέγοντας την κατάλληλη δομή.

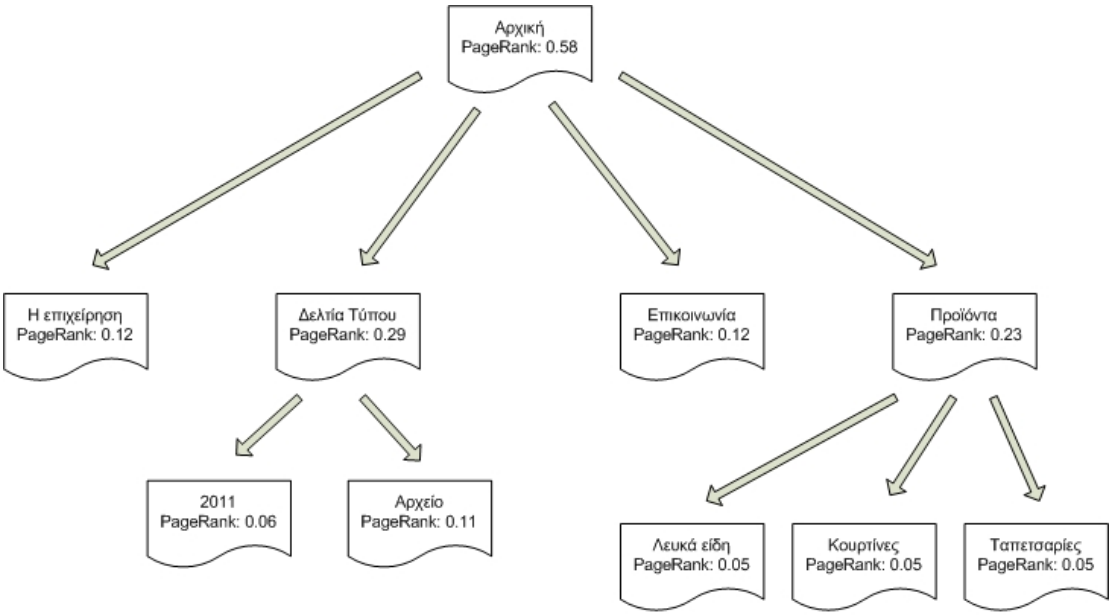
Έτσι, παρατηρείται ότι, παρόλο που η φυσική απόκτηση εισερχόμενων συνδέσμων πραγματοποιείται συνήθως για συγκεκριμένες σελίδες ενός ιστοχώρου (όπως, για παράδειγμα, επιστημονικά άρθρα, ειδησεογραφικές δημοσιεύσεις, εργαλεία διαδικτυακών πυλών), η αρχική σελίδα παρουσιάζει το μεγαλύτερο βαθμό PageRank σε ολόκληρο τον ιστότοπο. Αυτό συμβαίνει γιατί η κατασκευή (και όχι φυσική απόκτηση) εισερχόμενων

συνδέσμων συνήθως αφορά διαδικτυακούς καταλόγους που συνδέουν προς την αρχική σελίδα, από προεπιλογή και προτίμηση του ίδιου του διαχειριστή, αλλά και για δύο κυρίως λόγους που σχετίζονται με την εσωτερική δομή, διότι (α) όλες οι σελίδες ενός ιστοχώρου παρέχουν τουλάχιστον έναν σύνδεσμο προς την αρχική σελίδα (από το κεντρικό μενού και τον υπερσύνδεσμο του λογότυπου) και (β) ο βαθμός της αρχικής σελίδας διαιρείται στις σελίδες των συνδέσμων του κεντρικού μενού, από τις οποίες οι διαιρεμένοι βαθμοί διαιρούνται με τη σειρά τους προς τα κατώτερα στάδια του ιστοχώρου.



Εικόνα 30 Όλες οι σελίδες ενός ιστοχώρου παρέχουν σύνδεσμο προς την αρχική

Απόρροια του τελευταίου είναι ότι καθώς διανύεται κατά πλάτος ο γράφος ενός ιστοτόπου, από πάνω προς τα κάτω, ο βαθμός PageRank σημειώνει σταδιακή μείωση.



Εικόνα 31 Ο PageRank της αρχικής διαιρείται στις σελίδες του κεντρικού μενού

5.2.12 Αξιώματα του βαθμού PageRank

Δεδομένης της αρχικής διατύπωσης και της θεωρητικής και μαθηματικής ανάλυσης του όρου PageRank, προκύπτουν εύκολα τρία ιδιαίτερα αξιοποιήσιμα αξιώματα, γύρω από τα οποία αναπτύσσεται η φιλοσοφία της τεχνικής βελτιστοποίησης του βαθμού PageRank, μέσω της κατασκευής εισερχόμενων συνδέσμων.

5.2.12.1 Πρώτο αξίωμα: Ο αριθμός των εισερχόμενων συνδέσμων

Σύμφωνα με τον μαθηματικό τύπο υπολογισμού του βαθμού PageRank μίας σελίδας Α, $PageRank(A) = (1-d) + d\sum_{i=1}^{n} \frac{PageRank(T_i)}{C(T_i)}$, καθώς αυξάνονται οι εισερχόμενοι σύνδεσμοι στην ιστοσελίδα Α, δηλαδή οι σελίδες που παρέχουν εξερχόμενο σύνδεσμο προς την σελίδα Α, ο αριθμός *n* των συνδέσμων και, κατ' επέκτασιν, ο όρος $\sum_{i=1}^{n} \frac{PageRank(T_i)}{C(T_i)}$ αυξάνονται, βελτιώνοντας το βαθμό PageRank, ανεξαρτήτως των όρων *C(Ti)* και *PageRank(Ti)*.

Επομένως, συνίσταται η προσέλκυση και απόκτηση ή τεχνητή κατασκευή όσο το δυνατόν περισσότερων συνδέσμων προς μία ιστοσελίδα.

Παράλληλα, προσοχή πρέπει να δοθεί στο γεγονός ότι η τιμή του βαθμού PageRank αποδίδεται σε επίπεδο σελίδας και όχι ιστοτόπου ή διακομιστή και ως τέτοιου πρέπει να αντιμετωπίζεται από το διαχειριστή. Αυτό, στην πράξη, σημαίνει ότι ο προορισμός των εισερχόμενων συνδέσμων που κατασκευάζονται δεν πρέπει να περιορίζεται στην αρχική σελίδα, αλλά να αφορά οποιαδήποτε ενδιαφέρουσα σελίδα του ιστοτόπου, της οποίας η κατάταξη επιθυμείται να βελτιστοποιηθεί.

Τέλος, σημειώνεται σε αυτό το σημείο ότι, όπως έχει ήδη αναφερθεί, είναι επιθυμητός ο μεγάλος αριθμός των backlinks, όχι ο ρυθμός με τον οποίον αυτοί αυξάνονται. Οι αλγόριθμοι των μηχανών αναζήτησης είναι ιδιαίτερα ευαίσθητοι σε αυτό το θέμα, καθώς η αύξηση των εισερχόμενων συνδέσμων με ασυνήθιστη ταχύτητα αποτελεί ένδειξη χειραγώγησης του βαθμού PageRank κι επιδιώκεται τέτοιες σελίδες να ποινικοποιούνται, επιφέροντας τα αντίθετα αποτελέσματα. Επίσης, δε συνίσταται η απόκτηση συνδέσμων από σελίδες ή ιστοχώρους αμφιβόλου ποιότητας, όπως, για παράδειγμα, σελίδες οι διαχειριστές των οποίων εφαρμόζουν τεχνικές black – hat βελτιστοποίησης, καθώς και οι φάρμες συνδέσμων (link farms).

5.2.12.2 Δεύτερο αξίωμα: Η ροή PageRank των εισερχόμενων συνδέσμων

Πάντα σύμφωνα με τον μαθηματικό τύπο υπολογισμού του βαθμού PageRank, όσο πιο υψηλός είναι ο βαθμός PageRank των *n* σελίδων, *PageRank(i)*, που συνδέουν προς την ιστοσελίδα Α (ceteris paribus), τόσο πιο μεγάλη είναι και η ροή του βαθμού (PageRank flow) προς την σελίδα Α.

Άλλωστε, όσο πιο μεγάλος είναι ο βαθμός PageRank μίας σελίδας, τόσα περισσότερα διαφορετικά μονοπάτια μπορεί δυνητικά να ακολουθήσει ένας τυχαίος χρήστης του Διαδικτύου, έτσι και μία διαφορετική ιστοσελίδα που αποκλειστικά δέχεται έναν σύνδεσμο από τη σελίδα αυτή, αυξάνει (σχεδόν) εξίσου τις πιθανότητές της να την επισκεφθεί ο τυχαίος χρήστης αυτός. Οι μηχανές αναζήτησης εκλαμβάνουν το σύνδεσμο από τη σελίδα Α προς τη

σελίδα Β ως μία ψήφο της σελίδας Α προς τη σελίδα Β, ενώ η ψήφος από μία σελίδα υψηλού βαθμού PageRank θεωρείται σημαντική ψήφος.

Έτσι, συνίσταται η απόκτηση ή κατασκευή εισερχόμενων συνδέσμων από ιστοσελίδες με όσο το δυνατόν υψηλότερο βαθμό PageRank.

Το παραπάνω έχει ιδιαίτερη βαρύτητα κατά τη σύγκριση μεταξύ δύο διαφορετικών επενδυτικών σχεδίων για τον ιστότοπο μίας ηλεκτρονικής επιχείρησης. Έτσι, ανάμεσα σε δύο διαφορετικές οικονομικές προσφορές για αγορά συνδέσμου σε διαφορετικές ιστοσελίδες, η επιχείρηση οφείλει να αξιολογήσει όχι μόνο το κόστος της επένδυσης (ανηγμένο στο ίδιο χρονικό διάστημα) αλλά και το βαθμό απόδοσής της, ελέγχοντας, ανά περιόδους, τον ενδεικτικό βαθμό PageRank της κάθε σελίδας, μαζί με την επισκεψιμότητα και τη θέση της στα οργανικά αποτελέσματα των μηχανών αναζήτησης για τους όρους αναζήτησης που την ενδιαφέρουν και που, ενδεχομένως, συνδέουν την κάθε σελίδα με τον ιστότοπο της επιχείρησης.

Παράλληλα, κατά την κατασκευή συνδέσμων από διαδικτυακούς καταλόγους ή άλλους ιστοχώρους σχετικού περιεχομένου (κατόπιν αιτήσεως ή συμφωνίας), συνίσταται η επιλογή μίας σελίδας του ιστοτόπου που βρίσκεται, κατά την κατά βάθος διάσχιση του γράφου αυτού, όσο το δυνατόν πλησιέστερα στην αρχική σελίδα. Με τον τρόπο αυτό διασφαλίζεται η όσο το δυνατόν μεγαλύτερη ροή βαθμού PageRank. Έτσι, κατά την καταχώρηση στο Open Directory Project (http://www.dmoz.org) ενός ελληνικού ιστοτόπου που πραγματεύεται ηλεκτρονικά παιχνίδια για τις κονσόλες, είναι προτιμότερη η αίτηση δημιουργίας συνδέσμου στην κατηγορία «Top > World > Greek > Παιχνίδια > Ηλεκτρονικά Παιχνίδια», παρά στην κατηγορία «Top > World > Greek > Παιχνίδια > Ηλεκτρονικά Παιχνίδια > Κονσόλες».

5.2.12.3 Τρίτο αξίωμα: Ο αριθμός εξερχόμενων συνδέσμων

Από τον τύπο υπολογισμού του PageRank, γίνεται εμφανές ότι όσο περισσότεροι είναι οι εξερχόμενοι σύνδεσμοι της σελίδας *i*, *C(Ti)*, που συνδέει προς τη σελίδα Α, τόσο μικρότερη είναι η επίδραση της συγκεκριμένης σελίδας στο βαθμό της ιστοσελίδας Α. Μάλιστα, σε περίπτωση που οι σύνδεσμοι αυτοί *C(Ti)* αυξηθούν, ceteris paribus, τότε ο βαθμός PageRank της σελίδας Α θα μειωθεί.

Θεωρητικά, ο βαθμός PageRank μίας σελίδας εκπροσωπεί το πλήθος των μονοπατιών που ένας τυχαίος χρήστης μπορεί δυνητικά να ακολουθήσει για να επισκεφθεί τη σελίδα αυτή. Έτσι, καθώς οι εξερχόμενοι σύνδεσμοι της σελίδας αυτής αυξάνονται, η πιθανότητα, στο επόμενο από τη σελίδα βήμα, ο χρήστης να ακολουθήσει έναν από τους συνδέσμους της μειώνεται, καθώς ισούται με $1/C(Ti)$.

Έτσι, όσον αφορά την κατασκευή εισερχόμενων συνδέσμων από ξένους ιστοχώρους, συνίσταται η απόκτηση συνδέσμων από ιστοσελίδες που δεν περιλαμβάνουν πολλούς εξερχόμενους συνδέσμους, ενώ ο ρυθμός με τον οποίον αυτοί αυξάνονται είναι σχετικά χαμηλός.

Παράλληλα με το προηγούμενο αξίωμα του βαθμού PageRank, κατά την αξιολόγηση δύο επενδυτικών σχεδίων μίας ηλεκτρονικής επιχείρησης, πέρα από το οικονομικό κριτήριο, την επισκεψιμότητα, την κατάταξη της κάθε σελίδας στα οργανικά αποτελέσματα των μηχανών αναζήτησης για τις επιθυμητές λέξεις ή φράσεις – κλειδιά και το βαθμό PageRank, η επιχείρηση οφείλει να αξιολογεί και τον παράγοντα των συνολικών εξερχόμενων συνδέσμων από την κάθε σελίδα, δίνοντας προτεραιότητα σε εκείνη τη σελίδα που περιέχει τους λιγότερους εξερχόμενους συνδέσμους και παρουσιάζει μικρό ρυθμό προσθήκης νέων.

Το συγκεκριμένο συμπέρασμα, όμως, για τον αριθμό των εξερχόμενων συνδέσμων μπορεί να αξιοποιηθεί ιδιαίτερα και κατά την κατασκευή της εσωτερικής δομής ενός ιστοτόπου, στην διαδικασία της οποίας είναι προτιμότερο οι σελίδες με υψηλό PageRank να περιλαμβάνουν περιορισμένο αριθμό συνδέσμων.

Τέλος, ιδιαίτερη προσοχή πρέπει να δοθεί και κατά την παροχή εξωτερικών εξερχόμενων συνδέσμων από έναν ιστότοπο, από την οπτική γωνία του ίδιου του ιστοτόπου. Η ροή PageRank, όπως έχει ήδη αναλυθεί, δε διακρίνει εάν ο εξερχόμενος σύνδεσμος είναι εσωτερικός ή εξωτερικός, καθώς και οι δύο τύποι των συνδέσμων μοιράζονται την ίδια πιθανότητα ακολούθησης από έναν τυχαίο χρήστη, άρα και το ίδιο μέρος του βαθμού PageRank. Έτσι, συνίσταται να αποφεύγονται περιττοί εξωτερικοί σύνδεσμοι στις σελίδες ενός ιστοτόπου, ενώ η παροχή συνδέσμων προς εξωτερικές σελίδες να πραγματοποιείται σε συγκεκριμένες σελίδες του ιστοτόπου αυτού (π.χ. σε μία σελίδα «σχετικές σελίδες» ή «χρήσιμοι σύνδεσμοι», κ.λπ.). Εάν δεν επιθυμείται η ροή PageRank στους εξωτερικούς συνδέσμους (ή διαρροή βαθμού PageRank, όπως ονομάζεται στην περίπτωση της μη

επιθυμητής απόδοσης PageRank), παρά μόνο η προβολή των σελίδων αυτών, μπορεί να χρησιμοποιηθεί η τιμή **rel = ”nofollow”**, όπως αναλύθηκε στο προηγούμενο κεφάλαιο, στην ετικέτα του υπερσυνδέσμου <a href>. Η ετικέτα αυτή μπορεί να χρησιμοποιηθεί για την αποφυγή ροής PageRank και σε εσωτερικές σελίδες, η κατάταξη των οποίων στα οργανικά αποτελέσματα αναζήτησης και η βελτιστοποίηση αυτής δεν επιθυμείται ή δεν αποτελεί προτεραιότητα των διαχειριστών του ιστοτόπου, όπως για παράδειγμα η σελίδα της επικοινωνίας, διάφορες φόρμες, σελίδες με σχόλια των επισκεπτών ή των χρηστών μίας κοινότητας και σελίδες με τα προσωπικά στοιχεία πελατών.

5.3 Η διαδικασία της κατασκευής συνδέσμων

Όπως γίνεται κατανοητό από τα παραπάνω, η απόκτηση εισερχόμενων συνδέσμων αποτελεί ζωτικής σημασίας διαδικασία για τη βελτιστοποίηση της κατάταξης των ιστοσελίδων στα αποτελέσματα όλων των μηχανών αναζήτησης, είτε αυτές δίνουν αξία σε ποιότητα και δημοσιότητα, είτε αποκλειστικά στην ποσότητα των συνδέσμων αυτών. Ανεξάρτητα από τη φιλοσοφία της εκάστοτε μηχανής αναζήτησης, ο αλγόριθμος PageRank εκφράζει πετυχημένα την πιθανότητα ένας τυχαίος χρήστης να καταλήξει σε μία ορισμένη ιστοσελίδα. Υπάρχουν πέντε διαφορετικές στρατηγικές κατασκευής συνδέσμων, οι οποίες ακολούθως αναλύονται κατά σειρά προτεραιότητας.

5.3.1 Φυσική απόκτηση συνδέσμων

Πρόκειται για την πρώτη και κύρια στρατηγική δημιουργίας εισερχόμενων συνδέσμων, τη μόνη ενδεχομένως κοινώς αποδεκτή, η οποία έχει τα καλύτερα συγκριτικά αποτελέσματα αλλά, παράλληλα, δεν αφήνει πολλά περιθώρια τεχνικής βελτιστοποίησης. Περιλαμβάνει τη δημιουργία φρέσκου, αυθεντικού, ποιοτικού κι ενδιαφέροντος περιεχομένου το οποίο οι επισκέπτες ενός ιστοτόπου θα θελήσουν να διαδώσουν στους φίλους τους, στους επισκέπτες των προσωπικών τους σελίδων, στα μέσα κοινωνικής δικτύωσης, σε φόρουμ συζητήσεων και τα επιστημονικά τους άρθρα (ως αναφορές).

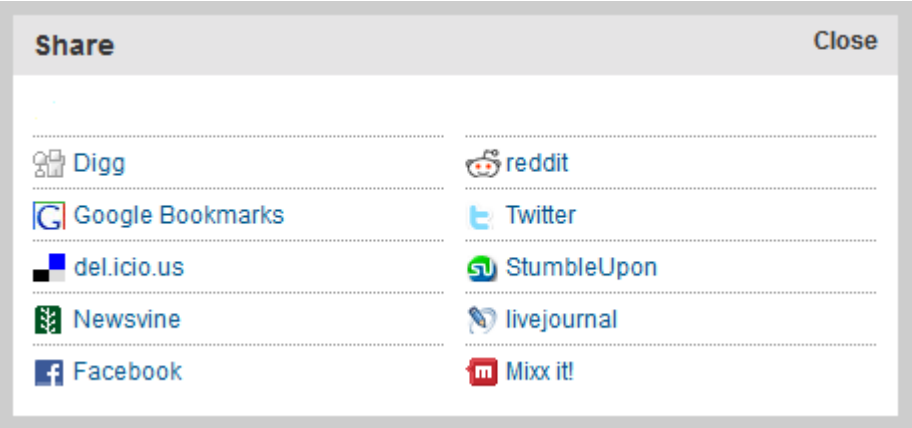
Μάλιστα, υπάρχει μία ιδιαίτερα τετριμμένη φράση στη βιομηχανία του Search Engine Optimization που λέει ότι το περιεχόμενο είναι ο βασιλιάς (“content is king”), την οποία πρώτος διατύπωσε ο Bill Gates, ιδρυτής της Microsoft.

Μία τακτική που συχνά ακολουθείται είναι η δημιουργία θεματικού περιεχομένου, όπως, για παράδειγμα, είναι οι κριτικές τέχνης ή προϊόντων και υπηρεσιών, επιστημονικά άρθρα ή άρθρα ερμηνευτικής δημοσιογραφίας πάνω στα πιο πρόσφατα νέα, χρήσιμοι ταξιδιωτικοί οδηγοί, συνταγές και συμβουλές.

Επίσης, πέρα από τεχνικές marketing (διαγωνισμούς, επιβραβεύσεις μελών κ.α.), μία στρατηγική που προσελκύει συνδέσμους είναι η δημιουργία κοινοτήτων σε έναν ιστοχώρο, μέσω της κατασκευής chat box (κουτιού συζητήσεων) ή ενός forum, με ενεργά μέλη και συντονιστές, πάνω σε θέματα που απασχολούν τους δυνητικούς επισκέπτες του ιστοτόπου, ή που σχετίζονται με τις υπηρεσίες και τα προϊόντα της επιχείρησης.

Τέλος, για την ενίσχυση του βαθμού PageRank της αρχική σελίδας ενός ιστοτόπου που, πολλές φορές, είναι και το ζητούμενο, συνίσταται η κατασκευή χρήσιμων πρόσθετων εφαρμογών κι εργαλείων, όπως πρόσθετων πρόγνωσης καιρού, θεματικών αριθμομηχανών, πρόσθετων ελέγχου διαθεσιμότητας εισιτηρίων, ενσωματωμένων στην αρχική σελίδα.

Στην κατηγορία της φυσικής απόκτησης συνδέσμων, εντάσσεται και η τεχνική της προσθήκης κοινωνικού σελιδοδείκτη (social bookmarking). Οι κοινωνικοί σελιδοδείκτες επιτρέπουν στους χρήστες του Διαδικτύου να αποθηκεύσουν, όπως οι σελιδοδείκτες των φυλλομετρητών, αλλά και να μοιραστούν τους αγαπημένους τους συνδέσμους. Τέτοιες διαδικτυακές υπηρεσίες που προσφέρουν την προβολή κοινωνική σελιδοδεικτών είναι οι Delicious, Digg, Technorati, StumbleUpon που εντάσσονται στα κοινωνικά μέσα δικτύωσης (social networks), καθώς και το ελληνικό FreeStuff. Σε αυτή την κατεύθυνση, είναι πλέον σύνηθες να κατασκευάζεται και να ενσωματώνεται σε ορισμένες σελίδες μία γραμμή εργαλειών στην οποία δίνεται η δυνατότητα δημοσίευσης της διεύθυνσης URL και του περιεχομένου της σελίδας, όπως αυτή που παρέχεται στην ιστοσελίδα της βρετανικής εφημερίδας “The Guardian”:



Εικόνα 32 Παράδειγμα γραμμής εργαλειών δημοσίευσης σε Social Media

Η διαδικασία του social bookmarking αποτελεί μία πολύ σημαντική τεχνική βελτιστοποίησης των ιστοσελίδων στα κοινωνικά μέσα δικτύωσης (Social Media Optimization).

5.3.2 Δημιουργία εισερχόμενων συνδέσμων

Πρόκειται για μία τεχνική βελτιστοποίησης που είναι ηθικά αμφισβητήσιμη, καθώς χειραγωγεί την ιδιότητα της φυσικής διασύνδεσης του Διαδικτύου και τη φιλοσοφία της θεώρησης του συνδέσμου ως αντικειμενικής ψήφου από μία σελίδα σε μία άλλη.

Μία ασφαλής και θεμιτή, κατά τα πρότυπα των μηχανών αναζήτησης, τεχνική είναι η κατασκευή ενός ξεχωριστού blog, ή αντίστοιχου θεματικού χώρου ως υποτομέα (sub-domain) του ιστοχώρου (π.χ. blog.site.com ή press.site.com), στον οποίο να περιγράφονται νέα προϊόντα ή υπηρεσίες, να ανακοινώνονται δελτία τύπου του οργανισμού ή της ψηφιακής επιχείρησης ή να διεξάγονται διαγωνισμοί και να συνδέουν προς διάφορες σελίδες του κεντρικού ιστοτόπου.

Αντίθετα, η δημιουργία ξεχωριστών ιστοσελίδων για την αναπαραγωγή του περιεχομένου του αρχικού ιστοτόπου, η εσωτερική (on-page) βελτιστοποίηση αυτών και η κατασκευή εξερχόμενων συνδέσμων προς τον αρχικό ιστότοπο θεωρείται ανήθικη τεχνική βελτιστοποίησης. Συν τοις άλλοις, η τεχνική αυτή ενέχει τον κίνδυνο ποινής στα αποτελέσματα αναζήτησης για διπλότυπο περιεχόμενο.

Στην κατηγορία της δημιουργίας εισερχόμενων συνδέσμων εντάσσεται, γενικά, ένα άλλο μεγάλο μέρος της βελτιστοποίησης των ιστοσελίδων στα Social Media και Social Networks. Το κομμάτι αυτό αφορά στην αξιοποίηση των εξερχόμενων συνδέσμων που δύναται ο χρήστης να προσθέσει στη σελίδα του σε αυτά, όπως, για παράδειγμα, το Facebook, το MySpace, το LinkedIn κ.α., καθώς όλες οι επιχειρήσεις και οι οργανισμοί έχουν τη δυνατότητα εκπροσώπησης σε αυτά, μία ή περισσότερες φορές. Προφανώς, αυτό που έχει ιδιαίτερη σημασία είναι η ύπαρξη ή μη του γνωρίσματος **rel = "no follow"** στους εξερχόμενους από αυτά συνδέσμους καθώς και η αξία που αποδίδει ή μπορεί να αποδώσει σε αυτούς η κάθε μηχανή αναζήτησης, ανά περιόδους, στα μέσα αυτά. Για παράδειγμα, το Facebook ανέκαθεν τοποθετούσε ένα «τείχος» που το καθιστούσε αόρατο και δεν επέτρεπε στις μηχανές αναζήτησης να ανιχνεύσουν το περιεχόμενο στις σελίδες των χρηστών του, εφόσον οι ανιχνευτές των μηχανών δε διαθέτουν προσωπική σελίδα, άρα ούτε και το δικαίωμα προβολής του περιεχομένου. Αντίθετα, παραδείγματα social media που δεν αφαιρούν την αξία των συνδέσμων (κάνοντας χρήση του **rel = "no follow"**) είναι τα παρακάτω:

- www.digg.com
- www.technorati.com
- www.Slashdot.org
- www.furl.net
- www.reddit.com
- http://del.icip.us
- www.stumbleupon.com
- www.flickr.com

Η συμμετοχή στα μέσα κοινωνικής δικτύωσης δεν αποδίδει μόνο σε επίπεδο συνδέσμων και PageRank, αλλά και σε επίπεδο κίνησης και επισκεψιμότητας, καθώς πολλές ψηφιακές επιχειρήσεις δρουν αποκλειστικά μέσω αυτών.

Τέλος, συνίσταται η συμμετοχή και εκπροσώπηση μίας επιχείρησης ή ενός οργανισμού που δραστηριοποιείται στο Διαδίκτυο σε διάφορα σχετικά blogs ή microblogs (όπως το twitter) και forum, με την προσθήκη σχολίων και δημοσιεύσεων σε αυτά. Στα σχόλια και τις δημοσιεύσεις, ο διαχειριστής ενός ιστοτόπου δύναται να προσθέσει συνδέσμους για την υποστήριξη των θέσεών του, την προώθηση υπηρεσιών και προϊόντων σχετικών με τα νήματα ή τα blogs στα οποία συμμετέχει. Στην ίδια κλίμακα, είναι δυνατή και η συμμετοχή σε σελίδες wiki (όπως το Wikipedia, ή άλλα εξειδικευμένα wikis), το περιεχόμενο των οποίων μπορεί να υποστεί επεξεργασία από οποιονδήποτε, ενώ μπορούν να προστεθούν σχετικοί σύνδεσμοι στα διάφορα άρθρα αυτών.

5.3.3 Αίτηση εισερχόμενων συνδέσμων (μίας κατεύθυνσης)

Με τον όρο αυτό, εννοούμε την αίτηση ενός διαχειριστή σε κάποιον ξένο ιστότοπο για τη δημιουργία ενός συνδέσμου από τον ιστότοπο αυτό σε μία ιστοσελίδα του διαχειριστή, χωρίς την ύπαρξη κάποιας οικονομικής συναλλαγής μεταξύ των δύο μερών.

Φυσικά, το αίτημα ικανοποιείται μόνο μετά την συγκατάθεση κάποιου φυσικού προσώπου, υπεύθυνου του εξωτερικού ιστοτόπου, η οποία έπεται της επίσκεψης αυτού στην ιστοσελίδα. Αυτό σημαίνει ότι η σελίδα πρέπει να πληροί κάποιες προϋποθέσεις, βασικότερη εκ των οποίων είναι το ενδιαφέρον περιεχόμενο, όπως αναλύθηκε στην πρώτη τεχνική κατασκευής συνδέσμων, αυτή της φυσικής απόκτησης. Ακόμη πιο σημαντικός παράγοντας για τη συγκατάθεση του υπεύθυνου του ιστοτόπου στον οποίο απευθύνεται ο διαχειριστής είναι η σχετικότητα του περιεχομένου της σελίδας με τον ιστότοπο αυτό.

Η πλέον δομημένη στρατηγική που εντάσσεται σε αυτή την κατηγορία ξεκινάει με τη σύγκριση των σελίδων που συνδέουν προς τον υπό εξέταση ιστότοπο και αυτών που συνδέουν προς κάποιον άλλον ανταγωνιστικό ιστότοπο. Αυτό μπορεί να πραγματοποιηθεί αναζητώντας στις μηχανές τους εισερχόμενους συνδέσμους των δύο ιστοτόπων, κάνοντας χρήση του τελεστή «link:», όπως έχει αναλυθεί σε προηγούμενο κεφάλαιο, ή, εφόσον οι μηχανές αναζήτησης τείνουν να μην ενημερώνουν ικανοποιητικά το ευρετήριό τους για τις ανάγκες αυτών των αναζητήσεων (παρέχοντας ελλιπή αποτελέσματα), του τελεστή των εισαγωγικών εντός των οποίων η ακριβής διεύθυνση URL του ιστοτόπου (“URL“, όπως

“http://www.ntua.gr”), αναζήτηση η οποία θα επιστρέψει τα ζητούμενα αποτελέσματα και, ενδεχομένως, μερικά ανεπιθύμητα. Μετά τη σύγκριση, μπορούν να εντοπισθούν οι ιστοσελίδες που συνδέουν προς τον ανταγωνιστή και όχι τον υπό εξέταση ιστότοπο και να προσεγγισθούν με κάποια αίτηση για σύνδεσμο, αφού πρώτα πραγματοποιηθεί μία ανάλυση των λόγων για τους οποίους οι ιστοσελίδες αυτές δε συνδέουν προς τον υπό εξέταση ιστότοπο και, κυρίως, των λόγων για τους οποίους οι ιστοσελίδες αυτές θα δέχονταν να συνδέουν προς τον ιστότοπο.

Μία παρόμοια στρατηγική περιλαμβάνει την αναζήτηση των κρίσιμων λέξεων ή φράσεων – κλειδιών, δηλαδή των όρων αναζήτησης για τους οποίους επιθυμείται η υψηλή θέση της ιστοσελίδας στα αποτελέσματα αναζήτησης, και ο εντοπισμός των πρώτων 100 αποτελεσμάτων. Στη συνέχεια, η αίτηση μπορεί να πραγματοποιηθεί στους διαχειριστές αυτών των ιστοσελίδων.

Στην ίδια κατηγορία εντάσσονται και οι καταχωρήσεις στους διαδικτυακούς καταλόγους (web directories / human – powered directories), όπως αυτοί μελετήθηκαν σε προηγούμενο κεφάλαιο. Η διαδικασία, συνήθως, περιλαμβάνει τον εντοπισμό των καταλόγων, την εύρεση της κατάλληλης κατηγορίας η οποία σχετίζεται με τον υπό εξέταση ιστότοπο, η καταχώρηση της διεύθυνσης URL της αρχικής σελίδας του ιστοτόπου, μαζί με ορισμένα άλλα στοιχεία (περιγραφή ιστοτόπου, λέξεις – κλειδιά, κάποιος λογαριασμός ηλεκτρονικής αλληλογραφίας εντός του ονόματος τομέα του ιστοτόπου και ένα τηλέφωνο επικοινωνίας), και ακολουθεί η εξέταση του ιστοτόπου από τους διαχειριστές του εκάστοτε καταλόγου και η έγκριση του συνδέσμου προς προσθήκη στον κατάλογο. Ιδιαίτερα σημαντικό αποτελεί το πρώτο στάδιο της διαδικασίας, αυτό της εύρεσης των βέλτιστων καταλόγων, καθώς η καταχώρηση του ιστοτόπου σε περιορισμένο αριθμό ποιοτικών (με υψηλό βαθμό PageRank) καταλόγων μπορεί να αποδειχθεί πολύ πιο αποδοτική από την αντίστοιχη καταχώρηση του ιστοτόπου σε πολλούς καταλόγους, αμφιβόλου ποιότητας, καθώς η τελευταία ενέχει και τον κίνδυνο της φάρμας συνδέσμων, ανάμεσα σε κάποιους εξ αυτών, που μπορεί να επιφέρει τα αντίθετα από τα επιθυμητά αποτελέσματα. Ένα ιδιαίτερα ποιοτικό διαδικτυακό κατάλογο αποτελεί ο Ανοικτός Κατάλογος, το **Open Directory Project** (http://www.dmoz.org), ο οποίος αποτελεί τον επίσημο σύμβουλο της μηχανής Google, καθώς τροφοδοτεί, ανά περιόδους, τον αντίστοιχο κατάλογο της Google (Google Directory), για τον οποίον δεν υπάρχει η

δυνατότητα προσθήκης και στον οποίον προκρίνονται μόνο όσες ιστοσελίδες εγκρίνονται στο Open Directory Project (ODP). Η έγκριση ενός συνδέσμου στον ODP είναι εξαιρετικά δύσκολη και ενδέχεται να μεσολαβήσει αρκετά μεγάλο διάστημα μεταξύ της καταχώρησης αιτήματος και της εξέτασης του ιστοτόπου. Υπενθυμίζεται ότι η καταχώρηση ενός συνδέσμου σε αρκετούς εξίσου ποιοτικούς καταλόγους με τον ODP ενδέχεται να απαιτεί κάποια οικονομική επιβάρυνση, που πολλές φορές δεν είναι καθόλου αμελητέα, με τον πιο ακριβό κατάλογο να είναι αυτός της Yahoo (Yahoo Directory) ο οποίος απαιτεί ένα ετήσιο τέλος των $299 (ή $600 για περιεχόμενο ιστοτόπου κατάλληλο μόνο για ενήλικες).

Τονίζεται, για άλλη μία φορά, ότι αυτή η τεχνική βελτιστοποίησης θα πρέπει να αφορά εξωτερικές ιστοσελίδες με υψηλό βαθμό PageRank και όσο το δυνατόν λιγότερους εξερχόμενους συνδέσμους, για τη μέγιστη δυνατή ροή βαθμού PageRank μέσω του συνδέσμου.

Τέλος, προτεραιότητα θα πρέπει να αποτελεί η αίτηση να γίνεται σε προσωπικό επίπεδο και όχι με κάποιον αυτόματο τρόπο (π.χ. μέσω λογισμικού αποστολής μαζικών μηνυμάτων ηλεκτρονικής αλληλογραφίας).

5.3.4 Αίτηση αμοιβαίων συνδέσμων

Η τεχνική αυτή βελτιστοποίησης περιλαμβάνει την ίδια ακριβώς διαδικασία με την προηγούμενη τεχνική, με τη διαφορά της παροχής του επιπλέον κινήτρου της αμοιβαίας ανταλλαγής συνδέσμων (reciprocal linking), έτσι ώστε να ωφεληθούν και οι δύο ιστότοποι από αυτήν.

Είναι δεδομένο ότι με τον τρόπο αυτό, ενδέχεται οι δύο σύνδεσμοι να αλληλοεξουδετερώνονται, καθώς ο ένας ιστότοπος αποδίδει PageRank στον άλλον, με αποτέλεσμα και οι δύο ιστότοποι να παρουσιάζουν διαρροή βαθμού PageRank. Παράλληλα, η αμοιβαία ανταλλαγή συνδέσμων αποτελεί στρατηγική που μηχανές όπως η Google την στοχοποιούν ως παράδειγμα προσπάθειας χειραγώγησης του βαθμού PageRank μίας σελίδας, οπότε ενδέχεται να αποδίδουν ποινές σε ιστοτόπους που πραγματοποιούν κατάχρηση αυτής της τεχνικής.

Στην περίπτωση, λοιπόν, της μη εξουδετέρωσης των αμοιβαίων συνδέσμων, συνίσταται η αναζήτηση εκείνων των ιστοτόπων, η δημιουργία συνδέσμου και η διαρροή βαθμού PageRank προς τους οποίους πρόκειται να αποσβέσει μέσω της λήψης συνδέσμου από αυτούς. Αυτό μπορεί να σημαίνει ότι η εξωτερική ιστοσελίδα με την οποία θα πραγματοποιηθεί η ανταλλαγή έχει υψηλότερο βαθμό PageRank, λιγότερους εξερχόμενους συνδέσμους, ή και τα δύο. Επίσης, είναι προτιμότερη η δημιουργία αμοιβαίων, σε επίπεδο ιστοτόπου και όχι σελίδας, συνδέσμων, η δημιουργία, δηλαδή, ενός συνδέσμου από τη σελίδα Α του ιστοτόπου Χ προς τη σελίδα Β του ιστοτόπου Υ και, παράλληλα, η λήψη ενός συνδέσμου από τη σελίδα Δ του ιστοτόπου Υ σε μία ορισμένη σελίδα Γ του ιστοτόπου Χ.

Παράλληλα, δίνεται η δυνατότητα στους διαχειριστές ιστοτόπων από δίκτυα συνδέσμων (link networks) να κάνουν ευρεία χρήση αυτής της τεχνικής. Πρόκειται για την ανταλλαγή συνδέσμων μεταξύ σχετικών ιστοσελίδων, μέσω μίας τρίτης υπηρεσίας που αναλαμβάνει το ρόλο του μεσολαβητή (ενδεχομένως κι επί πληρωμή). Η εφαρμογή της μεθόδου αυτής δε συνίσταται, αντιθέτως αντενδείκνυται, καθώς, σε περίπτωση που εντοπισθεί από τις μηχανές αναζήτησης, είναι πολύ πιθανή η τιμωρία των εμπλεκόμενων ιστοτόπων στα οργανικά και μη αποτελέσματα.

5.3.5 Αγορά συνδέσμων

Στην τεχνική αυτή ανήκουν όλες οι προηγούμενες τεχνικές δημιουργίας συνδέσμων (πλην της φυσικής απόκτησης), στην περίπτωση που απαιτείται κάποια οικονομική συναλλαγή μεταξύ των δύο ενδιαφερόμενων μερών. Έτσι, μπορούν να αγοραστούν σύνδεσμοι από κάποιον διαδικτυακό κατάλογο, από άλλον ιστότοπο αλλά και από κάποιον διαμεσολαβητή (link broker).

Σύμφωνα με τον Matt Cutts, μηχανικό και υπέυθυνο δελτίων τύπου της Google, ορισμένες περιπτώσεις δεν εξετάζονται αυτόματα από τους αλγορίθμους, αλλά διεξοδικά, με τη συμμετοχή του ανθρώπινου παράγοντα, για τον εντοπισμό φαινομένων που παραβιάζουν τους όρους που η μηχανή αναζήτησης της Google έχει θέσει, όσον αφορά τα πρότυπα και την ανταλλαγή συνδέσμων. Έτσι, ιστοσελίδες που εντοπίζεται ότι πωλούν συνδέσμους έναντι χρηματικής αμοιβής δε θα τιμωρούνται με ποινές στα αποτελέσματα αναζήτησης (θα

συνεχίσουν να κατέχουν τις ίδιες θέσεις στα αποτελέσματα, για τις διάφορες αναζητήσεις), αλλά θα χάνουν τη δυνατότητά τους να αποδίδουν αξία στους συνδέσμους τους, με αποτέλεσμα να διακόπτεται η ροή βαθμού PageRank. Παράλληλα, συνίσταται επισήμως η χρήση του γνωρίσματος **rel = "no follow"** όταν υπάρχει οικονομική συναλλαγή. Με άλλα λόγια, επιτρέπεται η αγορά συνδέσμων όταν αυτοί αποσκοπούν σε αύξηση της επισκεψιμότητας και σε branding, αλλά όχι όταν ο στόχος είναι η ροή PageRank (εφόσον αυτή μεταφράζεται ως αντικειμενική ψήφος μίας σελίδας σε μία άλλη) (Matt Cutts: Gadgets, Google, and SEO, 2005).

5.4 Συμπεράσματα και πρακτικές

Από την παραπάνω θεωρητική ανάλυση του βαθμού PageRank αλλά και τις τεχνικές βελτιστοποίησης αυτού, εξάγονται τα παρακάτω συμπεράσματα και πρακτικές:

α) Συνίσταται οι επιθυμητοί όροι για τους οποίους επιδιώκεται η βελτιστοποίηση ενός ιστοτόπου να τοποθετούνται σε μεγάλη συχνότητα και διαφορετικές θέσεις στις σελίδες με υψηλό βαθμό PageRank εντός του ιστοτόπου. Πρακτικά, αυτό σημαίνει ότι οι τεχνικές βελτιστοποίησης εντός της ιστοσελίδας που αναλύθηκαν στο προηγούμενο κεφάλαιο είναι καλό και πρέπει να αξιοποιούνται πλήρως στις σελίδες εκείνες που έχουν υψηλό βαθμό PageRank.

β) Οι σύνδεσμοι με το γνώρισμα **rel = "no follow"** δε μεταφέρουν καμία αξία, σε όρους PageRank, στη σελίδα και μπορούν μόνο να αποδώσουν σε όρους επισκεψιμότητας και branding.

γ) Οι μηχανές αναζήτησης δείχνουν ιδιαίτερη αυστηρότητα ως προς τη ροή του βαθμού PageRank ανά τις σελίδες. Κατά την καταχώρηση, για παράδειγμα, ενός συνδέσμου σε διαδικτυακούς καταλόγους, είναι προτιμότερο να ποικίλλει το anchor text, ανά τα αιτήματα, ώστε η σύνδεση προς τον υπό εξέταση ιστότοπο να φαίνεται όσο το δυνατόν πιο φυσική στις μηχανές αναζήτησης. Παράλληλα, είναι προτιμότερο η διαδικασία της κατασκευής συνδέσμων να γίνεται διαχρονικά και όχι μέσα σε ένα ορισμένο, μικρό διάστημα, καθώς επίσης και να γίνεται χειροκίνητα και σε προσωπικό επίπεδο, όχι με λογισμικά

αυτοματοποίησης. Ο ρυθμός απόκτησης των συνδέσμων είναι ένας παράγοντας που οι μηχανές αναζήτησης εξετάζουν.

δ) Τέλος, η συμμετοχή σε κάποιο forum ή blog οφείλει να πραγματοποιείται με τρόπο φυσικό, που να συμβάλλει στη συζήτηση, και όχι με την απευθείας παράθεση του συνδέσμου, χωρίς προηγούμενο συμμετοχής, καθώς, πέραν των μηχανών αναζήτησης, το λόγο έχουν επίσης και τα μέλη και οι συντονιστές των blogs και των νημάτων ενός forum.

6 *Συμπεράσματα & Προοπτικές*

6.1 Συμπεράσματα

Η παρούσα μελέτη αποτελεί μία επισταμένη προσπάθεια να συγκεντρωθούν και να συμπληρωθούν θεωρητικές αναλύσεις αλλά και εμπειρικές προσεγγίσεις που σχετίζονται με όλο το εύρος των θεμάτων που συνδέονται με τη διαδικασία της βελτιστοποίησης της θέσης που λαμβάνουν οι ιστοσελίδες στα οργανικά αποτελέσματα των μηχανών αναζήτησης.

Αρχικά, μελετήθηκε η λειτουργία των μηχανών αναζήτησης, όσον αφορά τα δύο στάδια που προηγούνται της κατάταξης των αποτελεσμάτων, δηλαδή την ανίχνευση των εγγράφων στο Διαδίκτυο και την ευρετηρίαση αυτών, αλλά και το στάδιο που έπεται της κατάταξης, αυτό της επεξεργασίας των ερωτημάτων. Έτσι, αναλύθηκαν οι πολιτικές που διέπουν και μοντελοποιούν τη διαδικασία της ανίχνευσης, όπως η δομημένη ανάθεση διευθύνσεων URL στον ανιχνευτή, η συχνότητα και τα κριτήρια επανεπίσκεψης των σελίδων με σκοπό τη βελτιστοποίηση της διαδικασίας αλλά και την αποφυγή υπερφορτώσεων στις σελίδες, αλλά και οι αλγόριθμοι που έχουν έως τώρα χρησιμοποιηθεί για την επίλυση όλων των προβλημάτων που συνοδεύουν τη διαδικασία της ανίχνευσης από τις μηχανές αναζήτησης. Έπειτα, μελετήθηκαν τα χαρακτηριστικά ενός ευρετηρίου, όσον αφορά τους χρονικούς και υλικούς πόρους που απαιτούνται για την υλοποίηση ενός, οι δομές που μπορεί να έχει το ευρετήριο μίας μηχανής αναζήτησης, όπως η ιδιαίτερα διαδεδομένη δομή του ανεστραμμένου ευρετηρίου που χρησιμοποιείται από πολλές μηχανές ελαφρώς τροποποιημένη, καθώς και οι γλωσσικές και μη προκλήσεις που αντιμετωπίζουν οι σύγχρονες μηχανές αναζήτησης κατά την ανάλυση των εγγράφων. Ιδιαίτερο ενδιαφέρον παρουσιάζει και η επεξεργασία των ερωτημάτων των χρηστών από τις μηχανές, ενώ αναλύεται εκτενέστερα ο τρόπος με τον οποίο επιτυγχάνεται η διαδικασία σε ένα ανεστραμμένο ευρετήριο, αλλά και οι διάφοροι τελεστές αναζήτησης, τόσο από την σκοπιά των χρηστών και των δυνατοτήτων

συγκεκριμενοποίησης των ερωτημάτων τους όσο κι από την σκοπιά των μηχανών αναζήτησης.

Στη συνέχεια και κατά τη μελέτη του σταδίου της κατάταξης των ιστοσελίδων – αποτελεσμάτων αναζήτησης, πραγματοποιήθηκε μία προσπάθεια προσέγγισης των εσωτερικών κι εξωτερικών παραγόντων που επιδρούν στους αλγορίθμους κατάταξης, όπως αυτοί εισήχθησαν από τους ιδρυτές της πλέον δημοφιλούς Google, αλλά και όπως προσπάθησαν να εντοπίσουν πρακτικά διάφοροι ερευνητές στο παρελθόν, κατά την προσπάθεια προσέγγισης του ακριβούς αλγορίθμου που χρησιμοποιούν οι μηχανές αναζήτησης. Παρατηρήθηκε, λοιπόν, ότι, πέραν της θέσης, της γραμματοσειράς, της διάκρισης των χαρακτήρων σε πεζά ή κεφαλαία και την εμφάνιση του όρου αναζήτησης στο anchor text, παράγοντες όπως το συνολικό περιεχόμενο και η γλώσσα στην οποία αυτό είναι γραμμένο, η συχνότητα και η σχετική θέση του όρου αναζήτησης (π.χ. σε επικεφαλίδες, τον τίτλο ή τη διεύθυνση URL), η μορφοποίησή των εικόνων και κειμένων και η σύνδεση του εγγράφου με το υπόλοιπο Διαδίκτυο (εισερχόμενοι κι εξερχόμενοι σύνδεσμοι, αναφορές σε καταλόγους) επηρεάζουν καθοριστικά τη θέση του εγγράφου στα αποτελέσματα αναζήτησης. Η προσέγγιση αυτή υπήρξε καθοριστική για την ακόλουθη ανάλυση των παραγόντων αυτών, καθώς και άλλων που εμπειρικά έχει αποδειχθεί πως σχετίζονται με τους αλγορίθμους κατάταξης.

Έτσι, σε πρώτο στάδιο, αναλύθηκαν διεξοδικά όλοι εκείνοι οι παράγοντες που αφορούν την εσωτερική οργάνωση της ιστοσελίδας και του διακομιστή φιλοξενίας αυτής, ενώ παρουσιάστηκαν οι αντίστοιχες τεχνικές βελτιστοποίησης των παραγόντων αυτών. Έτσι, αναλύθηκε η χρησιμότητα του αρχείου αποκλεισμού ανιχνευτών robots.txt και ορισμένες τεχνικές σύνταξης αυτού για την κατεύθυνση των ανιχνευτών αναζήτησης εκεί που επιθυμείται, μελετήθηκαν οι λειτουργίες όλων των meta ετικετών και κατά πόσο υποστηρίζονται και λαμβάνονται υπόψη από τις σημαντικότερες εκ των μηχανών αναζήτησης καθώς και η ορθή σύνταξη αυτών για κάθε επιθυμητή από το διαχειριστή περίπτωση, ενώ μελετήθηκαν εκτενέστερα όλοι οι παράγοντες που σχετίζονται με το περιεχόμενο της ιστοσελίδας, όπως ο τίτλος και οι επικεφαλίδες της σελίδας, τα διάφορα γνωρίσματα παράθεσης συνδέσμων και εικόνων, καθώς και οι ετικέτες μορφοποίησης του περιεχομένου που αποδίδουν σημασιολογικό βάρος άρα και χαρακτηρίζουν με μεγαλύτερη βαρύτητα το περιεχόμενο ενός εγγράφου. Έπειτα, ιδιαίτερη έμφαση δόθηκε στη δομή της διεύθυνσης URL και τη χρησιμότητά της τόσο από την σκοπιά των μηχανών όσο και από την οπτική γωνία των χρηστών, ενώ περιγράφηκαν οι προκλήσεις που συναντώνται στον έλεγχο αυτής, ανάλογα με το δυναμικό ή στατικό χαρακτήρα παραγωγής τους, καθώς και η μεθοδολογία που οφείλει να ακολουθείται για τη μετατροπή των δυναμικών και ανεξέλεγκτων διευθύνσεων URL σε στατικές και στοχευμένες στο περιεχόμενο της σελίδας που

εκπροσωπούν. Στη συνέχεια, αναλύθηκαν διεξοδικά όλα τα είδη χαρτών ιστοτόπων XML που υποστηρίζονται από τις μηχανές αναζήτησης, καθώς και οι τρόποι σύνταξης αυτών για κάθε περίπτωση. Τέλος, πραγματοποιήθηκε μία μοναδική ακαδημαϊκή αναφορά στη δυνατότητα βελτιστοποίησης των ιστοσελίδων περιεχομένου τεχνολογίας Flash, που, στη βιομηχανία του SEO, χαρακτηρίζονται και αντιμετωπίζονται συνήθως ως καταδικασμένες να αποτύχουν στα αποτελέσματα. Έτσι, παρουσιάστηκε εκτενώς η τεχνολογία SWFobject, μία δέσμη ενεργειών JavaScript, η οποία επιτρέπει τον παράλληλο προγραμματισμό του περιεχομένου σε Flash αλλά και κώδικα HTML, καθώς και όλες τις γνωστές δυνατότητες βελτιστοποίησης του τελευταίου, ενώ παρέχει την επιπλέον δυνατότητα προβολής του περιεχομένου HTML στους χρήστες που δε διαθέτουν καθόλου ή τον κατάλληλο Flash Player που απαιτείται για την αναπαραγωγή του Flash περιεχομένου.

Παράλληλα, αναλύθηκαν ορισμένα στοιχεία που κάποτε υπήρξαν καθοριστικά για τη θέση των ιστοσελίδων στις μηχανές αναζήτησης ενώ τώρα δεν επιδρούν ή επιδρούν σε αμελητέο βαθμό σε αυτήν. Έτσι, περιγράφηκε η meta ετικέτα λέξεων – κλειδιών και, μέσα από την παρουσίαση της ιστορίας και του κύκλου ζωής της, εξηγήθηκαν οι λόγοι που οδήγησαν τις μηχανές αναζήτησης στην παραίτησή τους από την ετικέτα αυτή, ενώ, για τους ίδιους ακριβώς λόγους (κυρίως την κατάχρηση, δηλαδή, της ιδιότητάς τους από τους υπεύθυνους marketing των ιστοσελίδων, μέσα στη δεκαετία του 1990 και τις αρχές της δεκαετίας του 2000), συμπεραίνεται μαθηματικά η αμελητέα ή ανύπαρκτη επίδραση της απόλυτης συχνότητας εμφάνισης ενός όρου (TF-IDF) σε μία σελίδα στην κατάταξή της στα αποτελέσματα αναζήτησης.

Σε επίπεδο διακομιστή, προσεγγίσθηκε μία καθολική στρατηγική για την επιλογή του τύπου και του ονόματος τομέα, μελετήθηκε η γεωγραφική τοποθέτηση αυτού και τον τρόπο με τον οποίο οι μηχανές αναζήτησης αξιοποιούν αυτή την πληροφορία για την παρουσίαση των πλέον γεωγραφικά σχετικών αποτελεσμάτων στους χρήστες ανά τον κόσμο, ενώ παρουσιάσθηκαν αναλυτικά οι τεχνικές της ανακατεύθυνσης και κανονικοποίησης που επιλύουν σημαντικά ζητήματα βελτιστοποίησης και ποινικοποίησης στα αποτελέσματα των μηχανών αναζήτησης, όπως αυτό του διπλότυπου περιεχομένου. Τέλος, προσεγγίσθηκαν παράγοντες που σχετίζονται με το χρόνο, αφού περιγράφηκε το φαινόμενο «sandbox» που αφορά το χρονικό διάστημα της αφομοίωσης των καινούριων ιστοσελίδων στο ευρετήριο και τη συμπεριφορά των μηχανών απέναντι σε αυτές όπως έχει εμπειρικά διαπιστωθεί, ενώ μελετήθηκαν τα θέματα της συχνότητας ανανέωσης του περιεχομένου, της μακροβιότητας των ιστοτόπων και της συχνότητας δημιουργίας εισερχόμενων κι εξερχόμενων συνδέσμων.

Σε δεύτερο στάδιο, υπολογίστηκε ο βαθμός PageRank που χρησιμοποιεί η Google, αλλά και οι υπόλοιπες μηχανές αναζήτησης με ορισμένες τροποποιήσεις, ενώ αναλύθηκε θεωρητικά η χρησιμότητα και λειτουργία του για την ιεράρχηση των σελίδων του Διαδικτύου. Έπειτα,

προσεγγίστηκαν τρεις τεχνικές εξατομίκευσης του βαθμού PageRank, μελετήθηκαν οι παράμετροι εκείνες που επιτρέπουν μελλοντικές προσπάθειες εξατομίκευσης κι ερμηνεύθηκε κατά πώς αυτές μπορούν να συμβάλλουν στην προσέγγιση του Σημασιολογικού Ιστού (Web 3.0). Σε τελική ανάλυση του βαθμού PageRank, διατυπώθηκαν ορισμένα αξιώματα που τον χαρακτηρίζουν, βάσει της θεωρητικής ανάλυσης και του μαθηματικού υπολογισμού του, και μελετήθηκαν πώς αυτά καθορίζουν τις τεχνικές βελτιστοποίησης του τρόπου διασύνδεσης των σελίδων μεταξύ τους. Έτσι, αναπτύχθηκαν οι βασικότερες τεχνικές κατασκευής εισερχόμενων συνδέσμων προς μία ιστοσελίδα ή έναν ιστότοπο, ενώ προσεγγίστηκε και η χρησιμότητα του παράγοντα PageRank στον τρόπο με τον οποίο δομούνται εσωτερικά οι ιστότοποι.

Στις επιμέρους αναλύσεις κάθε ξεχωριστού παράγοντα βελτιστοποίησης, εξάγονται ορισμένα συμπεράσματα για τη συμπεριφορά του εκάστοτε παράγοντα, τη χρησιμότητά και βαρύτητά του στους αλγορίθμους κατάταξης, ενώ αναπτύσσονται και οι καλύτερες πρακτικές – τεχνικές βελτιστοποίησης που απορρέουν από τα συμπεράσματα αυτά.

Σε ευρύτερα πλαίσια και βάσει των ευρημάτων της μελέτης, εξάγονται ορισμένα πολύ χρήσιμα συμπεράσματα, όσον αφορά την εφαρμογή της γνώσης και την οργανική βελτιστοποίηση:

α) Κατά πρώτον, γίνεται εμφανές ότι η κατάταξη των αποτελεσμάτων πραγματοποιείται σε επίπεδο ιστοσελίδας και όχι ιστοτόπου. Παρατηρούμε, έτσι, ότι τα αποτελέσματα των μηχανών αναζήτησης δεν περιορίζονται στην προβολή μίας σελίδας ανά ιστότοπο αλλά πολλές φορές σελίδες του ίδιου ιστοτόπου εμφανίζονται, κοντά ή μακριά μεταξύ τους, ότι δεν αποδίδεται αξία συνδέσμων σε διακομιστές ή ιστοχώρους αλλά ξεχωριστές σελίδες ή ότι δε λαμβάνεται υπόψη η διάκριση των εσωτερικών κι εξωτερικών συνδέσμων για τη ροή του βαθμού PageRank. Αυτό σημαίνει ότι ο ανταγωνισμός για μία θέση στις πρώτες σελίδες των αποτελεσμάτων αναζήτησης δεν υπάρχει μόνο μεταξύ ιστοτόπων αλλά και μεταξύ σελίδων του ίδιου ιστοτόπου και προς αυτή την κατεύθυνση θα πρέπει να πραγματοποιείται η βελτιστοποίηση. Έτσι, διαφορετικές καμπάνιες κατασκευής συνδέσμων θα πρέπει να πραγματοποιούνται για κάθε ιστοσελίδα και διαφορετικές τεχνικές βελτιστοποίησης υπάρχει η δυνατότητα να υιοθετούνται και να εφαρμόζονται ανά τις σελίδες του ίδιου ιστοτόπου. Η πιο σημαντική, όμως, διαφοροποίηση που οφείλει να γίνεται είναι κατά την εστίαση σε λέξεις ή φράσεις – κλειδιά.

Συγκεκριμένα, ένα από τα μεγαλύτερα προβλήματα που παρουσιάζονται σε βελτιστοποιημένους ιστοτόπους αποτελεί το φαινόμενο του **αυτο-κανιβαλισμού** των λέξεων – κλειδιών (**keyword cannibalization** ή **self-cannibalization**). Πρόκειται για το φαινόμενο κατά το οποίο δύο ή περισσότερες ιστοσελίδες του ίδιου ιστοτόπου βελτιστοποιούνται για τις ακριβώς ίδιες λέξεις ή φράσεις. Έστω μία επιχείρηση που πουλάει εσώρουχα ηλεκτρονικά

και έχει βελτιστοποιήσει τον ιστότοπο (σελίδες για ανδρικά, γυναικεία και παιδικά εσώρουχα) για τον όρο «εσώρουχα». Με τον τρόπο αυτό, οι μηχανές αναζήτησης καλούνται να προσδιορίσουν ποια από τις δύο (τουλάχιστον) σελίδες είναι πιο σχετική με το ερώτημα αναζήτησης (ή είναι αποτελεσματικότερα βελτιστοποιημένη) και, στην περίπτωση που επιλεχθεί η πλέον ανεπιθύμητη, ενδεχομένως η επιχείρηση χάσει δυνητικούς πελάτες ή επισκέπτες, ενώ δημιουργείται επιπρόσθετος ανταγωνισμός στα αποτελέσματα αναζήτησης που επιβαρύνει τον ήδη υπάρχων ανταγωνισμό. Μία εξίσου σημαντική επίπτωση, όμως, είναι αυτή του περιορισμού των δυνατοτήτων βελτιστοποίησης, όσον αφορά, για παράδειγμα, τους εξής παράγοντες:

- Βαθμός PageRank: Εφόσον x σελίδες συνδέουν προς την σελίδα με τα ανδρικά εσώρουχα, y σελίδες συνδέουν προς την σελίδα Β που παρουσιάζει τα γυναικεία και z σελίδες συνδέουν προς την ιστοσελίδα Γ με τα παιδικά εσώρουχα, ενδεχομένως χάνεται η δυνατότητα $s \leq x + y + z$ επιπλέον ιστοσελίδες να συνδέουν προς την αρχική σελίδα και ο δυνητικός βαθμός PageRank στην ουσία χάνεται.
- Anchor text εσωτερικών συνδέσμων: Εφόσον οι σελίδες του ιστοτόπου δείχνουν προς τις σελίδες Α, Β και Γ με τις ίδιες λέξεις – κλειδιά στο anchor text, χάνεται η δυνατότητα συγκέντρωσης αξίας του anchor text σε έναν στόχο.

Έτσι, η σωστή στρατηγική είναι εκείνη που βελτιστοποιεί τις σελίδες Α, Β και Γ για τους όρους «ανδρικά εσώρουχα», «γυναικεία εσώρουχα», «παιδικά εσώρουχα» αντίστοιχα.

β) Η διαδικασία της βελτιστοποίησης των ιστοσελίδων για τις μηχανές αναζήτησης αποδεικνύεται πως δεν είναι παρά η ορθή και κατά τα πρότυπα ανάδειξη του περιεχομένου των ιστοσελίδων. Δεν πρόκειται περί χειραγώγησης των παραγόντων που επηρεάζουν την κατάταξη των αποτελεσμάτων, αλλά αξιοποίησης αυτών στο βαθμό που διευκολύνονται οι μηχανές αναζήτησης και στην κατεύθυνση που οι ίδιες κατευθύνουν. Κατ' αυτή την έννοια, δεν τίθεται θέμα περί ηθικότητας της διαδικασίας της οργανικής βελτιστοποίησης, εφόσον αυτή πραγματοποιείται στα πλαίσια και κατά τα πρότυπα και τους κανονισμούς που οι ίδιες οι μηχανές αναζήτησης έχουν ορίσει (white-hat τεχνικές).

γ) Ως επέκταση του παραπάνω συμπεράσματος, παρατηρήσαμε ότι οι τεχνικές βελτιστοποίησης δεν αφορούν στην τροποποίηση του περιεχομένου καθαυτού αλλά στη βελτιστοποίηση του τρόπου με τον οποίο αυτό παρουσιάζεται στους χρήστες και τις μηχανές αναζήτησης. Όπως ήδη αναφέρθηκε, δηλαδή, το περιεχόμενο είναι αυτό που καθορίζει κατά πόσο σημαντική ή σχετική είναι μία ιστοσελίδα, καθώς και κατά πόσο ή πόσο συχνά οι χρήστες το επανεπισκέπτονται που είναι και ο σημαντικότερος στόχος της διαχείρισης ενός ιστοτόπου. Για παράδειγμα, μία σελίδα με εφαρμογή υψηλού επιπέδου τεχνικών βελτιστοποίησης και αρκετά ευνοϊκές θέσεις στα αποτελέσματα αναζήτησης που, όμως, παρουσιάζει πολύ μεγάλα ποσοστά νέων επισκεπτών δείχνει ότι δεν ικανοποίησε το αίτημα

του χρήστη, την πρώτη φορά που αυτός την επισκέφθηκε. Τέτοια στατιστικά ενδέχεται να αξιοποιούνται τώρα ή μελλοντικά από τις μηχανές αναζήτησης, γεγονός που σημαίνει ότι η υψηλή θέση στα αποτελέσματα δεν αντικατοπτρίζει την αποτελεσματικότητα της σελίδας κι ενδέχεται μελλοντικά να χαθεί.

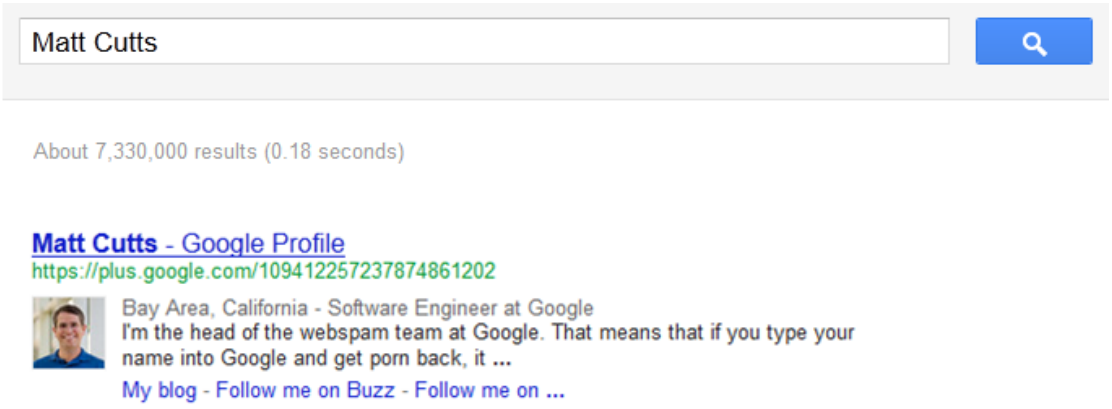
δ) Τέλος, είναι εμφανές ότι οι πρακτικές του SEO οφείλουν να προσαρμόζονται στα σύγχρονα δεδομένα, όπως τις αλλαγές που εμπειρικά παρατηρούνται στους αλγορίθμους κατάταξης των αποτελεσμάτων των μηχανών αναζήτησης ή τις νέες τεχνολογίες που εμφανίζονται και υιοθετούνται, κατά καιρούς, από τις μηχανές αναζήτησης, όπως υπήρξαν τα τελευταία χρόνια τα δεδομένα των μέσων κοινωνικής δικτύωσης (π.χ. tweets). Ο Παγκόσμιος Ιστός και ο τρόπος με τον οποίο ο χρήστης αποκτά πρόσβαση σε αυτόν συνδέονται άρρηκτα με τη λειτουργία των μηχανών αναζήτησης. Έτσι, η μετάβαση από το Web 2.0 στο Web 3.0 θα πραγματοποιηθεί όταν οι μηχανές αναζήτησης παύσουν να αναγνωρίζουν meta data (δεδομένα για τα δεδομένα) και επικεντρωθούν στα data, δηλαδή τα κριτήρια κατάταξης δε θα είναι πλέον η σύνταξη και η μορφοποίηση αλλά η ίδια η πληροφορία και η σημασιολογία (semantics). Προς την ίδια, λοιπόν, κατεύθυνση θα πρέπει να κινηθεί και προσαρμοστεί η βελτιστοποίηση για τις μηχανές αναζήτησης, όπως αναλύεται στην επόμενη παράγραφο.

6.2 Προοπτικές

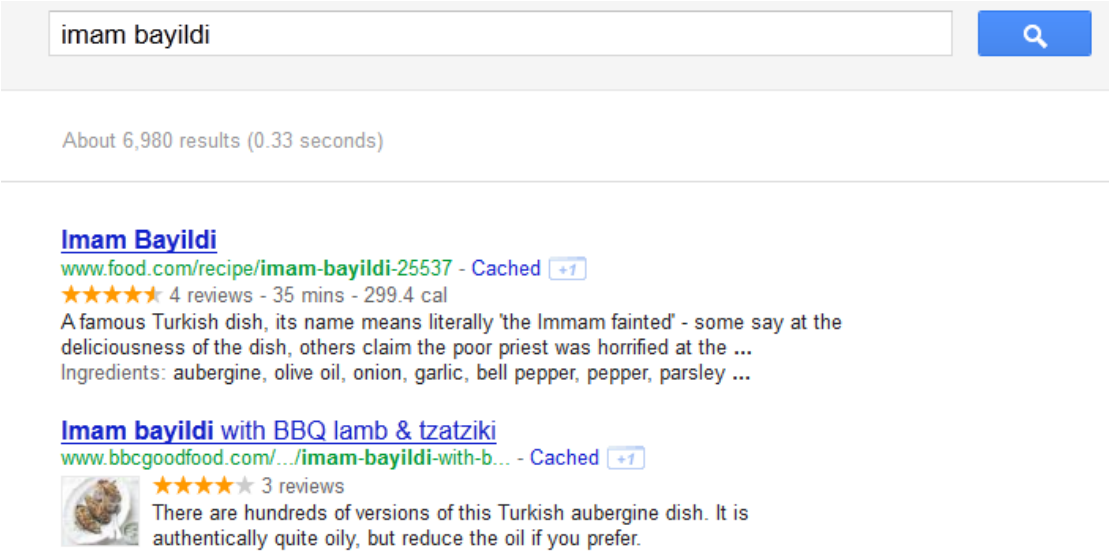
Πέραν της εξατομίκευσης του βαθμού PageRank που έχει ήδη, κατά μία έννοια, αξιοποιηθεί από τις μηχανές αναζήτησης, όπως φαίνεται από την παρουσίαση εξατομικευμένων αποτελεσμάτων, με κριτήριο τη γεωγραφική τοποθεσία του διακομιστή και του ιστοτόπου, τη γλώσσα στην οποία είναι γραμμένες οι σελίδες αυτού, το ιστορικό αναζητήσεων, τις προτιμήσεις και το προφίλ των χρηστών, ήδη οι μηχανές αναζήτησης έχουν κάνει ένα ακόμη βήμα προς την κατεύθυνση του σημασιολογικού ιστού. Έτσι, οι Google, Yahoo και Bing, παρότι ξεκίνησαν την ενσωμάτωση και τη δυνατότητα ευρετηρίασης τριών τύπων που σχετίζονται με το Σημασιολογικό Ιστό (RDFa, microformats, microdata) από το 2008, ανακοίνωσαν τον Ιούνιο του 2011, την από κοινού προώθηση του προτύπου «schema.org», ως συνέχειας της συλλογικής προσπάθειας που σήμανε η υιοθέτηση του προτύπου χαρτών ιστοτόπων «sitemaps.org». Κατά το πρότυπο αυτό, υποστηρίζονται και οι τρεις τύποι σήμανσης του Σημασιολογικού Ιστού, ενώ υιοθετείται επίσημα και δίνεται μεγαλύτερη έμφαση ο τύπος των microdata.

6.2.1 Microdata και rich snippets

Πλούσια αποσπάσματα ή περιγραφές (rich snippets) ονομάζονται οι περιγραφές των αποτελεσμάτων αναζήτησης που περιλαμβάνουν επιπλέον πληροφορίες από τις συνήθεις, οι οποίες εξαρτώνται αποκλειστικά από το περιεχόμενο της σελίδας (semantics) και τους νέους τρόπους σήμανσης (markup) που εισαγάγουν τον Σημασιολογικό Ιστό, όπως φαίνεται στα παρακάτω παραδείγματα:



Εικόνα 33 Πλούσια περιγραφή προσώπου στα αποτελέσματα της Google



Εικόνα 34 Πλούσια περιγραφή συνταγών στα αποτελέσματα της Google

Έτσι, ο κώδικας HTML παρουσίασης ενός προσώπου, γεγονότος, μέρους ή δραστηριότητας, μπορεί να εμπλουτιστεί με πληροφορίες σχετικές με το αντικείμενο παρουσίασης και το

περιεχόμενο της εκάστοτε σελίδας. Έστω, για παράδειγμα, το παρακάτω απόσπασμα μίας HTML σελίδας παρουσίασης του προσωπικού μου προφίλ:

```
<div>
      Ονομάζομαι Αλέξανδρος Νίκας και η προσωπική μου ιστοσελίδα είναι:
      <a href="http://www.alexnikas.gr">Alexandros Nikas Portfolio</a>
      Ζω στην Αθήνα, Ελλάδα και σπουδάζω ηλεκτρολόγος μηχανικός &
      μηχανικός υπολογιστών στο NTUA.
</div>
```

Στο ίδιο κομμάτι κώδικα HTML, μπορούμε να προσθέσουμε ετικέτες σήμανσης microdata, εμπλουτίζοντας την περιγραφή του αποτελέσματος, το οποίο οδηγεί στην σελίδα φιλοξενίας του προφίλ, των αναζητήσεων για το όνομά μου στις μηχανές αναζήτησης:

```
<div itemscope itemtype="http://data-vocabulary.org/Person">
      Ονομάζομαι <span itemprop="name">Αλέξανδρος Νίκας</span> και η
      προσωπική μου ιστοσελίδα είναι:
      <a href="http://www.alexnikas.gr" itemprop="url">Alexandros Nikas
      - Portfolio</a>
      Ζω στην
      <span itemprop="address" itemscope itemtype="http://data-
      vocabulary.org/Address">
            <span itemprop="locality">Αθήνα</span>,
            <span itemprop="country-name">Ελλάδα</span>
      </span>
      και σπουδάζω <span itemprop="title">ηλεκτρολόγος μηχανικός &
      μηχανικός υπολογιστών</span> στο <span itemprop="affiliation">
      NTUA</span>.
</div>
```

Κατά τη διαδικασία της βελτιστοποίησης, θα πρέπει να ληφθεί υπόψη και να αξιοποιηθεί αυτή η καινούρια δυνατότητα που προσφέρεται από τον Σημασιολογικό Ιστό και τις μηχανές αναζήτησης.

6.2.2 SEO και Σημασιολογικός Ιστός

Ο Σημασιολογικός Ιστός (Semantic Web ή Web 3.0) ορίζεται ως ένας ιστός από δεδομένα που μπορούν να υποβληθούν σε επεξεργασία, άμεσα ή έμμεσα, από τις μηχανές (αναζήτησης, εν προκειμένω).

Από τον ορισμό αυτό προκύπτει η πρώτη γενική ιδέα – ιδιότητα του Web 3.0 που θα καθορίσει τη λειτουργία των μηχανών αναζήτησης και, φυσικά, τις τεχνικές βελτιστοποίησης

των ιστοσελίδων για τα αποτελέσματα αναζήτησης. Το συγκεκριμένο μοντέλο περιλαμβάνει όλα τα είδη των πληροφοριών που δεν είναι, προς το παρόν, διαθέσιμα στο Web 2.0 (όπως, για παράδειγμα, έγγραφα σε οποιαδήποτε μορφή, πληροφορίες αποκλειστικά εντός βάσεων δεδομένων). Σε αυτές τις κατηγορίες των πληροφοριών προφανώς ανήκουν και οι ιστοσελίδες.

Η δεύτερη ιδιότητα του σημασιολογικού ιστού είναι αυτή της εξάρτησης από δομημένα δεδομένα (structured data). Όπως αναλύθηκε προηγουμένως, η μορφή RDF (RDFa, microformats, microdata) που βασίζεται στη γλώσσα XML αποτελεί τον πυρήνα των δομημένων δεδομένων και επιτρέπει τον επίσημο ορισμό αντικειμένων και σχέσεων στο Διαδίκτυο, όπως φάνηκε και από τα παραδείγματα της προηγούμενης παραγράφου (σε microdata). Το τρίπολο που περιγράφει η RDF (Resource Definition Framework) είναι αυτό του υποκειμένου – κατηγορήματος – αντικειμένου και αποτελεί έναν επίσημο κι εύχρηστο τρόπο περιγραφής *οποιασδήποτε* σχέσης μεταξύ υποκειμένου κι αντικειμένου. Καθώς ήδη δίνεται η νέα δυνατότητα στους διαχειριστές ιστοσελίδων να εμπλουτίζουν το περιεχόμενο με επιπλέον περιγραφές, μέσω των τριών τύπων που πλέον οι μηχανές υποστηρίζουν, αυτό που απομένει για την πλήρη μετάβαση στο Web 3.0 είναι η δυνατότητα διασύνδεσης όλων των δεδομένων (οντολογιών) με τρόπους που έχουν νόημα (semantics).

Η ιδιότητα αυτή της εξάρτησης από συνδεδεμένα μεταξύ τους δεδομένα (linked data) αποτελεί την τρίτη γενική ιδέα πίσω από το semantic web. Σε ένα τέτοιο μοντέλο, κάθε δεδομένο ή σύνολο δεδομένων μπορεί να ανιχνευθεί και να ευρετηριασθεί και, στη συνέχεια, να εντοπισθεί, να παρέχει χρήσιμες πληροφορίες αλλά και χρήσιμες συνδέσεις προς άλλα σύνολα δεδομένων. Η διασύνδεση αυτή που χαρακτηρίζει το τρίπολο των δεδομένων μπορεί θεωρητικά να βοηθήσει στην αποσαφήνιση δεδομένων με πολλαπλές ερμηνείες (γρύλλος, κλειδί, μηχανικός, φέτα), όπως, για παράδειγμα, στη βιομηχανία του SEO, στην αποσαφήνιση ιστοσελίδων με φαινομενικά ίδιο περιεχόμενο κι επί της ουσίας διαφορετικό σκοπό (διπλότυπο περιεχόμενο), γεγονός που θα έπαυε τη χρησιμοποίηση της μεθόδου κανονικοποίησης, ή το φαινόμενο του κανιβαλισμού των λέξεων – κλειδιών που παρατηρείται συχνά εντός των (βελτιστοποιημένων) ιστοτόπων.

Με τη μετάβαση, λοιπόν, αυτή από το Web 2.0 σήμερα στο Semantic Web αύριο, η διαδικασία της βελτιστοποίησης παύει να αφορά στις τεχνικές που εφαρμόζονται σε ιστοσελίδες με σκοπό τη βελτίωση της θέσης τους στα αποτελέσματα και πλέον μετατρέπεται στο σύνολο των τεχνικών που εφαρμόζονται σε προσβάσιμα δομημένα και διασυνδεδεμένα δεδομένα, μεταξύ των οποίων και ιστοσελίδες, με σκοπό την αύξηση της ορατότητας των δεδομένων αυτών στις μηχανές αναζήτησης. Η διαφοροποίηση αυτή θα σημάνει και τη μετάβαση του Search Engine Optimization σε ένα ευρύτερο πλαίσιο, το **Semantic SEO**. Προς αυτή την κατεύθυνση κινείται η πλειοψηφία των μηχανών με την ενσωμάτωση των rich

snippets, όπως και προς την ίδια κατεύθυνση κινήθηκε η Google από το 2003, με την τροποποίηση του αλγορίθμου PageRank και το συνδυασμό του με τον αλγόριθμο Hilltop (που επρόκειτο για μία προσπάθεια καθορισμού σχέσεων μεταξύ συνδέσμων), όπως εξηγήθηκε αναλυτικά στο 5^{ο} κεφάλαιο. Άλλωστε, και σε θεωρητικό επίπεδο, σημασία στη νέα πραγματικότητα δε θα έχουν τα δεδομένα για τα δεδομένα (meta data) αλλά τα δεδομένα καθαυτά (data), και σε αυτά θα πρέπει να εστιάσουν οι στρατηγικές βελτιστοποίησης στο άμεσο μέλλον και όχι στην προσπάθεια ανάδειξης των πρώτων.

7 *Βιβλιογραφία*

Apache. (2004). *Apache Module mod_rewrite.* Ανακτήθηκε από http://httpd.apache.org/docs/current/mod/mod_rewrite.html, [30/04/2011]

Baeza-yates, R., & Castillo, C. (Συγγρ.). (2002). Balancing Volume, Quality and Freshness in Web Crawling. *IN SOFT COMPUTING SYSTEMS - DESIGN, MANAGEMENT AND APPLICATIONS*, 565--572.

Baeza-Yates, R., Castillo, C., Marin, M., & Rodriguez, A. (Συγγρ.). (2005). Crawling a country: better strategies than breadth-first for web page ordering. *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05 (σσ 864–872). New York, NY, USA: ACM. doi:10.1145/1062745.1062768

Bharat, K. & Mihaila, G.A.. (1999). Hilltop: A Searh Engine based on Expert Documents. Ανακτήθηκε από http://ftp.cs.toronto.edu/pub/reports/csrg/405/hilltop.html

Bifet, A., Castillo, C., Chirita, P. A., & Weber, I. (Συγγρ.). (2005). An Analysis of Factors Used in Search Engine Ranking. Ανακτήθηκε από http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.5043

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (Συγγρ.). (2004). UbiCrawler: a scalable fully distributed Web crawler. *Software: Practice and Experience*, *34*(8), 711-726. doi:10.1002/spe.587

Brin, S., & Page, L. (Συγγρ.). (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Ανακτήθηκε από http://ilpubs.stanford.edu:8090/361/

Castillo, C. (Συγγρ.). (2005). Effective web crawling. *SIGIR Forum*, *39*(1), 55–56. doi:10.1145/1067268.1067287

Cho, J., & Garcia-Molina, H. (Συγγρ.). (2002). Parallel crawlers. *Proceedings of the 11th international conference on World Wide Web*, WWW '02 (σσ 124–135). New York, NY, USA: ACM. doi:10.1145/511446.511464

Cho, J., & Garcia-Molina, H. (Συγγρ.). (2003). Effective page refresh policies for Web crawlers. *ACM Trans. Database Syst.*, *28*(4), 390–426. doi:10.1145/958942.958945

Cho, J., Garcia-Molina, H., & Page, L. (Συγγρ.). (1998). Efficient Crawling Through URL Ordering. Ανακτήθηκε από http://ilpubs.stanford.edu:8090/347/

Coffman Jr., E. G., Liu, Z., & Weber, R. R. (Συγγρ.). (1998). Optimal robot scheduling for Web search engines. *Journal of Scheduling*, *1*(1), 15-29. doi:10.1002/(SICI)1099-1425(199806)1:1<15::AID-JOS3>3.0.CO;2-K

Cothey, V. (Συγγρ.). (2004). Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, *55*(14), 1228-1238. doi:10.1002/asi.20078

Davide Musella, D. M. (Συγγρ.). (1995). The META Tag of HTML. Ανακτήθηκε Νοέμβριος 1, 2011, από http://tools.ietf.org/html/draft-musella-html-metatag-01

Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., & Gori, M. (Συγγρ.). (2000). Focused crawling using context graphs. *IN 26TH INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, VLDB 2000*, 527--534.

Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (Συγγρ.). (2002). Self-similarity in the web. *ACM Trans. Internet Technol.*, *2*(3), 205–223. doi:10.1145/572326.572328

Edwards, J., McCurley, K., & Tomlin, J. (Συγγρ.). (2001). An adaptive model for optimizing performance of an incremental web crawler. *Proceedings of the 10th international conference on World Wide Web*, WWW '01 (σσ 106–113). New York, NY, USA: ACM. doi:10.1145/371920.371960

Google Blog. (2007). *Robots Exclusion Protocol: now with even more flexibility.* Ανακτήθηκε από http://googleblog.blogspot.com/2007/07/robots-exclusion-protocol-now-with-even.html, [21/04/2011]

Google Blog. (2008). *Google learns to crawl Flash.* Ανακτήθηκε από http://googleblog.blogspot.com/2008/06/google-learns-to-crawl-flash.html, [06/05/2011]

Google Webmaster Central Blog. (2009). *Google does not use the keywords meta tag in web ranking.* Ανακτήθηκε από http://googlewebmastercentral.blogspot.com/2009/09/google-does-not-use-keywords-meta-tag.html, [21/04/2011]

Google. (1998). *Google searches more sites more quickly, delivering the most relevant results.* Ανακτήθηκε από http://www.google.com/intl/en_uk/technology/index.html, [20/04/2011]

Gupta, V., Gomes, B., Lamping, J., McGrath, M., Singhal, A., & Tong, S. (Συγγρ.). (2003). System and method for providing preferred country biasing of search results. Ανακτήθηκε από http://www.google.com/patents/about?id=aoifAAAAEBAJ

Haveliwala, T., Kamvar, S., Kamvar, A., & Jeh, G. (Συγγρ.). (2003). An Analytical Comparison of Approaches to Personalizing PageRank. Ανακτήθηκε από http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.5438

Ipeirotis, P. G., Ntoulas, A., Cho, J., & Gravano, L. (Συγγρ.). (2005). Modeling and managing content changes in text databases. *21st International Conference on Data Engineering, 2005. ICDE 2005. Proceedings* (σσ 606- 617). Παρουσιάστηκε στο 21st International Conference on Data Engineering, 2005. ICDE 2005. Proceedings, IEEE. doi:10.1109/ICDE.2005.91

Kleinberg, J. M. (Συγγρ.). (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, *46*(5), 604–632. doi:10.1145/324133.324140

Langville, A. N., Meyer, C. D., & FernÁndez, P. (Συγγρ.). (2006). Google's PageRank and beyond: The science of search engine rankings. *The Mathematical Intelligencer*, *30*, 68-69. doi:10.1007/BF02985759

Manning, C., Raghavan, P., & Schütze, H. (Συγγρ.). (2008). Introduction to Information Retrieval. Cambridge University Press. Ανακτήθηκε από http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0521865719

Matt Cutts: Gadgets, Google, and SEO. (2005). *Text links and PageRank*. Ανακτήθηκε από http://www.mattcutts.com/blog/text-links-and-PageRank/, [22/09/2011]

Menczer, F., Pant, G., & Srinivasan, P. (Συγγρ.). (2004). Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.*, *4*(4), 378–419. doi:10.1145/1031114.1031117

Mi Islita. (2005). *The Keyword Density of Non-Sense.* Ανακτήθηκε από http://www.miislita.com/fractals/keyword-density-optimization.html, [21/04/2011]

Microsoft. (2004). *URL Rewriting in ASP.NET.* Ανακτήθηκε από http://msdn.microsoft.com/en-us/library/ms972974.aspx, [30/04/2011]

Motwani, R., & Raghavan, P. (Συγγρ.). (2010). Algorithms and theory of computation handbook. Στο M. J. Atallah & M. Blanton (Επιμ.), (σσ 12–12). Chapman & Hall/CRC. Ανακτήθηκε από http://dl.acm.org/citation.cfm?id=1882757.1882769

Najork, M., & Wiener, J. L. (Συγγρ.). (2001). Breadth-first crawling yields high-quality pages. *Proceedings of the 10th international conference on World Wide Web*, WWW '01 (σσ 114–118). New York, NY, USA: ACM. doi:10.1145/371920.371965

Page, L., Brin, S., Motwani, R., & Winograd, T. (Συγγρ.). (1999). The PageRank Citation Ranking: Bringing Order to the Web. Ανακτήθηκε από http://ilpubs.stanford.edu:8090/422/

Pant, G., Srinivasan, P., & Menczer, F. (Συγγρ.). (2004). Crawling the Web. *IN WEB DYNAMICS: ADAPTING TO CHANGE IN CONTENT, SIZE, TOPOLOGY AND USE. EDITED BY M. LEVENE AND A. POULOVASSILIS*, 153--178.

Pinkerton, B. (Συγγρ.). (2000). WebCrawler: Finding What People Want. Ανακτήθηκε από http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.890

Pringle, G., Allison, L., & Dowe, D. L. (Συγγρ.). (1998). What is a tall poppy among Web pages? *Proceedings of the seventh international conference on World Wide Web 7*, WWW7 (σσ 369–377). Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V. Ανακτήθηκε από http://dl.acm.org/citation.cfm?id=297805.297862

Search Engine Land. (2007). *Meta Keywords Tag 101: How to "Legally" Hide Words On Your Pages For Search Engines.* Ανακτήθηκε από http://searchengineland.com/meta-keywords-tag-101-how-to-legally-hide-words-on-your-pages-for-search-engines-12099, [25/04/2011]

Shi J. (2007). Linear Algebra behind Google PageRank. *Stanford University*.

Shkapenyuk, V., & Suel, T. (Συγγρ.). (2002). Design and implementation of a high-performance distributed Web crawler. *18th International Conference on Data Engineering, 2002. Proceedings* (σσ 357-368). Παρουσιάστηκε στο 18th International Conference on Data Engineering, 2002. Proceedings, IEEE. doi:10.1109/ICDE.2002.994750

Sisson, D. (2006). *Google SEO Secrets*. Redmond, WA: Blue Moose Webworks.

W3C. (2000). *HTML Techniques for Web Content Accessibility Guidelines 1.0.* Ανακτήθηκε από http://www.w3.org/TR/WCAG10-HTML-TECHS/, [22/04/2011]

W3C. (2010). *Techniques for WCAG 2.0.* Ανακτήθηκε από http://www.w3.org/TR/WCAG20-TECHS/#F41, [22/04/2011]

Zeller, T. (2006). A New Campaign Tactic: Manipulating Google data, The New York Times, October 26, 2006.

Παράρτημα Α Έρευνα κι ανάλυση των λέξεων - κλειδιών

Α.1 Τι να προτιμήσουμε, λέξη ή φράση;

Πολλοί χρήστες αλλά και web marketers πιστεύουν ότι είναι προτιμότερο να στοχεύουν σε μεμονωμένες λέξεις, καθώς αυτές ειναι πολύ πιο πιθανό να αποτελέσουν μέρος μίας φράσης που ένας χρήστης θα αναζητήσει. Αυτή η πεποίθηση έχει αποδειχθεί λάθος, αφενός γιατί οι μηχανές αναζήτησης δίνουν πολύ μεγαλύτερη βαρύτητα στους ακριβείς όρους – κλειδιά, προωθώντας σελίδες που φαίνονται να απαντούν σε αυτούς τους ακριβείς όρους, παρά σε σελίδες που σχετίζονται περισσότερο με μέρος της φράσης αναζήτησης, αφετέρου γιατί οι χρήστες αποδεδειγμένα αναζητούν για περισσότερες από μία λέξεις, όταν χρησιμοποιούν τις μηχανές αναζήτησης. Συγκεκριμένα, έρευνα της OneStat, τον Ιούλιο του 2006, έδειξε τα εξής αποτελέσματα:

July 2006

1.	2 word phrases	28.91%
2.	3 word phrase	27.85%
3.	4 word phrases	17.11%
4.	1 word phrases	11.43%
5.	5 word phrases	8.25%
6.	6 word phrases	3.68%
7.	7 word phrases	1.59%

Εικόνα 35 Έρευνα της OneStat για τον αριθμό των λέξεων των ερωτημάτων

Δηλαδή, η πλειοψηφία των αναζητήσεων αφορούσε φράσεις των δύο και τριών λέξεων, ενώ μόλις μία στις δέκα αναζητήσεις αφορούσαν μία μόνο λέξη.

Α.2 Κατανοώντας τη μακροσκελή ουρά της διαδικτυακής αναζήτησης

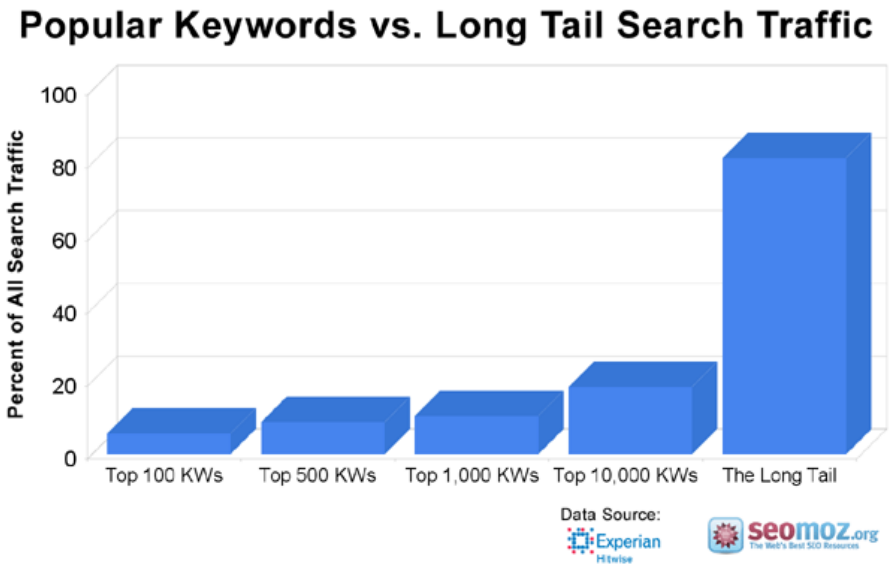
Είναι σπουδαίο κι αποτελεί πρόκληση να ασχολούμαστε με συνήθεις, καθημερινές, γενικές λέξεις – κλειδιά που έχουν χιλιάδες αναζητήσεις κάθε μέρα, ή ακόμη κι εκατοντάδες, αλλά, στην πραγματικότητα, αυτοί οι δημοφιλείς όροι αναζήτησης διαμορφώνουν συνολικά λιγότερο από το 30% των συνολικών αναζητήσεων που εκτελούνται στο Διαδίκτυο.

Το υπόλοιπο 70% βρίσκεται επάνω στην αποκαλούμενη μακροσκελή ουρά (long tail) των αναζητήσεων. Το long tail περιλαμβάνει εκατοντάδες εκατομμύρια μοναδικών αναζητήσεων που ενδέχεται να πραγματοποιηθούν ελάχιστες φορές μέσα σε μία μέρα αλλά που, εάν ενωθούν όλες μαζί, συναποτελούν την συντριπτική πλειοψηφία της παγκόσμιας αναζήτησης πληροφοριών μέσα από τις μηχανές αναζήτησης.

Αυτό, φυσικά, δε σημαίνει ότι μία ιστοσελίδα που καταλαμβάνει πολύ καλές θέσεις στις μηχανές αναζήτησης για πολύ γενικές (και δύσκολες) λέξεις δε θα έχει πολύ μεγαλύτερη κίνηση από μία άλλη που προσελκύει κίνηση για πολύ ειδικευμένους όρους, αλλά ότι ακόμη και αυτή η σελίδα δε θα εμφανιζόταν στα αποτελέσματα για τη συντριπτική πλειοψηφία των αναζητήσεων. Άλλωστε, τα ποσοστά αποδοτικότητας των όρων αναζήτησης είναι καλύτερα όταν πρόκειται για ιδιαίτερα εξειδικευμένους ή σπάνιους όρους, αρκεί να αναλογιστεί κανείς ότι μία σελίδα που βρίσκεται στις πρώτες θέσεις για το ερώτημα «κομμωτήρια» δύσκολα θα εξυπηρετεί το χρήστη, σε τοπικό επίπεδο, κι ακόμη πιο δύσκολα θα τον προσελκύσει, δεδομένου του τεράστιου όγκου των αποτελεσμάτων που θα έχει στη διάθεσή του, ενώ, αντίθετα, μία σελίδα που βρίσκεται στην πρώτη θέση για το ερώτημα «κομμωτήρια γλυφάδα» σίγουρα θα τραβήξει το ενδιαφέρον του χρήστη, θα τον εξυπηρετεί (κατά πάσα πιθανότητα) και, ως εκ τούτου, θα οδηγήσει σε όφελος (οικονομικό) του συγκεκριμένου κομμωτηρίου. Αντίστοιχα, μία αναζήτηση για τη φράση «music sharing social network» θα εξυπηρετήσει τον επισκέπτη του myspace.com, ενώ ο ίδιος χρήστης, στο ερώτημα «social network» δε θα έπαιρνε τη ζητούμενη απάντηση, μπαίνοντας στο πρωτοπόρο facebook.

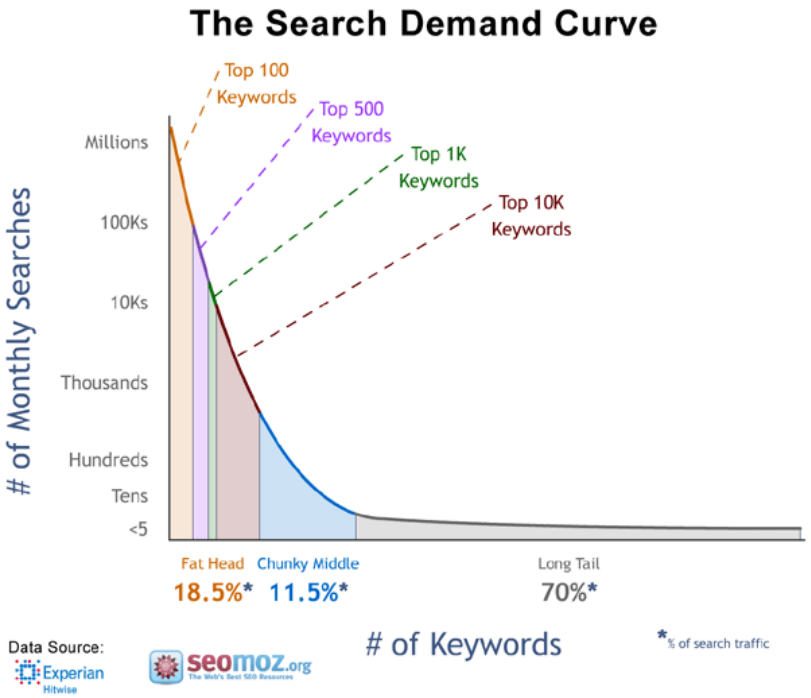
Εάν, λοιπόν, η διαδικτυακή αναζήτηση μπορούσε να παρασταθεί από μία μικροσκοπική σαύρα με κεφάλι διαμέτρου 2,5 εκατοστών, η ουρά αυτής της σαύρας θα είχε μήκος 221 μίλια.

Η σημασία της μακροσκελούς ουράς φαίνεται καλύτερα στα διαγράμματα που ακολουθούν:



Εικόνα 36 Η επισκεψιμότητα συναρτήσει της διασημότητας των λέξεων – κλειδιών

Πηγή: hitwise.com



Εικόνα 37 Οι δημοφιλέστερες λέξεις – κλειδιά ως προς το σύνολο των αναζητήσεων

Πηγή: SEOMoz.org

Τα διαγράμματα δείχνουν ότι ακόμη κι αν μονοπωλούμε τις δημοφιλέστερες 1000 λέξεις (που είναι αδιανόητο), πάλι θα χάναμε το 90% των αναζητήσεων.

Γι’ αυτό, φροντίζουμε να μην εξειδικεύουμε τις λέξεις – κλειδιά στις οποίες στοχεύουμε, αλλά να επιτρέπουμε στην σελίδα μας να αναδεικνύει και να εκθέτει το μεγαλείο της ανθρώπινης σκέψης, έρευνας και άποψης στους ανιχνευτές των μηχανών αναζήτησης, πετυχαίνοντας καλύτερα αποτελέσματα.

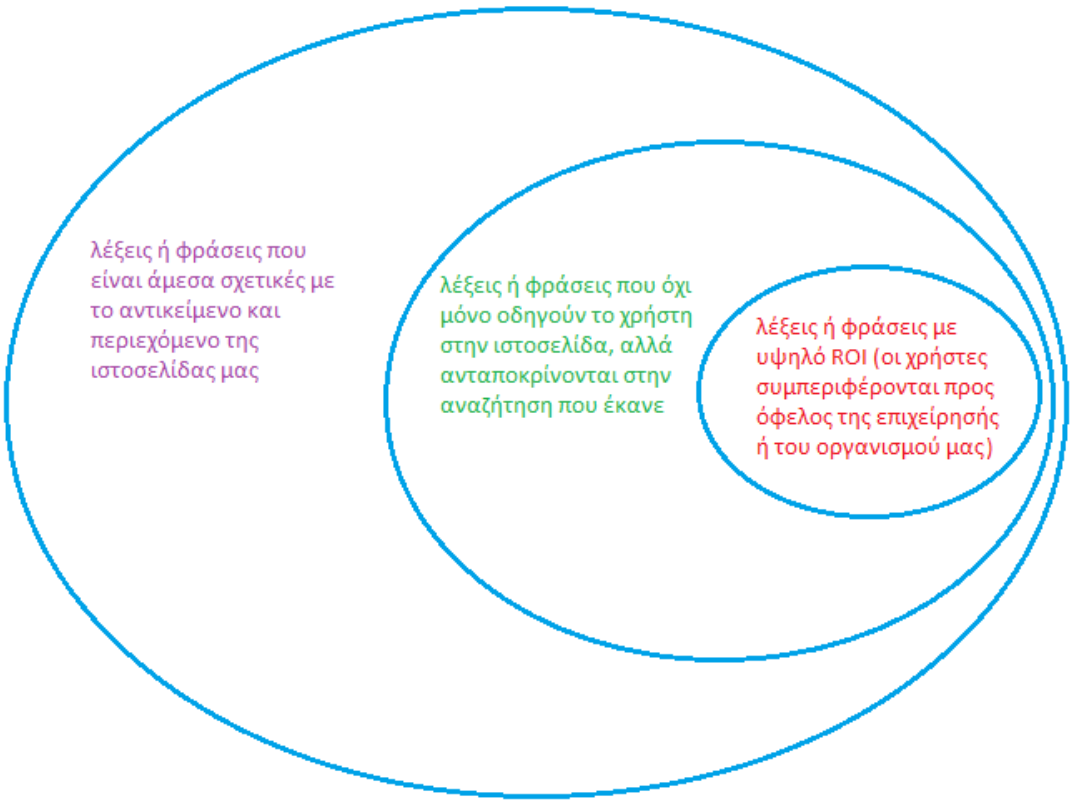
Ακολουθούν τα 9 βήματα που συνίσταται να ακολουθούνται για τη δομημένη έρευνα κι ανάλυση των λέξεων και φράσεων για τις οποίες μία επιχείρηση (ή οργανισμός) στοχεύει να βελτιστοποιήσει την ιστοσελίδα της.

Α.3 Η διαδικασία της έρευνας και ανάλυσης των λέξεων – κλειδιών

Α.3.1 Σχεδιασμός

Α.3.1.1 Προσδιορισμός των επιθυμητών όρων – κλειδιών

Αρχικά, προσδιορίζονται οι όροι εκείνοι που είναι άμεσα σχετικοί με το περιεχόμενο της υπό εξέταση ιστοσελίδας. Στη συνέχεια, απομονώνουμε εκείνες τις λέξεις ή φράσεις που, δεδομένης της συνάφειας με το αντικείμενο της ιστοσελίδας και υποθέτοντας ότι αυτή εμφανίζεται στα αποτελέσματα αναζήτησης αυτών των όρων, θα προσελκύσουν τους χρήστες αλλά και θα ανταποκρίνονται πλήρως στην αναζήτηση που πραγματοποίησαν. Για παράδειγμα, μία σελίδα που σχετίζεται αποκλειστικά με τουριστικές εκδρομές στη Βαρκελώνη κι εμφανιστεί στα αποτελέσματα της αναζήτησης «Βαρκελώνη», είναι πολύ πιθανό να μην ανταποκρίνεται στα κριτήρια του χρήστη. Τέλος, ακολουθεί η επιλογή εκείνων των λέξεων ή φράσεων που θα αποφέρουν τα μεγαλύτερα κέρδη για την επιχείρηση ή τον οργανισμό, είτε ως κέρδος μεταφράζεται η εγγραφή σε κάποια κοινότητα ή newsletter, είτε η αγορά μίας υπηρεσίας ή προϊόντος, ή η δωρεά σε κάποιο φιλανθρωπικό ίδρυμα.



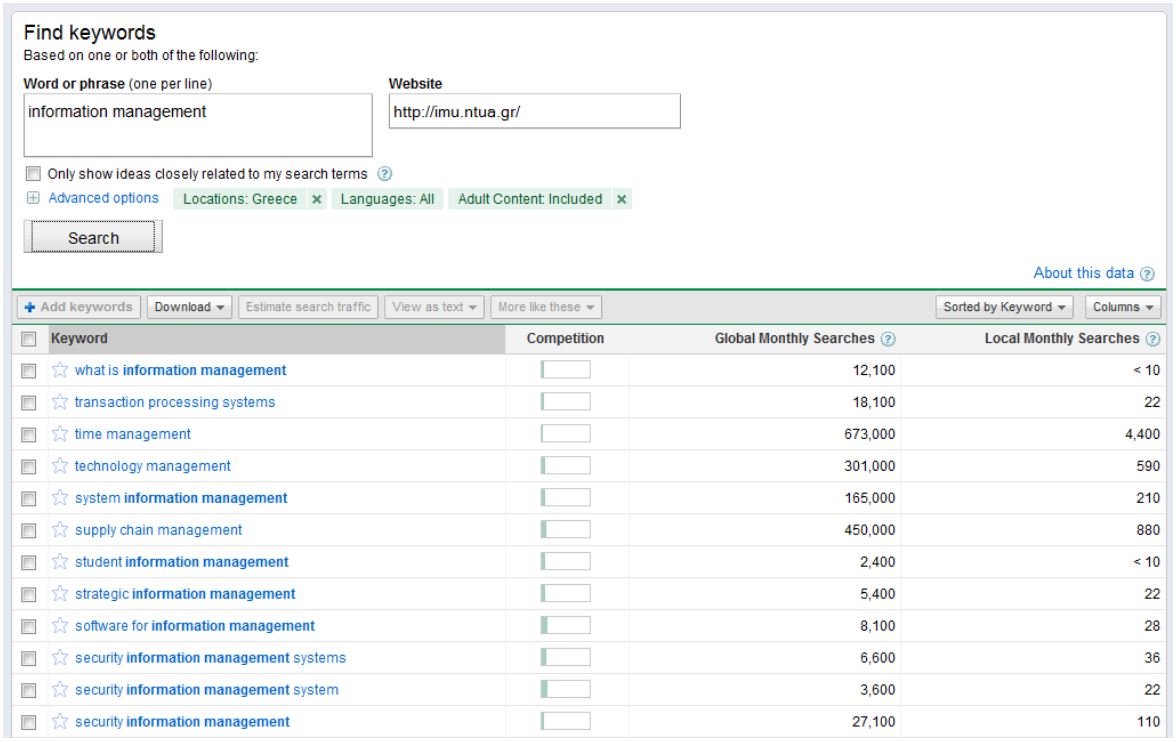
Εικόνα 38 Προσδιορισμός των επιθυμητών λέξεων ή φράσεων - κλειδιών

Α.3.1.2 Σχετικός όγκος αναζήτησης

Μετά τον προσδιορισμό των κατάλληλων φράσεων ή λέξεων – κλειδιών, συνίσταται η εκτίμηση του όγκου αναζήτησης και του σχετικού ανταγωνισμού για κάθε έναν επιθυμητό όρο. Η διαδικασία αυτή προσανατολίζει τους διαχειριστές του ιστοτόπου ποιοι όροι οφείλουν να βαραίνουν περισσότερο τη διαδικασία της βελτιστοποίησης, καθώς αυτή μπορεί να εξατομικευτεί για κάθε λέξη ή φράση. Τα πλέον χρήσιμα εργαλεία για τη διαδικασία αυτή είναι τα εξής:

1. Google AdWords Keyword Tool

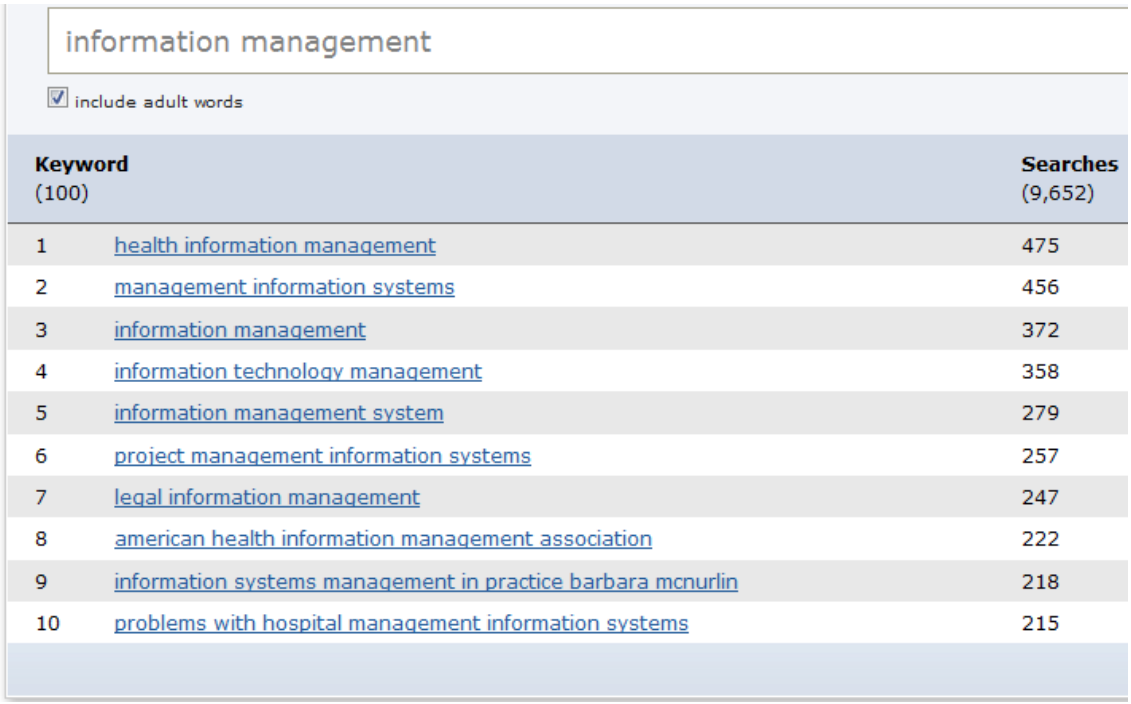
Το εργαλείο αυτό της Google παρουσιάζει στοιχεία σχετικά με τον ανταγωνισμό των σχετικών φράσεων, τις μηνιαίες αναζητήσεις παγκοσμίως αλλά κι εγχώρια, όσον αφορά αποκλειστικά τη μηχανή αναζήτησης της Google.



Εικόνα 39 Google AdWords Keyword Tool

2. WordTracker Keyword Tool

Αντίθετα με το AdWords Keyword Tool, το συγκεκριμένο εργαλείο παρέχει τις παγκόσμιες αναζητήσεις σχετικών φράσεων στις μηχανές αναζήτησης της InfoSpace (WebCrawler, MetaCrawler, Dogpile και WebFetch).

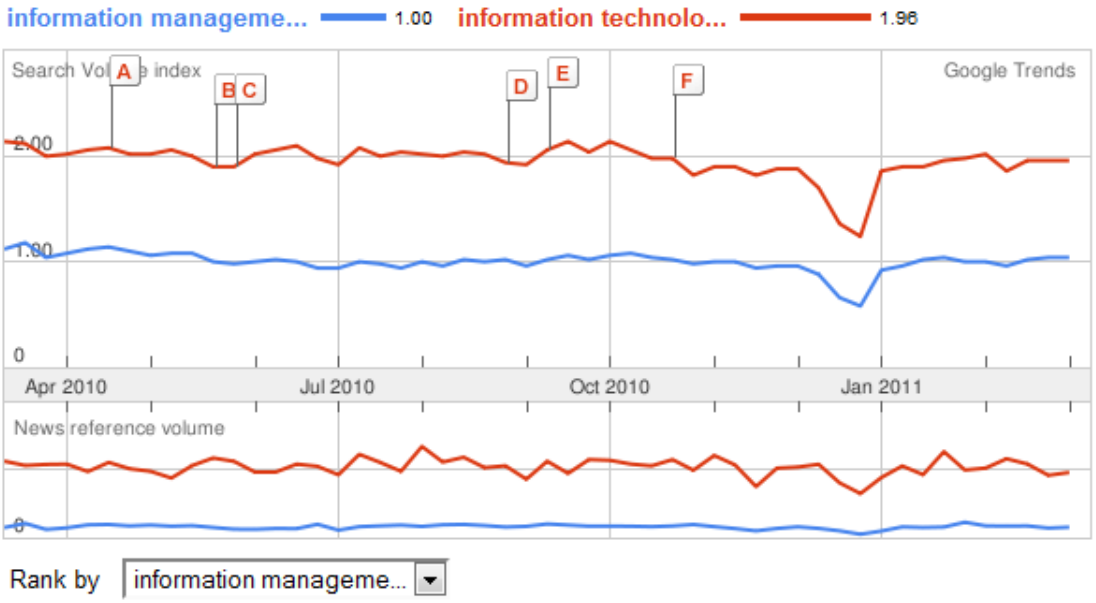


Εικόνα 40 WordTracker Keyword Tool

A.3.1.3 Τάση και χρονικές διακυμάνσεις

Έχει μεγάλη σημασία να προσδιορίσουμε πότε οι λέξεις μας έχουν τη μεγαλύτερη ζήτηση, πότε παύουν να χρησιμοποιούνται, όχι μόνο για να προβλέπουμε τη συνέχεια της πορείας τους, αλλά και για να αναλύσουμε την ίδια την αγορά για τα προϊόντα ή τις υπηρεσίες (π.χ. εάν η ζήτησή τους πέφτει σταδιακά τα τελευταία χρόνια ή μήνες, ίσως μία εταιρία χρειάζεται να επαναπροσδιορίσει τις λέξεις – κλειδιά ή ακόμη και τα ίδια τα προϊόντα της).

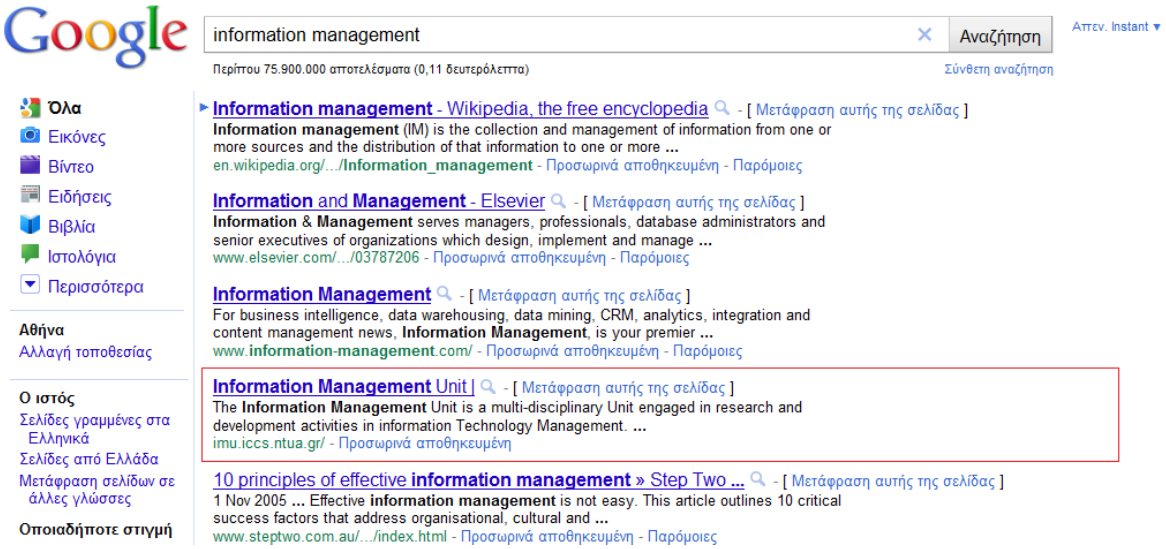
Παράδειγμα τέτοιου εργαλείου είναι το Google Trends, το οποίο παρουσιάζει τις διακυμάνσεις στην αναζήτηση μίας ή περισσότερων φράσεων – κλειδιών στη μηχανή της Google:



Εικόνα 41 Google Trends και τάση λέξεων - κλειδιών

A.3.1.4 Ανάλυση ανταγωνισμού

Προκειμένου να κατανοήσουμε σε μεγάλο βαθμό το ανταγωνιστικό τοπίο, είναι απαραίτητο να γνωρίζουμε ποιοι ανταγωνιστές μας πετυχαίνουν τις υψηλότερες θέσεις στα αποτελέσματα αναζήτησης (για τους επιθυμητούς όρους – κλειδιά). Πρόκειται για μία διαδικασία ανάλυσης του ανταγωνισμού, με κριτήριο τη θέση των ανταγωνιστών στα αποτελέσματα αναζήτησης και όχι τα οικονομικά στοιχεία της αγοράς και μπορεί να πραγματοποιηθεί με μία απλή αναζήτηση στις μηχανές



Εικόνα 42 Ανάλυση ανταγωνισμού στις μηχανές αναζήτησης

A.3.1.5 Ανάλυση δραστηριότητας στα Social Networks και Media

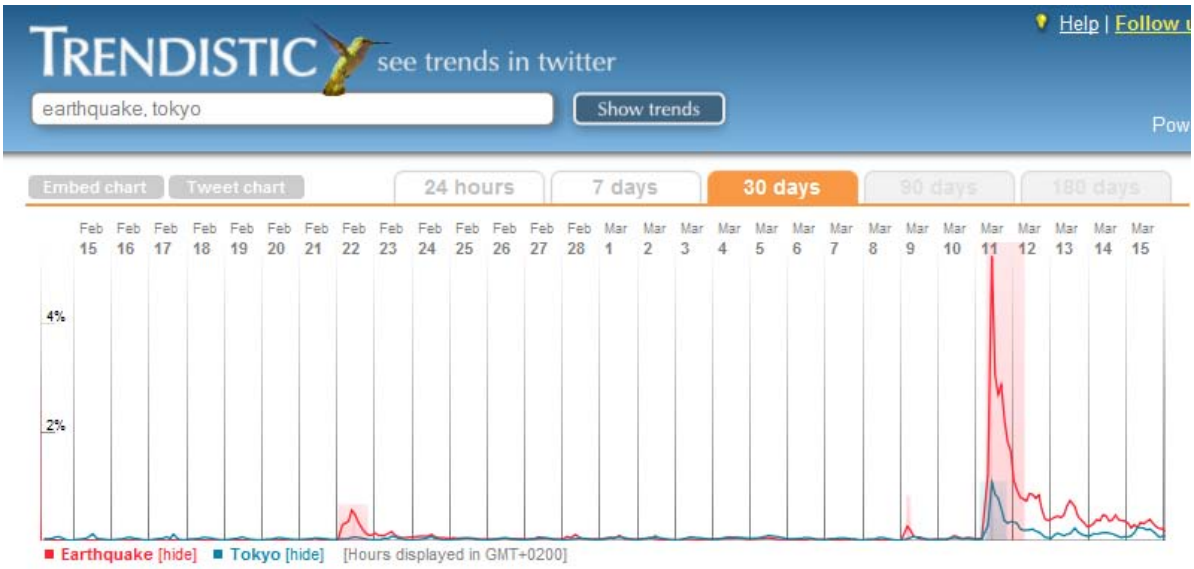
Η μελέτη της δραστηριότητας της σφαίρας των μέσων κοινωνικής δικτύωσης (social media & networks) είναι καθοριστική για τον προσδιορισμό της συχνότητας με την οποία αναφέρονται οι επιθυμητές λέξεις ή φράσεις κλειδιά στο περιεχόμενο και τις συνομιλίες σε ιστολόγια, forums, feeds και τα πολυσύχναστα μέσα δικτύωσης (facebook, twitter κλπ).

Η διαδικασία αυτή είναι ιδιαίτερα πολύτιμη για να αναγνωρίζουμε, έγκαιρα, αφενός τις αναδυόμενες τάσεις (κυρίως όσον αφορά τη νεολαία), αφετέρου την επικείμενη δυσκολία που θα αντιμετωπίσουμε στη βελτιστοποίηση της ιστοσελίδα μας στα αποτελέσματα αναζήτησης για λέξεις – κλειδιά που, πλέον, χρησιμοποιούνται κατά κόρον από σελίδες – κολοσσούς.

Τέτοια εργαλεία χρήσιμα για τη διαδικασία της ανάλυσης της δραστηριότητας στα κοινωνικά μέσα δικτύωσης είναι τα παρακάτω:

1. Trendistic:

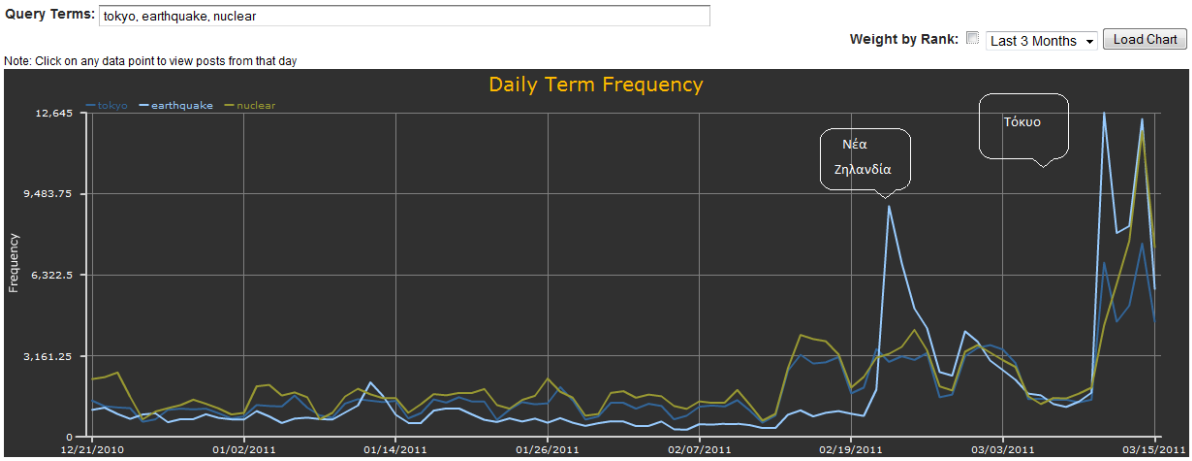
Το εργαλείο αυτό δείχνει τις τάσεις για μία ή περισσότερες λέξεις – κλειδιά ανά τους χρήστες του Twitter:



Εικόνα 43 Trendistic και δραστηριότητα στο Twitter

2. SEOMoz BlogScape

Το BlogScape παρατηρεί, για ένα αρκετά μεγάλο διάστημα, τις τάσεις στην παγκόσμια μπλογκόσφαιρα.



Εικόνα 44 Μελέτη των τάσεων της παγκόσμιας μπλογκόσφαιρας

3. Google News Trends

Το News Trends της μηχανής της Google παρατηρεί την τάση σε όλο το δίκτυο περιεχομένου της Google:

Εικόνα 45 Google News Trends

Η πραγματική αξία και ο λόγος για να αφιερώσει κάποιος χρόνο σε αυτούς τους τομείς έγκειται:

- Στην αναγνώριση προτύπων ή τάσεων που υποδεικνύουν ότι μία λέξη – κλειδί ή ένα προϊόν ή μία ιδέα βρίσκονται σε ακμή ή παρακμή
- Στην εύρεση περιεχομένου που, στο παρελθόν, έχει προσελκύσει πολύ μεγάλη προσοχή γύρω από ορισμένες λέξεις – κλειδιά
- Στον προσδιορισμό διαδικτυακών πυλών και διαδικτυακών κοινοτήτων όπου το θέμα με το οποίο ασχολούμαστε ενδέχεται να αποσπάσει πολλή προσοχή (ή όπου υπάρχουν δυνατότητες προώθησης)

Α.3.1.6 Προθέσεις των χρηστών του Διαδικτύου

Είναι πολύ σημαντικό να καταλάβουμε ποιος είναι ο στόχος των χρηστών εκείνων που πραγματοποιούν αναζητήσεις για τις επιθυμητές μας λέξεις – κλειδιά, σε ποιο στάδιο απόφασης βρίσκονται. Με τον τρόπο αυτό, μπορούμε να αποκτήσουμε μία αρκετά καλή ιδέα του δυνητικού ROI από την προσέλκυση κίνησης στη ιστοσελίδα μας, μέσω ενός καθορισμένου όρου αναζήτησης. Πολλές φορές, οι λέξεις – κλειδιά με το μεγαλύτερο όγκο αναζήτησης δεν προσελκύουν την πιο κερδοφόρο κίνηση επισκεπτών.

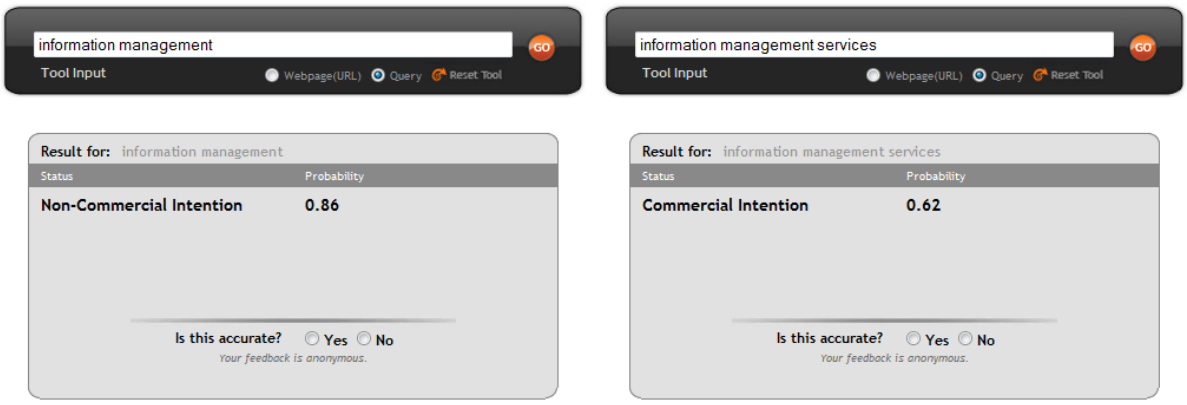
Ένας συνήθης ιστότοπος ενδέχεται να δεχθεί επίσκεψη από ένα χρήστη που πραγματοποιεί αναζήτηση συγκεκριμένου τύπου και περιεχομένου:

- Αναζήτηση πλοήγησης
- Αναζήτηση πληροφοριών
- Έρευνα αγοράς
- Αναζήτηση συναλλαγών

Όρος αναζήτησης	Μηνιαίες Αναζητήσεις	Πρόθεση	Δυνητική (εμπορική) αξία
barcelona airport	1,300	Πλοήγηση	Χαμηλή
barcelona hotels	2,900	Έρευνα Αγοράς	Μέτρια
Πακέτα εκδρομών βαρκελώνη	50	Συναλλαγή	Υψηλή
Συνταγή παέγια	?	Πληροφορίες	Χαμηλή
Sagrada Familia	1,300	Πληροφορίες	Πολύ χαμηλή

Πίνακας 11 Παράδειγμα προσδιορισμού των προθέσεων των χρηστών

Ένα ιδιαίτερα εύχρηστο και ακριβές εργαλείο που βασίζεται στη δραστηριότητα των χρηστών της msn και bing είναι το Microsoft AdCenter Labs Online Commercial Intention Tool:



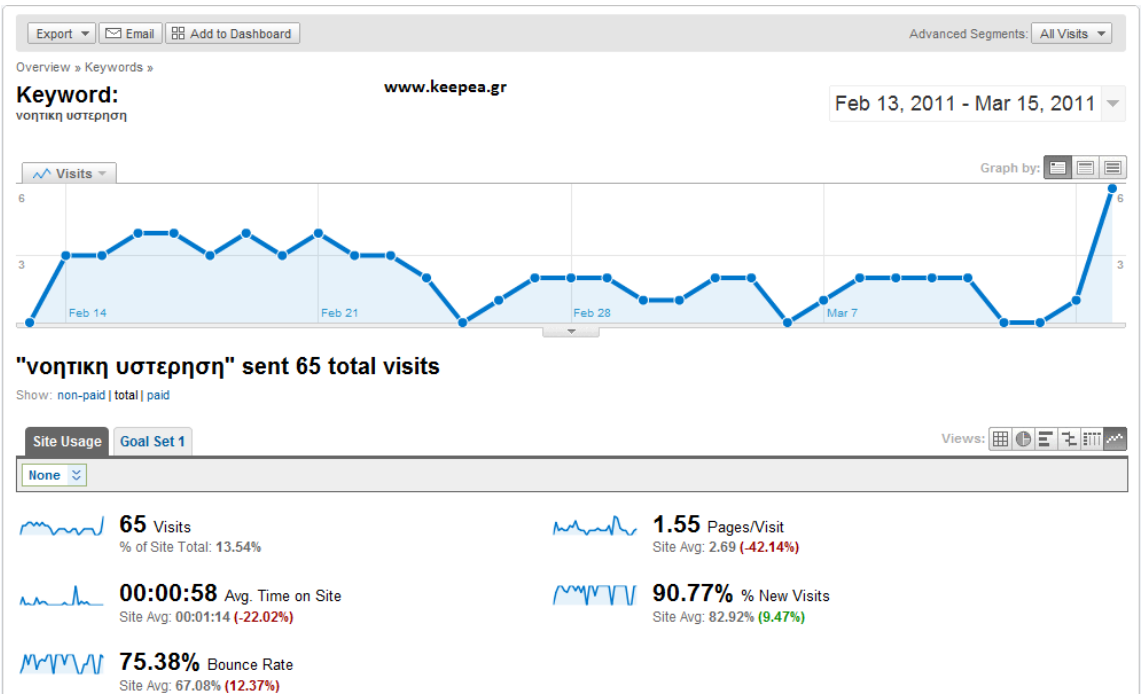
Εικόνα 46 Πρόβλεψη προθέσεων των χρηστών, βάσει ερωτήματος, από τη Microsoft

A.3.2 Ανατροφοδότηση

A.3.2.1 Δυνητική σχέση λέξεων – κλειδιών και περιεχομένου

Ο πλέον κατάλληλος τρόπος για να προσδιορίσουμε κατά πόσο σχετική είναι η λέξη – κλειδί με το περιεχόμενο ή την υπηρεσία ή το προϊόν που παρέχουμε είναι ο έλεγχος των πραγματικών αποτελεσμάτων. Μπορούμε να φιλτράρουμε τις πληροφορίες από τον αντίστοιχο λογαριασμό Google Analytics για την ιστοσελίδα μας και να επιλέξουμε μόνο εκείνους τους χρήστες που επισκέφθηκαν βάσει αυτής της συγκεκριμένης λέξης – κλειδιού, ή

να διεξαγάγουμε μία καμπάνια PPC (από την Google AdWords π.χ.). Δεδομένα όπως ο χρόνος παραμονής, το bounce rate, το conversion rate, η επίτευξη στόχου είναι μερικά από τα οποία, συνήθως, διαφοροποιούνται σε πολύ μεγάλο βαθμό, ανάλογα από τη λέξη κλειδί που προσέλκυσε τους επισκέπτες.



Εικόνα 47 Google Analytics και εποπτεία της αποτελεσματικότητας των keywords

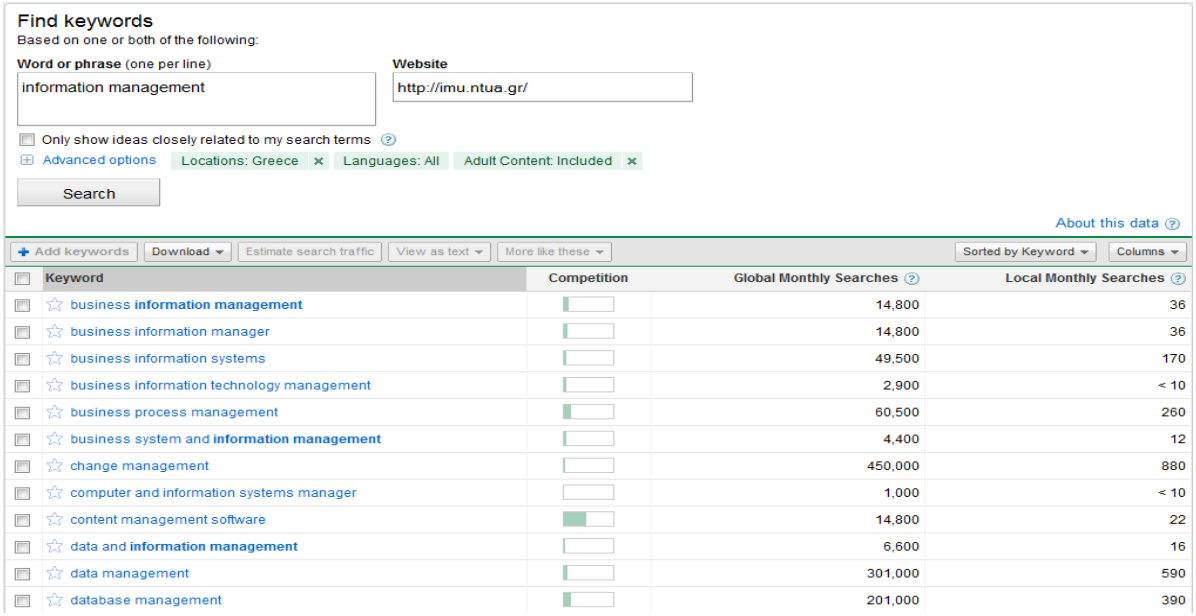
A.3.2.2 Σχετικοί όροι αναζήτησης

Είναι δεδομένο ότι το feedback από την επισκεψιμότητα στην ιστοσελίδα μας, μετά από λίγους μήνες, θα μας υποδείξει ποιες λέξεις ή φράσεις αποδίδουν περισσότερο και ποιες λιγότερο. Οι λιγότερο αποδοτικοί όροι, λόγω δυσκολίας κατάληψης κάποιας υψηλής θέσεις στα αποτελέσματα (SERPs), δε χρειάζεται να απορριφθούν, εφόσον έχουν κριθεί χρήσιμοι ή και κρίσιμοι για την επίτευξη οικονομικών και άλλων στόχων, αλλά να παραλλαχθούν σε άλλους παρόμοιους ή σχετικούς. Τη διαδικασία αυτή διευκολύνουν πολύ οι ίδιες οι μηχανές αναζήτησης με συγκεκριμένα χρήσιμα εργαλεία, καθώς είναι πολύ πιο αρμόδιες να μας υποδείξουν ποιους όρους αναζήτησης χρησιμοποιούν εναλλακτικά οι χρήστες για παρόμοιες ή σχετικές αναζητήσεις. Τέτοια εργαλεία είναι τα εξής:

1. Google Adwords Keyword Tool

Το εργαλείο της Google, όπως παρουσιάσθηκε και προηγουμένως για την ανάλυση του σχετικού όγκου αναζήτησης, μπορεί να εξυπηρετήσει τον σκοπό του εντοπισμού σχετικών

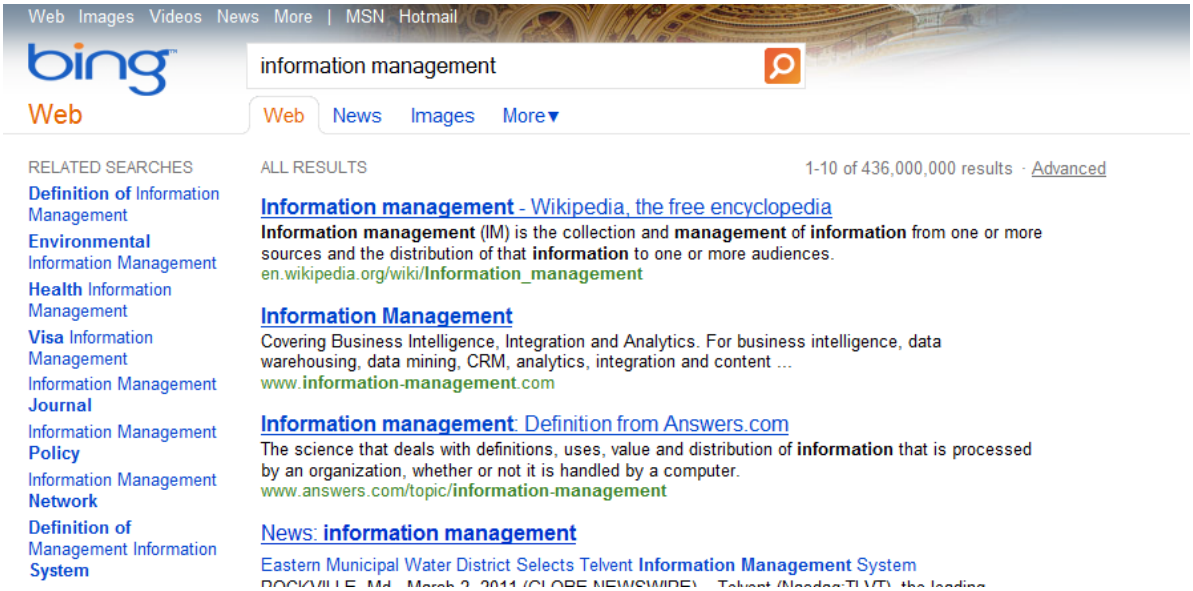
όρων αναζήτησης, εφόσον δεν επιλεχθεί η ρύθμιση της εμφάνισης ιδεών στενά συνδεδεμένων στους όρους αναζήτησης:



Εικόνα 48 Προτάσεις λέξεων και φράσεων προς βελτιστοποίηση από την Google

2. Bing Search Engine

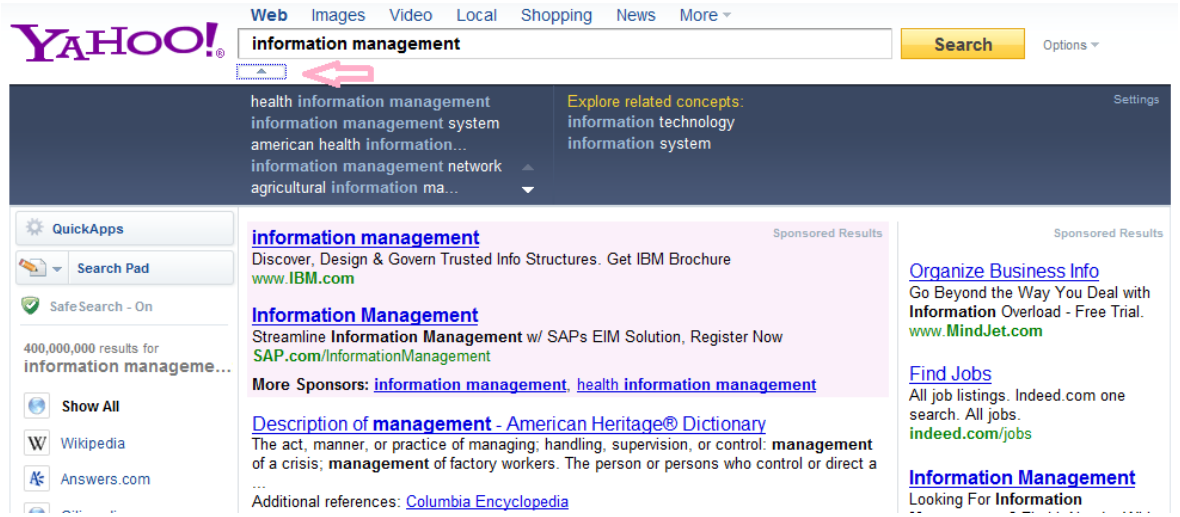
Η μηχανή Bing της Microsoft δίνει τις σχετικές αναζητήσεις απευθείας στα αποτελέσματα αναζήτησης της μηχανής:



Εικόνα 49 Σχετικοί όροι αναζήτησης από την Bing

3. Yahoo! Search Engine

Παρόμοια με την Bing, η Yahoo! προτείνει αναζητήσεις κάνοντας ένα κλικ στο κουμπί πάνω από τα αποτελέσματα:



Εικόνα 50 Προτάσεις σχετικών όρων από τη μηχανή της Yahoo!

Α.3.2.3 Μέριμνα για τη νομική κάλυψη

Καλό είναι να ελέγχουμε, κατά καιρούς, για τις λέξεις ή φράσεις που χρησιμοποιούμε στον ιστότοπό μας, καθώς ορισμένες από αυτές ενδέχεται κάποια στιγμή να γίνουν trademark από τις εταιρείες κολοσσούς. Για παράδειγμα, η εταιρεία που κρύβεται πίσω από τη διασημότερη, πλέον, σελίδα κοινωνικής δικτύωσης facebook έχει γνωστοποιήσει τις προθέσεις της για την κατοχύρωση της λέξης «book», ενώ έχει ήδη προχωρήσει τις διαδικασίες για την κατοχύρωση της λέξης «face», έχοντας ήδη μηνύσει (με επιτυχία) τη διαδικτυακή πύλη μίας κοινότητας καθηγητών με όνομα teachbook.com.

Προφανώς το βήμα αυτό δεν αφορά μόνο τις λέξεις – κλειδιά αλλά όλες τις λέξεις που περιέχονται στην ιστοσελίδα μας, αλλά έχοντας βελτιστοποιήσει την κατάταξή της στις μηχανές αναζήτησης για σχετικούς όρους – εμπορικά σήματα, είναι πολύ πιο εύκολο να την εντοπίσουν οι νομικές ομάδες των εταιρειών – κατόχων.

Παράρτημα Β Εποπτεία Ανίχνευσης

Παρακάτω, παρατίθεται ένα κομμάτι κώδικα PHP για κάθε μία εκ των σπουδαιότερων μηχανών αναζήτησης που, μετά την ενσωμάτωσή του στην αρχική σελίδα ενός ιστοτόπου, εντοπίζει κι ενημερώνει τον διαχειριστή, μέσω της αποστολής ενός e-mail, για την επισκεψιμότητα των ανιχνευτών της εκάστοτε μηχανής αναζήτησης στον ιστότοπο, τη συχνότητα των επισκέψεων αυτών καθώς και την ακριβή σελίδα που δέχεται την επίσκεψη.

Στο πεδίο της διεύθυνσης ηλεκτρονικής αλληλογραφίας, "Διεύθυνση e-mail" τοποθετείται ο επιθυμητός λογαριασμός – προορισμός στον οποίον θα αποστέλλονται οι ενημερώσεις των ανιχνεύσεων από την ορισμένη μηχανή αναζήτησης.

Β.1 Ανιχνευτές της Google (Googlebot)

```
<?php
//Στη μεταβλητή $emailaddress πρέπει να εισαχθεί η διεύθυνση e-mail
//του διαχειριστή

if ($_SERVER["HTTP_X_FORWARDED_FOR"]!= "")
{
$host = @gethostbyaddr($_SERVER["HTTP_X_FORWARDED_FOR"]);
}else{
$IP = $_SERVER["REMOTE_ADDR"];
$host = @gethostbyaddr($_SERVER["REMOTE_ADDR"]);
}

if(eregi("googlebot",$host))
{
$emailaddress = "Διεύθυνση e-mail";
mail("".$emailaddress."", "Εντοπίσθηκε επίσκεψη από την Google", "Η
διεύθυνση του ανιχνευτή είναι: " . $host . "\n και η σελίδα που
ανιχνεύθηκε είναι: " . $_SERVER['REQUEST_URI']."");
}
?>
```

Έτσι, σε περίπτωση ανίχνευσης μίας σελίδας του http://www.imu.ntua.gr από συγκεκριμένο ανιχνευτή της Google, στο λογαριασμό ηλεκτρονικής αλληλογραφίας του διαχειριστή θα αποσταλεί το αντίστοιχο μήνυμα, όπως για παράδειγμα το εξής:

Θέμα: **Εντοπίσθηκε επίσκεψη από την Google**

Η διεύθυνση του ανιχνευτή είναι: crawl-66-249-72-161.googlebot.com
και η σελίδα που ανιχνεύθηκε είναι: /?q=taxonomy/term/10

B.2 Ανιχνευτές της Yahoo (Slurp)

```php
<?php
//Στη μεταβλητή $emailaddress πρέπει να εισαχθεί η διεύθυνση e-mail
//του διαχειριστή

if ($_SERVER["HTTP_X_FORWARDED_FOR"]!= "")
{
$host = @gethostbyaddr($_SERVER["HTTP_X_FORWARDED_FOR"]);
}else{
$IP = $_SERVER["REMOTE_ADDR"];
$host = @gethostbyaddr($_SERVER["REMOTE_ADDR"]);
}

if(eregi("yahoo",$host))
{
$emailaddress = "Διεύθυνση e-mail";
mail("".$emailaddress."", "Εντοπίσθηκε επίσκεψη από την Yahoo!", "Η
διεύθυνση του ανιχνευτή είναι: " . $host . "\n και η σελίδα που
ανιχνεύθηκε είναι: " . $_SERVER['REQUEST_URI']."");
}
?>
```

B.3 Ανιχνευτές της Bing (Bingbot)

```php
<?php
//Στη μεταβλητή $emailaddress πρέπει να εισαχθεί η διεύθυνση e-mail
//του διαχειριστή

if ($_SERVER["HTTP_X_FORWARDED_FOR"]!= "")
{
$host = @gethostbyaddr($_SERVER["HTTP_X_FORWARDED_FOR"]);
}else{
$IP = $_SERVER["REMOTE_ADDR"];
$host = @gethostbyaddr($_SERVER["REMOTE_ADDR"]);
}

if(eregi("msn",$host))
{
$emailaddress = "Διεύθυνση e-mail";
mail("".$emailaddress."", "Εντοπίσθηκε επίσκεψη από την Bing", "Η
διεύθυνση του ανιχνευτή είναι: " . $host . "\n και η σελίδα που
ανιχνεύθηκε είναι: " . $_SERVER['REQUEST_URI']."");
}
?>
```