ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΟΙΚΟΝΟΜΙΚΗ ΚΑΙ ΔΙΟΙΚΗΣΗ ΤΩΝ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βελτιστοποίηση Αποτελεσμάτων Μηχανών Αναζήτησης σε Δυναμικές Ιστοσελίδες

Βίκτωρ Κ. Κυρίτσης

Επιβλέποντες: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής ΕΚΠΑ

ΑΘΗΝΑ

ΑΠΡΙΛΙΟΣ 2010

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βελτιστοποίηση Αποτελεσμάτων Μηχανών Αναζήτησης σε Δυναμικές Ιστοσελίδες

Βίκτωρ Κ. Κυρίτσης

Α.Μ.: ΜΟΠ 116

Επιβλέποντες: Ευστάθιος Χατζηευθυμιάδης, Επίκουρος Καθηγητής ΕΚΠΑ

ΑΠΡΙΛΙΟΣ 2010

ΠΕΡΙΛΗΨΗ

Στα πλαίσια της παρούσας διπλωματικής εργασίας εξετάζονται οι παράμετροι στη δημιουργία των δυναμικών ιστοσελίδων οι οποίες έχουν βαρύνουσα σημασία για την επίτευξη καλύτερης θέσης στους καταλόγους των αποτελεσμάτων των μηχανών αναζήτησης. Η χρήση νέων και αποδοτικότερων τεχνολογιών στον εξυπηρετητή ευνοεί την παραγωγή και την προσφορά δυναμικού περιεχομένου σε δικτυακές εφαρμογές οι οποίες προσπελαύνονται από μεγάλο αριθμό χρηστών. Επίσης, η δυνατότητα μεταβολής του περιεχομένου των ιστοσελίδων με εύκολο τρόπο μέσω των συστημάτων διαχείρισης περιεχομένου έχει διαμορφώσει την τάση της επικράτησης των δυναμικών ιστοσελίδων έναντι των στατικών, οι οποίες αποτελούσαν την καθιερωμένη λύση στα αρχικά στάδια ανάπτυξης του παγκόσμιου ιστού.

Η δημιουργία δυναμικών ιστοσελίδων, οι οποίες κατέχουν υψηλή θέση στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης προϋποθέτει την κατανόηση του τρόπου λειτουργίας των τελευταίων. Συνεπώς, αρχικώς παρουσιάζονται τα λογικά τμήματα που συνθέτουν μία μηχανή αναζήτησης. Κατόπιν, επιχειρείται η μοντελοποίηση μίας δυναμικής ιστοσελίδας μέσω της διαδικασίας της τμηματοποίησης ενώ κατηγοριοποιείται και ποσοτικοποιείται όπου είναι εφικτό η έννοια της μεταβολής για μία δεδομένη ιστοσελίδα. Τέλος δίνονται ευριστικοί κανόνες για τη σύνταξη δυναμικών ιστοσελίδων «φιλικών» προς τις μηχανές αναζήτησης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Εφαρμογές Διαδικτύου

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: δυναμική ιστοσελίδα, μηχανές αναζήτησης, γράφος διαδικτύου, βαθμός δημοτικότητας, βαθμός περιεχομένου, κατάλογος αποτελεσμάτων, μαρκετινγκ μηχανών αναζήτησης

ABSTRACT

In this master thesis, we examine the most important factors on the creation of dynamically generated web pages in the context of the search engine optimization. The advent of new and more effective technologies on the server side results in the creation and the serving of dynamic content accessed by a large number of users. Furthermore, the prevalence of the dynamic web pages against the static ones, which were supposed to be the only feasible solution at the beginning of the web, is due to the easy way that content changes through the use of the content management systems.

The understanding of the functionality of the constituent modules of a search engine is a prerequisite for the creation of dynamic web pages with a high search-engine ranking. Thus, we firstly present the logical parts of a search engine. Additionally, a way to model a dynamic web page through a process called fragmentation is exhibited. This approach helps to identify the stable and the dynamic content within each page. Also, a classification of the changes, that take place in a given web page, as well as some quantification measures are given. At last, we propose some empirical rules for the creation of dynamic web pages that are “friendly” to the search engines.

SUBJECT AREA: Web Applications

KEYWORDS: dynamic web page, search engines, web graph, popularity score, content score, search engine results, search engine marketing

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

Η λειτουργία των σύγχρονων μηχανών αναζήτησης πληροφορίας στο διαδίκτυο βασίζεται σε τεχνικές των παραδοσιακων συστημάτων ανάκτησης πληροφοριών (Information Retrieval System), όπως η λεκτική και σημασιολογική επεξεργασία των κειμένων, η δημιουργία αποδοτικών δομών ευρετηρίασης με όρους δεικτοδότησης, η γρήγορη και εύκολη πρόσβαση σε αυτές, η ανάκτηση πληροφοριών σχετικών με το ερώτημα του χρήστη. Ωστόσο, ο μεγάλος όγκος πληροφορίας του διαδικτύου προς επεξεργασία σε συνδυασμό με τον δυναμικό χαρακτήρα των ιστοσελίδων καθώς και η ύπαρξη αναφορών μεταξύ των ιστοσελίδων με την μορφή συνδέσμων καθιστούν την εργασία της ανάκτησης πληροφορίας ιδιαιτέρως απαιτητική. Ακολουθεί μία γρήγορη αναφορά στα κύρια γνωρίσματα του παγκόσμιου ιστού (World Wide Web) δεδομένου ότι συνιστά μία τεράστια βάση δεδομένων κειμένου, κάθε εγγραφή της οποίας αντιστοιχεί σε μία ιστοσελίδα (web page).

- *Μεγάλος Ογκος Δεδομένων*: Ο ακριβής αριθμός των ιστοσελίδων που συνιστούν τον Παγκόσμιο Ιστό είναι αγνωστός. Τον Ιανουάριο του 2004 εκτιμήθηκε κατά προσέγγιση ότι ο συνολικός αριθμός είναι της τάξης του 10^9 ενώ το μέσο μέγεθος της ιστοσελίδας είναι 500 Kb. Αν και ο ρυθμός μεγέθυνσης του Παγκόσμιου ιστού έχει επιβραδυνθεί, παραμένει η μεγαλύτερη υπαρκτή συλλογή κειμένων. Ωστόσο, πρέπει να σημειωθεί ότι από την προηγούμενη εκτίμηση απουσιάζουν οι ιστοσελίδες οι οποίες παράγονται δυναμικά κατόπιν αίτησης του χρήστη. Η εταιρεία Bright-Planet, η οποία δραστηριοποιείται στο χώρο των μηχανών αναζήτησης, εκτίμησε ότι η συλλογή των δυναμικώς παραγόμενων ιστοσελίδων είναι τουλάχιστον δύο φορές τάξεις μεγέθους μεγαλύτερη από την αντίστοιχη συλλογή των στατικών ιστοσελίδων.
- *Δυναμικός Χαρακτήρας του Παγκόσμιου Ιστού*: Στις παραδοσιακές συλλογές κειμένων διακρίνονται δύο μορφές στατικότητας. Η πρώτη μορφή στατικότητας αποδίδεται στο γεγονός ότι με την προσθήκη ενός κειμένου σε μία συλλογή το τελευταίο δεν υπόκειται σε περαιτέρω αλλαγές, π.χ., η συλλογή κειμένων–βιβλίων σε μία βιβλιοθήκη. Σε έρευνα των Junghoo Cho και Hector Garcia-Molina (1) σε ένα αντιπροσωπευτικό υποσύνολο του παγκόσμιου ιστού εκτιμήθηκε ότι το περιεχόμενο του 40% των ιστοσελίδων τροποποιούταν μια φορά την εβδομάδα,

ενώ η αλλαγή του περιεχομένου των ιστοσελίδων, των οποίων η διεύθυνση λήγει σε .com ήταν ημερήσια. Επίσης, η συχνότητα τροποποίησης μίας ιστοσελίδας συναρτάται του μεγέθους της (2). Το περιεχόμενο σελίδων με μεγάλο μέγεθος τροποποιείται συχνότερα σε σύγκριση με το περιεχόμενο σελίδων μικρού μέγεθος.

Η δεύτερη μορφή στατικότητας στις παραδοσιακές συλλογές κειμένων αποδίδεται στο γεγονός ότι το μέγεθός τους μεταβάλλεται με αργό ρυθμό σε σύγκριση με το αντίστοιχο μέγεθος του Παγκόσμιου Ιστού. Η προσθαφαίρεση εκατοντάδων ή έστω μερικών χιλιάδων στοιχείων σε μία παραδοσιακή συλλογή θεωρείται αμελητέα στην περίπτωση του Παγκόσμιου Ιστού.

- *Απουσία Κεντρικού Ελέγχου*: Ο παγκόσμιος ιστός χαρακτηρίζεται από την απουσία προτύπων τα οποία καθορίζουν το περιεχόμενο, την μορφοποίηση και τη δομή των ιστοσελίδων. Επίσης, κάθε χρήστης του διαδικτύου με πρόσβαση στις κατάλληλες τεχνολογίες είναι εν δυνάμει συντάκτης περιεχομένου μίας ιστοσελίδας. Τα δεδομένα του διαδικτύου είναι ετερογενοί, δημοσιευμένα σε πολλαπλές μορφές και γλώσσες. Το φαινόμενο της «αναρχίας» στον Παγκόσμιο Ιστό επιτείνεται από το γεγονός της απουσίας ελέγχου της ορθότητας του περιεχομένου των ιστοσελίδων. Τέλος, η απουσία κεντρικού ελέγχου έχει ως συνέπεια την δημιουργία σελίδων διαφορετικής θεματολογίας, όπως η διεκπεραίωση συναλλαγών ηλεκτρονικού εμπορίου, η γρήγορη αναζήτηση πληροφοριών, κ.α. ενώ συνεπάγεται την δυσκολία κατηγοριοποίησης των ιστοσελίδων.
- *Υπαρξη Συνδέσμων μεταξύ των Ιστοσελίδων*: Οι τεχνολογίες διαδικτύου δίνουν τη δυνατότητα στον δημιουργό μίας ιστοσελίδας της συμπερίληψης αναφορών, υπερσυνδέσμων (hyperlink) σε άλλες ιστοσελίδες. Κατά την περιήγησή του σε μία σελίδα ο χρήστης μπορεί να ανακατευθυνθεί άμεσα σε μία άλλη ιστοσελίδα επιλέγοντας έναν εκ των υπερσυνδέσμων της αρχικής ιστοσελίδας. Η ύπαρξη των συνδέσμων δημιουργεί ένα πολύπλοκο πλέγμα αλληλοσυνδεόμενων ιστοσελίδων.

Στο συγκεκριμένο περιβάλλον, η κύρια λειτουργία των μηχανών αναζήτησης (search engines) συνοψίζεται στην αλγοριθμική, συστηματική επεξεργασία του μεγάλου όγκου πληροφορίας του διαδικτύου με απώτερο σκοπό τη διευκόλυνση του χρήστη στην προσπάθεια του εύρεσης ποιοτικής πληροφορίας σε σύντομο χρονικό διάστημα. Σε

γενικές γραμμές, οι πηγές πληροφορίας στο διαδίκτυο εντοπίζονται αφενός στο περιεχόμενο κειμένου των ιστοσελίδων και αφετέρου στο τρόπο διασύνδεσής τους μέσω των υπερσυνδέσμων. Η διαδικασία ανάλυσης της πληροφορίας από την μηχανή αναζήτησης περιλαμβάνει την ανάκτηση του συνόλου ή μέρος των ιστοσελίδων του διαδικτύου, την επεξεργασία του περιεχομένου τους και του συνόλου των υπερσυνδέσμων καθώς και την ευρετηρίασή τους σε κατάλληλες δομές δεδομένων.

Η πρόσβαση του χρήστη στον παγκόσμιο ιστό πραγματοποιείται μέσω των προγραμμάτων περιήγησης ιστού (φυλλομετρητές – web browser), τα οποία εμφανίζουν το περιεχόμενο των ιστοσελίδων κωδικοποιημένο σε HTML ή XML. Συνήθως ο χρήστης δεν γνωρίζει την ακριβή διεύθυνση της ιστοσελίδας της οποίας το περιεχόμενο τον ενδιαφέρει, ενώ η τυχαία αναζήτηση είναι φύσει αδύνατη λόγω του μεγάλου όγκου πληροφορίας. Οι μηχανές αναζήτησης ικανοποιούν τη συγκεκριμένη πληροφοριακή ανάγκη. Ο χρήστης υποβάλλει το ερώτημά του σε κατάλληλο πεδίο κειμένου στην ιστοσελίδα της μηχανής αναζήτησης, το οποίο αποτελείται από μία η περισσότερες λέξεις. Εν γένει, οι ιστοσελίδες των μηχανών αναζήτησης έχουν μινιμαλιστική διεπαφή χρήστη ενώ η διεύθυνση τους είναι ευκολομνημόνευτη, π.χ., `www.google.com` και `www.bing.com`. Ως απάντηση στο ερώτημα η μηχανή αναζήτησης επιστρέφει ένα κατάλογο αποτελεσμάτων με υπερσυνδέσμους προς ιστοσελίδες, οι οποίες σχετίζονται με το ερώτημα του χρήστη. Επειδή η πληθυκότητα του συνόλου των αποτελεσμάτων ενδέχεται να είναι μεγάλη, ο χρόνος εξέτασης του περιεχομένου όλων των ιστοσελίδων από το χρήστη είναι απαγορευτικά μεγάλος. Συνεπώς, οι μηχανές αναζήτησης ταξινομούν κατά φθίνουσα σειρά σημαντικότητας τα αποτελέσματα, τα οποία αντιστοιχούν στην ερώτηση. Το περιεχόμενο των ιστοσελίδων στον κατάλογο αποτελεσμάτων είναι ανταποκρίνεται με μεγάλη πιθανότητα στις πληροφοριακές ανάγκες του χρήστη. Σύμφωνα με εμπειρικές μελέτες ο μέγιστος αριθμός ιστοσελίδων, τις οποίες θα εξετάσει ο μέσος χρήστης μίας μηχανής αναζήτησης, αντιστοιχούν στους τριάντα πρώτους υπερσυνδέσμους του καταλόγου αποτελεσμάτων. Σημειώνεται ότι δεδομένου ενός ερωτήματος, τα αποτελέσματα καθώς και η κατάταξή τους εξαρτώνται της υλοποίησης της μηχανής αναζήτησης. Κάθε μηχανή αναζήτησης διαθέτει το δικό της εξελισσόμενο αλγόριθμο με ένα αρκετά μεγάλο αριθμό συντελεστών στάθμισης, οι οποίοι καθορίζουν την κατάταξη των αποτελεσμάτων.

Η διείσδυση των πρακτικών του ηλεκτρονικού επιχειρείν και του ηλεκτρονικού εμπορίου επιβάλλει την δικτυακή παρουσία των εταιρικών οργανισμών με απώτερο σκοπό τον αυτοματισμό των εμπορικών συναλλαγών και των ροών εργασίας, την μείωση του κόστου συναλλαγών με παράλληλη αύξηση της ταχύτητας και της ποιότητας των παρεχόμενων υπηρεσιών.

Η άμεση σύνδεση των τεχνολογιών του παγκόσμιου ιστού με την κερδοφορία της επιχείρησης απαιτούν την εφαρμογή μεθόδων προώθησης στο διαδίκτυο με τη χρήση διαφόρων τεχνικών, μεθόδων προσέλκυσης επισκεπτών στον εταιρικό ιστότοπο. Χαρακτηριστικά παραδείγματα τεχνικών αποτελούν οι διαφημίσεις στο διαδίκτυο, το ηλεκτρονικό ταχυδρομείο καθώς και το μάρκετινγκ των μηχανών αναζήτησης. Στη συγκεκριμένη εργασία μελετάται η τεχνική του μάρκετινγκ των μηχανών αναζήτησης (search engine marketing) στο πλαίσιο της εκστρατείας δημιουργίας επισκεψιμότητας ενός ιστοτόπου (traffic-building campaign), αποτελούμενο κυρίως από ιστοσελίδες, οι οποίες παράγονται με δυναμικό τρόπο. Είναι σαφές ότι αν οι ιστοσελίδες ενός ιστοτόπου (website) δεν καταλαμβάνουν υψηλη θέση στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης, τότε ο οργανισμός ενδέχεται να απωλέσει σημαντικές ευκαιρίες, π.χ., απώλεια μεριδίου αγοράς στην περίπτωση ιστοτόπων ηλεκτρονικού εμπορίου.

Η βελτιστοποίηση αποτελεσμάτων των μηχανών αναζήτησης ορίζεται ως μία δομημένη μέθοδος βελτίωσης της θέσης των ιστοσελίδων στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης για επιλεγμένες φράσεις κλειδιά (keyword). Εσωτερικοί παράγοντες, οι οποίοι καθορίζονται πλήρως από τον διαχειριστή του ιστοτόπου και άπτονται του περιεχομένου, σε συνδυασμό με την ύπαρξη αναφορών, υπερσυνδέσμων προσδιορίζουν την θέση των ιστοσελίδων του ιστοτόπου στους καταλόγους αποτελεσμάτων. Γενικώς, η βελτιστοποίηση της θέσης μίας ιστοσελίδας χαρακτηρίζεται ως μία διαδικασία με μεγάλο βαθμό δυσκολίας, η οποία μπορεί τελικώς να μην είναι επιτυχής.

Ωστόσο, η κατάληψη μίας αξιοπρόσεκτης θέσης στον κατάλογο αποτελεσμάτων απαιτεί την ικανοποίηση ενός συνόλου κριτηρίων καθώς και τη ρύθμιση ορισμένων συντελεστών στάθμισης. Ενδεικτικά αναφέρεται η επιλογή κατάλληλων φράσεων κλειδιών δηλωτικών του αντικειμένου που διαπραγματεύεται ο ιστότοπος, η συχνότητα

επανάληψης μίας φράσης κλειδί (keyword frequency) στο περιεχόμενο της ιστοσελίδας, το κείμενο στην ετικέτα τίτλου της, το περιεχόμενο των μετα-ετικετών (meta-tag) στο αρχείο HTML, ο αριθμός των εισερχόμενων συνδέσμων, το κείμενο αγκύρωσης σε υπερσυνδέσμους (anchor text) κ.α. Κάθε κριτήριο έχει διαφορετικό συντελεστή βαρύτητας στη διαμόρφωση της τελικής θέσης μίας ιστοσελίδας. Επίσης, οι συντελεστές βάρους μεταβάλλονται διαρκώς από τους μηχανικούς των μηχανών αναζήτησης, οπότε στερείται νοήματος οποιαδήποτε προσπάθεια απόδοσης μονομερούς κατεύθυνσης στην διαδικασία βελτιστοποίησης.

Οι κυρίοτεροι λόγοι υλοποίησης ενός στρατηγικού πλαισίου βελτιστοποίησης της κατατακτήριας θέσης συνοψίζονται στα ακόλουθα δύο σημεία

- Αύξηση του ποσοστού επισκεψιμότητας του δικτυακού ιστοτόπου και κατά συνέπεια μεγέθυνση της εμβέλειας, η οποία αναφέρεται στον πιθανό αριθμό επισκεπτών – πελατών με τους οποίους ένας οργανισμός ή ένα φυσικό πρόσωπο μπορεί να έρθει σε επαφή.
- Σε επίπεδο εταιρικών οργανισμών, αύξηση του περιθωρίου κέρδους μέσω της προσφοράς υπηρεσιών και της πώλησης προϊόντων.

Η τροποποίηση του περιεχομένου της ιστοσελίδας για την επίτευξη καλύτερης θέσης στους καταλόγους αποτελεσμάτων είναι μία συνεχής διαδικασία, η οποία προϋποθέτει τον έλεγχο, την επίβλεψη και τη συχνή αναθεώρηση της ακολουθούμενης στρατηγικής βελτιστοποίησης. Η δυσκολία εύρεσης της κατάλληλης στρατηγικής οφείλεται στο γεγονός ότι ο βαθμός επιτυχίας της δεν προσδιορίζεται άμεσα αλλά απαιτεί την πάροδο ενός εύλογου χρονικού διαστήματος. Συνεπώς οι διαχειριστές των ιστοτόπων θέτουν μία σειρά προτεραιότητας στις ιστοσελίδες στο πλαίσιο της εφαρμογής των τεχνικών βελτιστοποίησης. Η αρχική ιστοσελίδα ενός ιστοτόπου, οι ιστοσελίδες με ποιοτικό και ενδιαφέρον περιεχόμενο αποτελούν τους πρωταρχικούς «στόχους» μίας συντονισμένης διαδικασίας βελτιστοποίησης.

Διακρίνονται οι ακόλουθες κύριες συνιστώσες στην προσπάθεια επίτευξης καλής θέσης στους καταλόγους αποτελεσμάτων.

- *Δημιουργία ποιοτικού περιεχομένου*: Ο μέσος χρήστης του διαδικτύου αναζητά ιστοσελίδες με ενδιαφέρον περιεχόμενο για την ικανοποίηση των πληροφοριακών

του αναγκών. Η σαφής και ακριβής διατύπωση του περιεχομένου, η εύστοχη επιλογή φράσεων κλειδιών, η τοποθέτησή τους σε κατάλληλα σημεία της ιστοσελίδας συντελούν στη δημιουργία αξιόπιστου περιεχομένου και κατά συνέπεια στη κατάληψη υψηλότερης θέσης στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης.

- *Υπερσυνδέσμων από και προς ιστοσελίδες με ποιοτικό περιεχόμενο*: Με την επιλογή ενός υπερσυνδέσμου ο χρήστης του διαδικτύου κατευθύνεται σε μία νέα ιστοσελίδα, η οποία βρίσκεται εντός ή εκτός του ιστοτόπου. Οι υπερσύνδεσμοι σε μία ιστοσελίδα διακρίνονται σε εξερχόμενους και εισερχόμενους. Οι εξερχόμενοι σύνδεσμοι σε μία ιστοσελίδα καθορίζονται αποκλειστικά από τον δημιουργό του περιεχομένου της ιστοσελίδας και αποτελούν αναφορές-παραπομπές στο περιεχόμενο τρίτων ιστοσελίδων. Το σύνολο των εισερχομένων συνδέσμων σε μία ιστοσελίδα αποτελείται από τις αναφορές ιστοσελίδων του διαδικτύου προς αυτή. Οι σύγχρονες μηχανές αναζήτησης αξιοποιούν την παρουσία συνδέσμων μεταξύ των ιστοσελίδων για τη βαθμολόγηση των τελευταίων. Προς αυτή την κατεύθυνση οι αλγόριθμοι HITS και PageRank έχουν προταθεί.
- *Συμπερίληψη ετικετών μεταδεδομένων στο κώδικα HTML*: Το περιεχόμενο των μεταδεδομένων (metadata) δεν είναι ορατό στο χρήστη. Ωστόσο, οι τύποι μεταδεδομένων, στους οποίους δίνεται μία σύντομη περίληψη του περιεχομένου μίας ιστοσελίδας και παρατίθονται οι φράσεις κλειδιά, παίζουν ρόλο στην κατάληψη καλύτερης θέσης της στους καταλόγους αποτελεσμάτων.

Στο παρόν πόνημα παρουσιάζονται αναλυτικά οι προηγούμενοι τρεις άξονες στη βελτιστοποίηση αποτελεσμάτων των μηχανών αναζήτησης στα πλαίσια των ιστοσελίδων οι οποίες παράγονται με δυναμικό τρόπο. Η δομή της εργασίας είναι η ακόλουθη. Στο κεφάλαιο 2 περιγράφεται η δομή των δυναμικών ιστοσελίδων. Επίσης γίνεται αναφορά στα χαρακτηριστικά των αλλαγών στις οποίες υπόκεινται οι δυναμικές ιστοσελίδες. Το κεφάλαιο 3 αναφέρεται στην αρχιτεκτονική των μηχανών αναζήτησης. Περιγράφονται συνοπτικά τα δομικά στοιχεία μίας μηχανής αναζήτησης ενώ κατηγοριοποιούνται με κριτήριο το βαθμό συσχέτισής τους με το ερώτημα του χρήστη. Η κατατακτήρια θέση μίας ιστοσελίδας στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης καθορίζεται από τον συνδυασμό δύο επιμέρους βαθμών, του βαθμού

περιεχομένου και του βαθμού δημοτικότητας. Στο κεφάλαιο 4 παρουσιάζονται αναλυτικά οι επαναληπτικοί αλγόριθμοι HITS και PageRank, οι οποίοι προσδιορίζουν ποσοτικά τον βαθμό δημοτικότητας των ιστοσελίδων. Δίνονται λεπτομέρειες για το γράφο διαδικτύου και στις ιδιότητές του, εφόσον η δομή καθορίζει το χρόνο εκτέλεσης των αλγορίθμων HITS και PageRank. Στο κεφάλαιο 5 παρουσιάζονται τεχνικές και μέθοδοι για το κτίσιμο του συνόλου των εισερχομένων και εξερχομένων συνδέσμων μίας ιστοσελίδας ώστε να βελτιστοποιηθεί ο βαθμός δημοτικότητάς της. Στα κεφάλαια 6 και 7 προσδιορίζονται οι παράμετροι περιεχομένου και μεταδεδομένων αντίστοιχα οι οποίοι καθορίζουν το βαθμό περιεχομένου. Στο κεφάλαιο 6 δίνονται ευριστικοί κανόνες για τη σύνταξη ποιοτικού περιεχομένου, ενώ στο κεφάλαιο 7 παρουσιάζονται τα αποτελέσματα πειραματικής μελέτης για τον ρόλο των μεταδεδομένων περιεχομένου στην κατάταξη των ιστοσελίδων στους καταλόγους των αποτελεσμάτων.

2 Δυναμικές Ιστοσελίδες

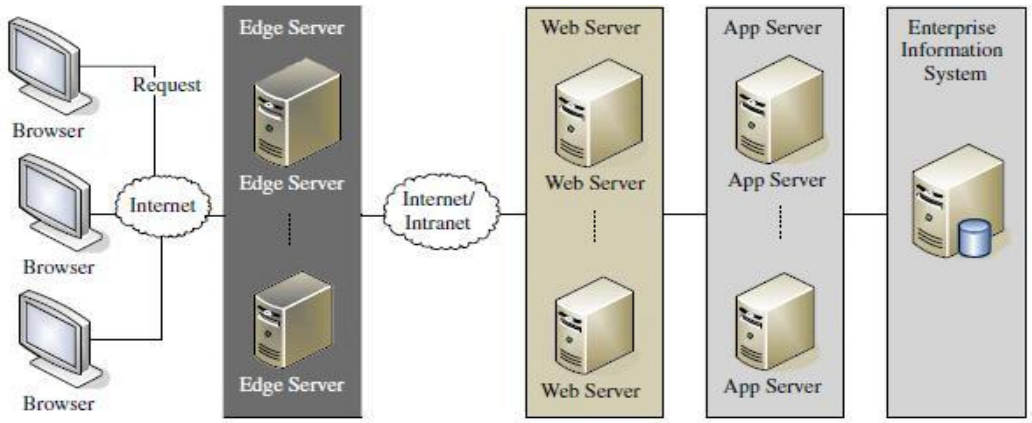
Στα αρχικά στάδια ανάπτυξης των τεχνολογιών διαδικτύου η πλειοψηφία των ιστοτόπων «δημοσίευε» ιστοσελίδες με στατικό περιεχόμενο (static web page). Οι εξυπηρετητές διαδικτύου φιλοξενούσαν αρχεία HTML με στατικό περιεχόμενο και εικόνες. Ωστόσο, στο διάστημα των τελευταίων δέκα ετών οι τεχνολογίες διαδικτύου εξελίσσονται με σταθερό και γρήγορο ρυθμό για την ενσωμάτωση δυναμικού περιεχομένου στις ιστοσελίδες. Οι εφαρμογές διαδικτύου (application web) βασίζονται στην εκτέλεση προγραμμάτων στον εξυπηρετητή για την παραγωγή δυναμικών ιστοσελίδων ανταποκρινομένων στα αιτήματα των χρηστών. Η ανάπτυξη των εφαρμογών του διαδικτύου αφενός προσέφερε τη δυνατότητα ολοκλήρωσης των διαδικασιών ηλεκτρονικού επιχειρείν, και αφετέρου συντέλεσε στην ανάπτυξη των υπηρεσιών ιστού (service web) οι οποίες βασίζονται στην τεχνολογία XML. Οι εφαρμογές διαδικτύου καθώς και οι υπηρεσίες ιστού προϋποθέτουν την παραγωγή δυναμικού περιεχομένου στον εξυπηρετητή για κάθε αίτημα του χρήστη.

Το μοντέλο συστήματος για την παραγωγή και δημοσίευση του δυναμικού περιεχομένου των ιστοσελίδων αποτελείται από τα ακόλουθα μέρη

1. **Εξυπηρετητής ακμής (Edge Server)**: Ο συγκεκριμένος εξυπηρετητής επεκτείνει την λειτουργικότητα του εξυπηρετητή διαδικτύου με την εναποθήκευση περιεχομένου και την δυνατότητα εκτέλεσης κώδικα εφαρμογής με απώτερο σκοπό την μείωση αφενός του χρόνου απόκρισης και αφετέρου του φόρτου εργασίας του εξυπηρετητή.
2. **Εξυπηρετητής διαδικτύου (Web Server)**: Ο εξυπηρετητής διαδικτύου δέχεται αιτήματα σύμφωνα με το πρωτόκολλο HTTP και στέλνει την απάντηση στον φυλλομετρητή του χρήστη.
3. **Εξυπηρετητής εφαρμογών (Application Server)**: Ο συγκεκριμένος εξυπηρετητής εκτελεί τον κώδικα των προγραμμάτων. Συνήθως συνδέεται με ένα ή περισσότερα σύστηματα διαχείρισης βάσης δεδομένων κα εκτελεί πράξεις ανάγνωσης και εγγραφής δεδομένων σε αυτά. Το δυναμικό περιεχόμενο των ιστοσελίδων παράγεται κυρίως στους εξυπηρετητές εφαρμογών.

4. **Σύστημα διαχείρισης βάσης δεδομένων (DataBase Management Systems)**: Αποτελεί τον κύριο χώρο αποθήκευσης του περιεχομένου των ιστοσελίδων.

Σχηματικά



Σχήμα 1 Τυπική αρχιτεκτονική των εφαρμογών διαδικτύου.

2.1 Χαρακτηριστικά των ιστοτόπων με δυναμικό περιεχομένο

Στην παρούσα ενότητα παρουσιάζονται τα κύρια χαρακτηριστικά των δυναμικών ιστοσελίδων (dynamic web page).

- Το μέσο μέγεθος των ιστοσελίδων είναι της τάξης των 10-15 Kb.
- Το ποσοστό δημοτικότητας των ιστοσελίδων δεν είναι ομοιόμορφα κατανεμημένο. Ενα μικρό ποσοστό των ιστοσελίδων είναι υπεύθυνο για την πλειοψηφία των αιτημάτων των χρηστών.
- Ο ρυθμός πρόσβασης των ιστοσελίδων από τους χρήστες είναι πολύ υψηλότερος της συχνότητας τροποποίησης των ιστοσελίδων.
- Οι δυναμικώς παραγόμενες ιστοσελίδες τροποποιούνται συχνότερα των στατικών ιστοσελίδων.

Οι ιστοσελίδες με δυναμικό περιεχόμενο μπορούν να ταξινομηθούν σε δύο κατηγορίες. Η πρώτη κατηγορία, η οποία καλείται *δυναμικές ιστοσελίδες*, περιέχει τις ιστοσελίδες οι οποίες παράγονται χωρίς να λαμβάνονται υπόψη πληροφορίες σχετικές με το χρήστη. Στη συγκεκριμένη κατηγορία δεν απαιτείται η γνώση στοιχείων για το χρήστη, ο οποίος προσπελάυνει την ιστοσελίδα. Συνεπώς, μία ιστοσελίδα της συγκεκριμένης κατηγορίας είναι η ίδια σε δεδομένη χρονική στιγμή για οποιονδήποτε χρήστη. Ως παράδειγμα

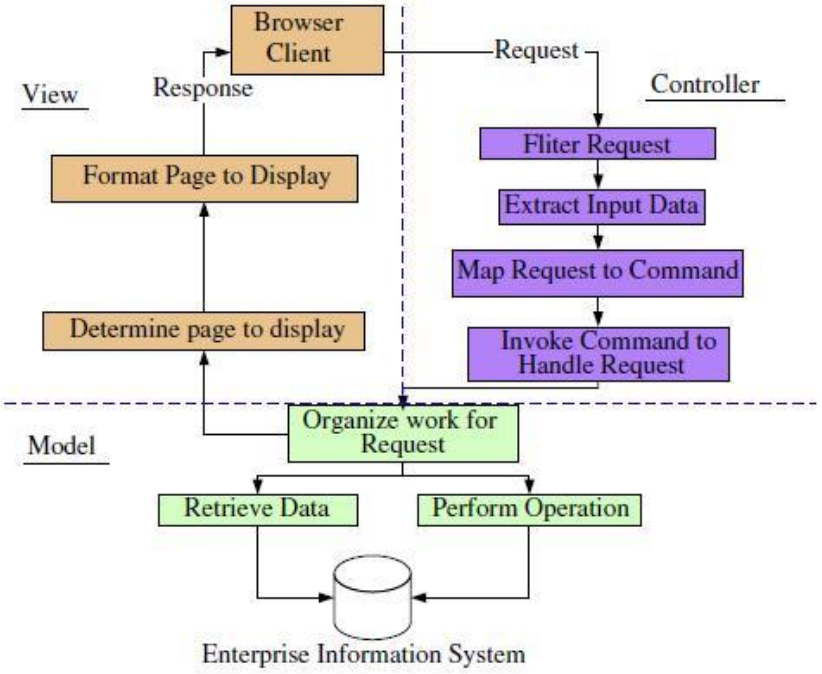
αναφέρονται οι ιστοσελίδες στις οποίες παρουσιάζονται δεδομένα πραγματικού χρόνου, όπως η τιμή του δείκτη αξιών στο χρηματιστήριο. Κύριο γνώρισμά τους αποτελεί το γεγονός ότι εντός μικρού χρονικού διαστήματος δύο διαφορετικές εκδόσεις της ίδιας ιστοσελίδας παρουσιάζουν μεγάλο βαθμό ομοιότητας.

Οι *εξατομικευμένες ιστοσελίδες* αποτελούν την δεύτερη κατηγορία ιστοσελίδων με δυναμικό περιεχόμενο. Οι ιστοσελίδες της συγκεκριμένης κατηγορίας παράγονται κατόπιν αιτήματος του χρήστη μέσω ενός ασφαλούς συστήματος (όνομα χρήστη κωδικός πρόσβασης). Ολοι οι κύριοι πάροχοι υπηρεσιών διαδικτύου προσφέρουν δυνατότητα πρόσβασης σε εξατομικευμένες ιστοσελίδες. Χαρακτηριστικά παραδείγματα αποτελούν οι ιστοσελίδες iGoogle και My Yahoo! καθώς και οι ιστότοποι κοινωνικής δικτύωσης. Η χρήση των πληροφοριών χρήστη στην παραγωγή των ιστοσελίδων καθιστά σχεδόν κάθε απάντηση του εξυπηρετητή διαδικτύου μοναδική και μερικές φορές εξαρτώμενη από την ακολουθία προηγούμενων αιτήσεων του χρήστη.

2.2 Παραγωγή δυναμικού περιεχομένου

Η διαδικασία της παραγωγής δυναμικού περιεχομένου περιλαμβάνει την εκτέλεση κώδικα προγράμματος, την πρόσβαση σε βάσεις δεδομένων και τη δημιουργία του πλαισίου παρουσίασης. Οι προηγούμενες υπο-διαδικασίες πραγματοποιούνται στον εξυπηρετητή εφαρμογών. Η διαδικασία παραγωγής δυναμικών ιστοσελίδων μοντελοποιείται με τη υιοθέτηση της αρχιτεκτονικής πρότυπο-όψη-ελεγχος **M**odel-**V**iew-**C**ontrol (MVC). Στο Σχήμα 2 Το μοντέλο μοντέλο-όψη-έλεγχος. παρουσιάζεται η αρχιτεκτονική MVC. Το τμήμα ελέγχου δέχεται τις HTTP αιτήσεις, ερμηνεύει την ακολουθία χαρακτήρων URL, και τις απεικονίζει στην εντολή η οποία πρόκειται να εκτελεστεί από το πρότυπο. Επίσης, το τμήμα ελέγχου καθορίζει τον τρόπο παρουσίασης της απάντηση στον φυλλομετρητή του χρήστη. Το πρότυπο αντιπροσωπεύει τα δεδομένα καθώς και τις αλγοριθμικές διαδικασίες, οι οποίες τα επεξεργάζονται. Οπως έχει ήδη αναφερθεί τα δεδομένα συνήθως αποθηκεύονται σε συστήματα διαχείρισης βάσεων δεδομένων. Τέλος, η όψη στο μοντέλο MVC είναι αρμόδια για την εποίκηση των προτύπων παρουσίασης με τα δεδομένα. Η χρήση της συγκεκριμένης αρχιτεκτονικής για την μοντελοποίηση της διαδικασίας δημιουργίας δυναμικών ιστοσελίδων είναι κατάλληλη εφόσον διαχωρίζει με σαφή τρόπο τα δεδομένα

από τις αλγοριθμικές διαδικασίες καθώς και τα περιεχόμενα από το τρόπο παρουσίασής τους.



Σχήμα 2 Το μοντέλο μοντέλο-όψη-έλεγχος.

2.3 Δομή των δυναμικών ιστοσελίδων

Οι δυναμικές ιστοσελίδες συντίθενται απο απλούστερες οντότητες γνωστές ως τμήματα (fragment). Εξ' ορισμού ένα τμήμα συνιστά ένα μέρος μίας ιστοσελίδας το οποίο έχει διακριτή θεματολογία ή/και λειτουργικότητα και το οποίο είναι διαχωρίσιμο από τα υπόλοιπα μέρη της ιστοσελίδας. Είναι δυνατή η ενσωμάτωση ενός τμήματος σε ένα άλλο. Η τμηματοποίηση μίας δυναμικώς παραγόμενης ιστοσελίδας ορίζεται από τον χρήστη με την δημιουργία προτύπων παρουσίασης στη γλώσσα υπερκειμένου HTML.

Για την κατανόηση της σύνθεσης μίας δυναμικής ιστοσελίδας από τα επιμέρους τμήματά της παρουσιάζεται το παράδειγμα μίας ιστοσελίδας από τον επίσημο ιστότοπο του γαλλικού πρωταθλήματος αντισφαίρισης. Στο σχήμαΣχήμα 3 Παράδειγμα δυναμικής ιστοσελίδας., στο οποίο απεικονίζεται η ιστοσελίδα μίας γνωστής αθλήτριας, διακρίνονται τα τμήματα της επικεφαλίδας (header) και της υποσέλιδο (footer), ο χώρος πλοήγησης αποτελούμενος απο υπερσυνδέσμους σε άλλες ιστοσελίδες, καθώς και ένα

σύντομο βιογραφικό σημείωμά της συνοδευόμενο από πρόσφατα αποτελέσματα αγώνων της.

Σχήμα 3 Παράδειγμα δυναμικής ιστοσελίδας.

Το βασικό σχεδιάγραμμα των ιστοσελίδων ορίζεται στον ακόλουθο κώδικα HTML. Ο επαρκής σχολιασμός του κώδικα υποδεικνύει στον αναγνώστη τα τμήματα της ιστοσελίδας.

```
<html>
<!-- %include header.frg -->
   <table>
      <tr>
         <td><!-- %include sidebr.frg --></td>
         <td><table>
```

```
            <tr><!-- %fragment graf_bio.frg --></tr>
            <tr><!-- %fragment graf_score.frg --></tr>
         </td></table>
      </tr>
   </table>
   <!-- %include footer.frg -->
   <!-- %fragment factoid.frg -->
   <!-- %fragment copyr.frg -->
</html>
```

Πίνακας 1 Κωδικας HTML

Η ευκολή υλοποίηση καθολικών αλλαγών σε όλες τις ιστοσελίδες ενός ιστοτόπου είναι δυνατή λόγω της τμηματοποίησης των δυναμικών ιστοσελίδων. Ως παράδειγμα αναφέρεται η αλλαγή του παρουσιαστικού των ιστοσελίδων του ιστοτόπου του γαλλικού πρωταθλήματος αντισφαίρισης, η οποία απαιτεί την τροποποίηση των τμημάτων "header.frg" και "footer.frg". Είναι εμφανές ότι ο ρυθμός τροποποίησης των τμημάτων μίας δυναμικής ιστοσελίδας ποικίλει. Οι πληροφορίες με στατικότερο περιεχόμενο, όπως το βιογραφικό σημείωμα του αθλητή στο συγκεκριμένο παράδειγμα, μπορούν να επικαιροποιηθούν σε ένα σημείο και να αποτελέσουν τμήματα σε μία ή περισσότερες ιστοσελίδες. Η αντικατάσταση της υπάρχουσας φωτογραφίας από μία νέα ή η εισαγωγή/διαγραφή πληροφοριών από το βιογραφικό σημείωμα της αθλήτριας απαιτεί την τροποποίηση του τμήματος "athlete_bio.frg". Συνήθως, το χρονικό διάστημα μεταξύ δύο τροποποιήσεων του τμήματος "stef_bio.frg" είναι μεγάλο συγκρινόμενο προς το αντίστοιχο διάστημα του τμήματος "stef_score.frg" στο οποίο παρουσιάζονται τα αποτελέσματα των αγώνων (πληροφορίες σχετικώς με το αποτέλεσμα ενός αγώνα τροποποιούνται συχνότερα στην περίπτωση κατά την οποία αυτός είναι σε εξέλιξη). Δεδομένου ότι ένα παιχνίδι ξεκινά (έστω ο τελικός αγώνας του γαλλικού πρωταθλήματος αντισφαίρισης), το τμήμα "athlete_score.frg" συμπεριλαμβάνει τα ακόλουθα δύο υπο-τμήματα

```
<!-- %fragment final_match.frg -->
<!-- %fragment semi_final_match.frg
-->
```

Καθώς ο τελικός αγώνας είναι σε εξέλιξη, μόνο το περιεχόμενο του τμήματος "final_match.frg" ενημερώνεται σε τακτά χρονικά διαστήματα. Επίσης το συγκεκριμένο τμήμα είναι δυνατόν να ενσωματωθεί σε περισσότερες ιστοσελίδες εκτός της ιστοσελίδας της αθλήτριας, π.χ., στην ιστοσελίδα της αντίπαλης αθλήτριας.

Στο Σχήμα 4 παρουσιάζεται ένα δεύτερο παράδειγμα δυναμικής ιστοσελίδας με πέντε διακριτά τμήματα. Δίνεται συνοπτική περιγραφή του περιεχομένου κάθε τμήματος εντός ενός πλαισίου.

Σχήμα 4

Τα τμήματα μίας ιστοσελίδας παρουσιάζουν διαφορές μεταξύ τους όσον αφορά το θέμα το οποίο διαπραγματεύονται, την λειτουργικότητά τους, το ρυθμό τροποποίησής τους καθώς και την πηγή προέλευσης του περιεχομένου τους. Επανερχόμενοι στο παράδειγμα της ιστοσελίδας του Σχήμα 4 το τμήμα των αποτελεσμάτων τροποποείται με διαφορετικό ρυθμό από το τμήμα του καταλόγου των μεταλλίων, το οποίο με τη σειρά του υπόκειται σε αλλαγές συχνότερα από το τμήμα στο οποίο παρουσιάζεται το ημερήσιο πρόγραμμα των αγώνων. Αντιθέτως, οι επιλογές πλοήγησης στην επικεφαλίδα καθώς και ο κατάλογος των υπερσυνδέσμων στο αριστερό τμήμα της

ιστοσελίδας αποτελούν τμήματα με σχετικά στατικό περιεχόμενο και είναι πιθανό να περιλαμβάνονται σε πολλές δυναμικές ιστοσελίδες του ιστοτόπου.

Μία ιστοσελίδα ενδέχεται να περιέχει πληροφορίες, οι οποίες παράγονται είτε από μία αυτόματη διαδικασία είτε από το συντάκτη του περιεχομένου. Χαρακτηριστικό παράδειγμα της πρώτης περίπτωσης αποτελεί η συνεχής ενημέρωση του τμήματος αποτελεσμάτων των αγώνων, η οποία πραγματοποιείται από μία αυτόματη διαδικασία, η έξοδος της οποίας δεν υπόκειται σε καμία διαδικασία επαλήθευσης της ορθότητας του περιεχομένου από το διαχειριστή της ιστοσελίδας. Σημειώνεται η απαίτηση της άμεσης επικαιροποίησης των τμημάτων με περιεχόμενο, το οποίο παράγεται με αυτόματο τρόπο, στην ανάπτυξη ιστοσελίδων με δεδομένα πραγματικού χρόνου. Τα παραπάνω εισάγουν ένα δεύτερο επίπεδο κατηγοριοποιήσης των τμημάτων μίας ιστοσελίδας με δυναμικό περιεχόμενο

1. *Τμήμα άμεσης δημοσίευσης (immediate fragment)*: Περιέχει πληροφορίες ζωτικής σημασίας, οι οποίες πρέπει να δημοσιευθούν το συντομότερο δυνατό. Συνήθως δεν προηγείται έλεγχος της δημοσίευσης από τον διαχειριστή.
2. *Τμήμα ελεχόμενης δημοσίευσης (quality controlled fragment)*: Τμήμα με δυναμικό περιεχόμενο του οποίου η δημοσίευση δεν επίγει. Το περιεχόμενο ενός τμήματος ελεγχόμενης δημοσίευσης υπόκειται σε έλεγχο από το διαχειριστή του ιστοτόπου πριν τη δημοσίευσή του.

Είναι προφανές από τα δύο προηγούμενα παραδείγματα ότι ο προσδιορισμός των τμημάτων με διαφορετική θεματολογία και λειτουργικότητα προϋποθέτει τη γνώση του αντικειμένου που διαπραγματεύεται το περιεχόμενο της δυναμικής ιστοσελίδας.

Στη βιβλιογραφία (3) προτείνονται αλγόριθμοι εντοπισμού των *ενδιαφερόντων τμημάτων* μίας ιστοσελίδας. Ενα τμήμα χαρακτηρίζεται ως ενδιαφέρον αν ικανοποιείται τουλάχιστον μία εκ των δύο επόμενων συνθηκών: (α) εμφανίζεται σε ένα πλήθος ιστοσελίδων του ιστοτόπου και (β) έχει διακριτά χαρακτηριστικά από τα υπόλοιπα τμήματα.

Αναλυτικότερα, δίνεται αναδρομικός ορισμός του συνόλου των εν δυνάμει ενδιαφερόντων τμημάτων μίας ιστοσελίδας.

i. Κάθε ιστοσελίδα ενός ιστοτόπου μπορεί να αποτελέσει ένα πιθανό ενδιαφέρον τμήμα.
ii. Ενα υπο-τμήμα ενός πιθανού ενδιαφέροντος τμήματος μπορεί να αποτελέσει πιθανό ενδιαφέρον τμήμα αν ικανοποιούνται μία εκ των δύο παρακάτω συνθηκών
    a. Το συγκεκριμένο υπό-τμήμα εμφανίζεται σε M το πλήθος πιθανά ενδιαφέροντα τμήματα, όπου $M > 1$.
    b. Το συγκεκριμένο υπο-τμήμα έχει διαφορετικά χαρακτηριστικά από το γονικό τμήμα στο οποίο περικλείεται, π.χ., διακριτή διάρκεια ζωής, διαφορετικό ρυθμό αλλαγής του περιεχομένου.

Τα πλεονεκτήματα της τμηματοποίησης των ιστοσελίδων συνοψίζονται στα ακόλουθα σημεία

- Δυνατότητα ενσωμάτωσης κοινών πληροφοριών σε πολλές ιστοσελίδες με ομοιόμορφο και αποδοτικό τρόπο.
- Ευκολία στο σχεδιασμό των ιστοσελίδων με τη δημιουργία ενός κοινού προτύπου παρουσίασης.
- Ευκολη διαχείριση συνόλων ιστοσελίδων τα οποία περιέχουν παρόμοιες πληροφορίες.
- Δυνατότητα δημιουργίας αποδοτικότερων μηχανισμών caching με άμεση συνέπεια την μείωση του χρόνου απόκρισης.

  Στις απομακρυσμένες caches αποθηκεύονται τα στατικά τμήματα των δυναμικών ιστοσελίδων. Κατά τη προετοιμασία της απάντησης, η cache στέλνει αίτημα προς τον εξυπηρετητή για την δημιουργία των τμημάτων με δυναμικό περιεχόμενο. Η κοινή πρακτική στη δημιουργία των ιστοσελίδων καταδεικνύει ότι τα τμήματα με δυναμικό περιεχόμενο αντιστοιχούν σε ένα μικρό ποσοστό των τμημάτων της ιστοσελίδας.

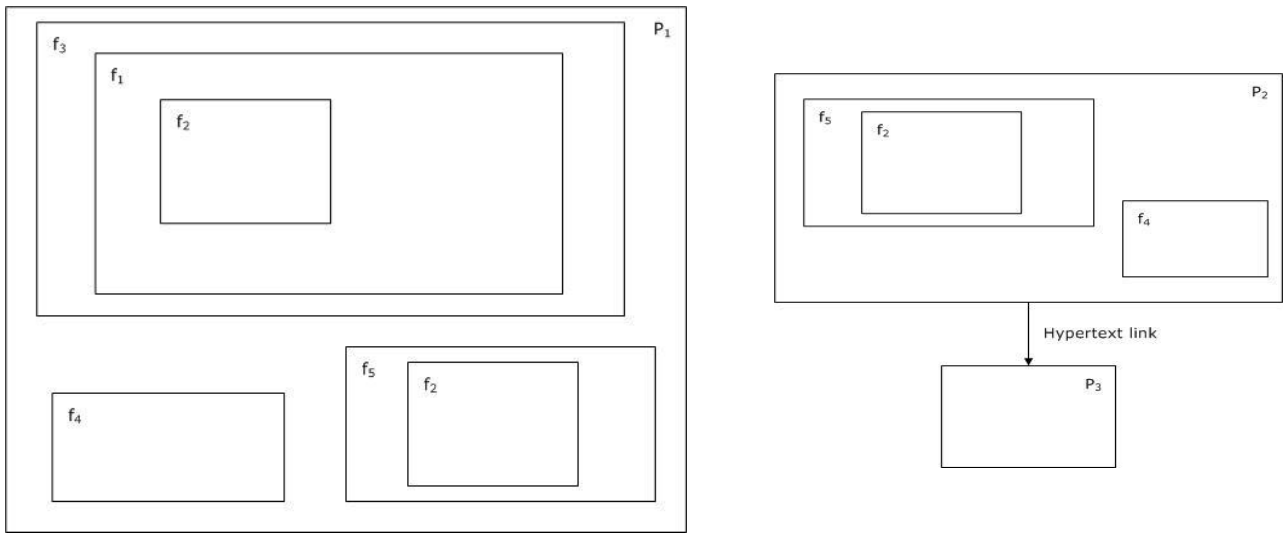
2.3.1 Μοντελοποιώντας τη δομή των δυναμικών ιστοσελίδων

2.3.1.1 Ο γράφος εξάρτησης των τμημάτων

Ο γράφος εξάρτησης των τμημάτων (object dependence graph) χρησιμοποιείται για την επισκόπηση με αφαιρετικό τρόπο της δομής μίας ιστοσελίδας. Στο συγκεκριμένο γράφο

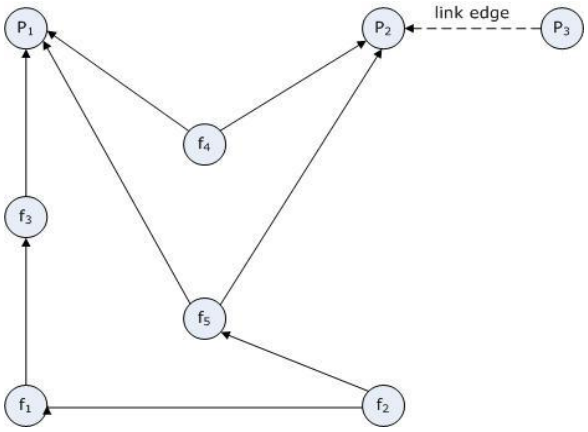
οι κόμβοι αντιστοιχούν στα τμήματα των ιστοσελίδων ενώ ορίζονται δύο κατηγορίες ακμών. Η ακμή σύνδεσης μεταξύ δύο τμημάτων δηλώνει ότι υπάρχει ένας υπεσύνδεσμος από το ένα τμήμα στο άλλο. Η ακμή ενσωμάτωσης δηλώνει ότι ένα τμήμα περικλείεται σε ένα γενικότετο τμήμα.

Στο σχήμα παρουσιάζεται με γραφικό τρόπο η δομή των ιστοσελίδων P_1, P_2 και P_3, ενώ το όνομα κάθε τμηματος ορίζεται από την συνένωνση του γράμματος “f” και ενός αύξοντα αριθμού. Σημειώνεται η ύπαρξη υπερσυνδέσμου στο περιεχόμενο της σελίδας P_2, ο οποίος «οδηγεί» στην ιστοσελίδα P_3.



Σχήμα 5 Τρεις ιστοσελίδες και τα τμήματά τους. Σημειώνεται η ύπαρξη μίας ακμής σύνδεσης μεταξύ των ιστοσελίδων P_2 και P_3.

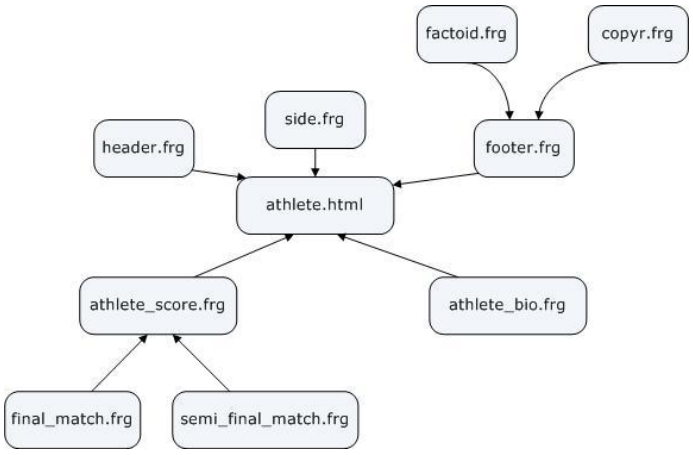
Ο γράφος εξάρτησης των τμημάτων για τις τρείς ιστοσελίδες δίνεται στο ακόλουθο σχήμα. Ολες οι ακμές εκτός μίας είναι ακμές ενσωμάτωσης.



Σχήμα 6 Ο γράφος εξάρτησης των τμημάτων

Η ακμή η οποία εκκινά από το τμήμα f_4 και προσπίπτει στο τμήμα P_1 δηλώνει ότι η ιστοσελίδα P_1 περιέχει το f_4. Η ακμή από την ιστοσελίδα P_3 στην P_2 είναι μία ακμή σύνδεσης και δηλώνει ότι η τελευταία περιέχει έναν υπερσύνδεσμο πρός την πρώτη.

Ο γράφος εξάρτησης των τμημάτων για την ιστοσελίδα του Σχήμα 3 Παράδειγμα δυναμικής ιστοσελίδας. είναι ο ακόλουθος



Σχήμα 7 Ο γράφος εξάρτησης των τμημάτων για την ιστοσελίδα του σχήματος

2.3.1.2 Μοντέλο DOM

Το μοντέλο αντικειμένων εγγράφου **D**ocument **O**bject **M**odel (DOM). είναι ένα πρότυπο του οργανισμού W3C [http://www.w3.org/DOM] το οποίο ορίζει μία διεπαφή προγραμματισμού εφαρμογών για HTML και XML αρχεία. Συνιστά ένα τρόπο αναπαράστασης της λογικής δομής των προαναφερθέντων αρχείων καθορίζοντας ένα προγραμματιστικό μοντέλο το οποίο είναι ανεξάρτητο της γλώσσας προγραμματισμού.

Η ύπαρξη προγραμματιστικής διεπαφής ανεξάρτητης της γλώσσας προγραμματισμού δίνει την δυνατότητα σε διάφορες γλώσσες σεναρίου (π.χ., JavaScript) να προσπελαύνουν τα τμήματα μίας ιστοσελίδας καθώς και να προσθέτουν, αφαιρούν τμήματα και να τροποποποιουν το περιεχόμενό τους. Η εφαρμογή του μοντέλου προϋποθέτει ότι ο κώδικας HTML και XML είναι καλώς ορισμένος. Η μετατροπή του υφιστάμενου κώδικα σε καλώς ορισμένο πραγματοποιείται με τη χρήση κατάλληλου προγραμματιστικού εργαλείου.

Στο μοντέλο DOM, η δομή του κώδικα HTML της ιστοσελίδας περιεχομένου μίας ιστοσελίδας παριστάνεται από μία ιεραρχική δομή δενδρικής μορφής. Π.χ., ο ακόλουθος κώδικας HTML

```
<P>Κείμενο παραγράφου</P>
```

οδηγεί στη δημιουργία δύο κόμβων στο δένδρο DOM: ένα στοιχείο `P` και ένα κόμβο κειμένου με περιεχόμενο ‘`Κείμενο παραγράφου`’. Το στοιχείο `P` περικλείει τον κόμβο κειμένου και συνεπώς αποτελεί γονικό κόμβο του δεύτερου στο δένδρο DOM. Κατ’ επέκταση το στοιχείο P είναι κόμβος – παιδί ενός άλλου στοιχείου στο οποίο περικλείεται, π.χ., ενός στοιχείου `DIV` ή του στοιχείου `body`. Εκτός των στοιχείων και των κόμβων κειμένου, ο τρίτος τύπος κόμβων, ο οποίος απαντάται σε ένα DOM δένδρο, είναι οι κόμβοι ιδιοτήτων. Στο ακόλουθο τμήμα κώδικα HTML

```
<P ALIGN="right">Κείμενο <B>παραγράφου</B></P>
```

Ο κόμβος `ALIGN` αποτελεί κόμβο – παιδί του στοιχείου `P` στο δένδρο DOM.

Η ρίζα του δένδρου DOM αντιστοιχεί στο στοιχείο `HTML`. Οι εσωτερικοί κόμβοι καθώς και τα φύλλα του δένδρου DOM προσπελαύνονται με την κλήση κατάλληλων προγραμματιστικών συναρτήσεων.

Δεν υπάρχει ακριβής αντιστοιχία μεταξύ των τμημάτων μίας δυναμικής ιστοσελίδας και της δομής της όπως αυτή ορίζεται σύμφωνα με το μοντέλο DOM. Ωστόσο, η αναπαράσταση της δομής μίας ιστοσελίδας με το μοντέλο DOM χρησιμοποιείται αφενός στον εντοπισμό με αλγοριθμικές μεθόδους των ενδιαφερόντων τμημάτων και αφετέρου στον προσδιορισμό του δυναμικού χαρακτήρα της δομής των δυναμικών ιστοσελίδων (βλ. Παράγραφο Μεταβολή της δομής της ιστοσελίδας). Στο επιστημονικό άρθρο (3)

προτείνεται αλγόριθμος για τον αυτόματο εντοπισμό τμημάτων σε μία δυναμική ιστοσελίδα στα πλαίσια του οποίου χρησιμοποιείται μία τροποποιημένη εκδοχή του ιεραρχικού μοντέλου DOM. Συγκεκριμένα από το δένδρο του μοντέλου DOM αφαιρούνται οι κόμβο, οι οποίοι αντιστοιχούν σε ετικέτες μορφοποίησης κειμένου (π.χ., <BIG>, <BOLD>, <I> κ.α.).

2.4 Μεταβολή του περιεχομένου της ιστοσελίδας

Εμπειρικές μελέτες σε δείγμα ιστοσελίδων του διαδικτύου αναφέρουν ότι ιστοσελίδες με υψηλά ποσοστά επισκεψιμότητας υφίστανται συχνότερα αλλαγές στο περιεχόμενό τους από ιστοσελίδες, οι οποίες δεν είναι δημοφιλείς. Ωστόσο δεν έχει παρατηρηθεί συσχέτιση μεταξύ του βαθμού αλλαγής του περιεχομένου μίας ιστοσελίδας και του ποσοστού επισκεψιμότητάς της. Επίσης, η συχνότητα αλλαγής του περιεχομένου συναρτάται του ονόματος περιοχής στο οποίο ανήκει η ιστοσελίδα, π.χ., .com, .edu. Η συχνότητα αλλαγής του περιεχομένου των κυβερνητικών και εκπαιδευτικών ιστοτόπων είναι σχετικά χαμηλή. Η εξήγηση αποδίδεται στο γεγονός ότι οι συγκεκριμένοι ιστότοποι παρέχουν συνήθως ποιοτικότερο και μονιμότερο περιεχόμενο, το οποίο απαιτεί επικαιροποίηση σε αραιά χρονικά διαστήματα. Δεν ισχύει το ίδιο για ιστοτόπους οι οποίοι παρέχουν υπηρεσίες ηλεκτρονικού εμπορίου καθώς και για ιστοτόπους εταιρικών οργανισμών.

Επίσης, η συχνότητα αλλαγής εξαρτάται της θεματολογίας των ιστοσελίδων. Ως χαρακτηριστικό παράδειγμα αναφέρονται ιστοσελίδες ενημερωτικού χαρακτήρα, (ιστοσελίδες αθλητικών γεγονότων, ιστολόγια, ειδησεογραφικές δικτυακές πύλες), των οποίων ο ρυθμός αλλαγής του περιεχομένου είναι υψηλότερος του μέσου όρου. Τέλος, οι εσωτερικές ιστοσελίδες ενός ιστοτόπου, των οποίων το μονοπάτι προσπέλασης ελαχίστου μήκους από την ιστοσελίδα πύλης του ιστοτόπου είναι μεγάλο, υπόκεινται σε αλλαγές σπανιότερα. Ωστόσο η έκταση της αλλαγής είναι μεγάλη όταν αυτή τελικώς πραγματοποιείται. Μία πιθανή εξήγηση του συγκεκριμένου προτύπου αλλαγής είναι ότι οι ιστοσελίδες του ιστοτόπου, οι οποίες βρίσκονται εγγύτερα της ιστοσελίδας πύλης (δηλαδή απαιτείται μικρός αριθμός μεταβάσεων για την προσπέλασή τους) και υφίστανται συχνότερα τροποποιήσεις στο περιεχόμενό τους, συνήθως περιέχουν υπερσυνδέσμους προς τις υπόλοιπες ιστοσελίδες (του ιστοτόπου).

2.4.1 Ποσοτικός προσδιορισμός του βαθμού αλλαγής περιεχομένου

Μία απλή προσέγγιση της ποσοτικοποίησης του βαθμού αλλαγής του περιεχομένου αποτελεί ο έλεγχος αθροίσματος (checksum) σε επίπεδο περιεχομένου των στιγμιοτύπων μίας ιστοσελίδας. Κύριο μειονέκτημα της συγκεκριμένης προσέγγισης είναι ότι δεν προσφέρει τη δυνατότητα ορισμού διαβαθμίσεων της υφιστάμενης αλλαγής. Αποδίδεται ο ίδιος βαθμός σε μικρές και μεγάλες αλλαγές, οι οποίες λαμβάνουν χώρα είτε στο περιεχόμενο είτε στη δομή της ιστοσελίδας.

Προσεγγίσεις οι οποίες βασίζονται στον υπολογισμό της διαφοράς του περιεχομένου μεταξύ των δομικών τμημάτων των στιγμιοτύπων μίας ιστοσελίδας έχουν προταθεί στη βιβλιογραφία (2). Ο συντελεστής Dice (Dice coefficient), ο οποίος ανήκει στη συγκεκριμένη κατηγορία προσεγγίσεων, ποσοτικοποιεί τον βαθμό ομοιότητας του περιεχομένου διαφόρων εκδόσεων της ιστοσελίδας. Ο μαθηματικός ορισμός του συντελεστή Dice δίνεται στην ακόλουθη σχέση

$$Dice(W_i, W_k) = 2 \cdot \frac{|W_i \cap W_k|}{|W_i| + |W_k|}$$

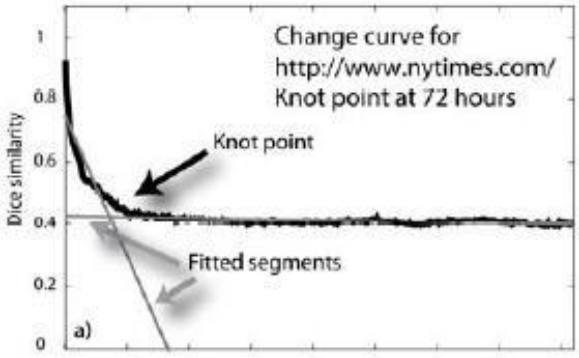
όπου W_i και W_k είναι τα σύνολα λέξεων του κειμένου των εκδόσεων i και k αντίστοιχα. Υψηλή τιμή του συντελεστή Dice, π.χ., $Dice(W_i, W_k) \approx 1$, δηλώνει μεγάλο βαθμό ομοιότητας μεταξύ των δύο διαφορετικών εκδόσεων. Αντιθέτως, η τιμή $Dice(W_i, W_k) = 0$ υποδηλώνει ότι οι εκδόσεις της ιστοσελίδας είναι ανόμοιες στον μέγιστο βαθμό. Χρησιμοποιώντας τον συντελεστή Dice ορίζεται μετρική για τον προσδιορισμό του βαθμού αλλαγής σε επίπεδο όρων, λέξεων (term). Συνεπώς είναι δυνατός ο προσδιορισμός των λέξεων και φράσεων, οι οποίες είναι κοινές στις διαφορετικές εκδόσεις της ιστοσελίδας και οι οποίες συνήθως είναι δηλωτικές του περιεχομένου (φράσεις κλειδιά).

2.4.2 Καμπύλες μεταβολής

Στις καμπύλες μεταβολής (change curve) του περιεχομένου μίας ιστοσελίδας αναπαριστάνεται με γραφικό τρόπο η τιμή του συντελεστή Dice στη διάρκεια του χρόνου. Συγκεκριμένα μετράται ο βαθμός αλλαγής του κειμένου μίας ιστοσελίδας σε σχέση με ένα σταθερό σημείο αναφοράς στο ιστορικό των εκδόσεων της ιστοσελίδας. Στο άρθρο (4) αναφέρεται η χρήση πέντε σημείων αναφοράς στον υπολογισμό της

μέσης τιμής του συντελεστή Dice, τα οποία αντιστοιχούν σε πέντε διαφορετικές αρχικές εκδόσεις της ιστοσελίδας. Αναλυτικότερα, η τιμή της καμπύλης μεταβολής σε καθε χρονικό σημείο t ισούται με τον μέσο όρο πέντε τιμών. Κάθε τιμή αντιστοιχεί στο βαθμό ομοιότητας του περιεχομένου της ιστοσελίδας την χρονική στιγμή t με το περιεχόμενο ενός εκ των πέντε αρχικών εκδόσεων, όπως αυτός προσδιορίζεται από το συντελεστή Dice.

Οι καμπύλες μεταβολής συνοψίζουν την εξέλιξη του περιεχομένου μίας ιστοσελίδας στο χρόνο. Στην ακόλουθη γραφική παράσταση παρουσιάζεται η μορφή της καμπύλης μεταβολής για τον ειδησεογραφικό ιστότοπο www.nytimes.com.

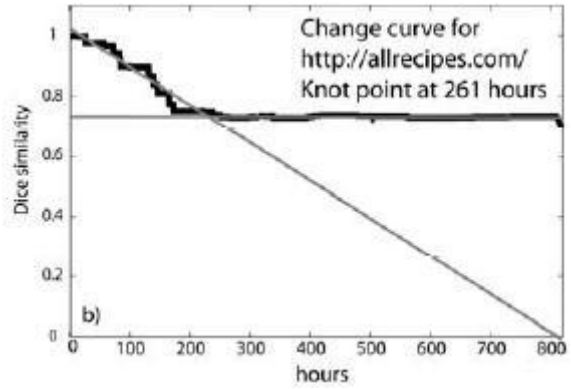


Σχήμα 8 Η καμπύλη μεταβολής του ιστότοπου www.nytimes.com

Στις περισσότερες των περιπτώσεων το περιεχόμενο των δυναμικών ιστοσελίδων διαφοροποιείται σε μεγάλο βαθμό εντός μικρού χρονικού διαστήματος από το περιεχόμενο της ιστοσελίδας αναφοράς. Η συγκεκριμένη διαπίστωση βρίσκει πλήρη εφαρμογή στην περίπτωση των ιστοσελίδων ειδησεογραφικού περιεχομένου (Σχήμα 8 Η καμπύλη μεταβολής του ιστότοπου www.nytimes.com), όπου άρθρα με περιορισμένο χρόνο παρουσίας απομακρύνονται από το κείμενο της ιστοσελίδας με συγκεκριμένο ρυθμό προκειμένου να προστεθούν νέα. Η υψηλή συχνότητα αλλαγής αποτυπώνεται με μία σχεδόν βηματική πτώση της καμπύλης μεταβολής. Πέραν κάποιας χρονικής στιγμής t' η καμπύλη γίνεται επίπεδη, και ο βαθμός ομοιότητας του περιεχομένου των επόμενων εκδόσεων για $t > t'$ και της αρχικής ιστοσελίδας έχει σχεδόν σταθερή τιμή. Στο σημείο κλισης t' της καμπύλης (όπου αυτή γίνεται επίπεδη) όλα τα άρθρα τα οποία ήταν παρόντα στην αρχική έκδοση της ιστοσελίδας έχουν απομακρυνθεί. Η σταθερή τιμή του συντελεστή Dice μετά τη χρονική στιγμή t' δηλώνει ότι κάθε μεταγενέστερη

έκδοση της ιστοσελίδας μετά το σημείο κλίσης περιέχει διαφορετικά άρθρα από την αρχική έκδοση. Η μη μηδενική τιμή του συντελεστή Dice για $t > t'$ οφείλεται αφενός στα τμήματα της ιστοσελίδας το περιεχόμενο των οποίων παραμένει αναλλοίωτο και αφετέρου στη διαχρονική παρουσία ορισμένων λέξεων και φράσεων. Παραδείγματα τμημάτων με σταθερό περιεχόμενο αποτελούν η επικεφαλίδα και το κατώτερο τμήμα της ιστοσελίδας, οι υπερσύνδεσμοι πλοήγησης στις υπόλοιπες ιστοσελίδες του ιστοτόπου κ.α. Εν γένει, ο μικρός βαθμός ομοιότητας μεταξύ του περιεχομένου των στιγμιοτύπων της ιστοσελίδας μετά το σημείο καμπής (knot point) και του περιεχομένου της αρχικής ιστοσελίδας αναφοράς αποδίδεται κυρίως στην ύπαρξη ενός καθολικού τρόπου παρουσίασης των ιστοσελίδων.

Η τιμή του σημείου καμπής εξαρτάται της θεματικής κατηγορίας στην οποία ανήκει η ιστοσελίδα. Η τιμή του σημείου καμπής για ιστοσελίδες ενημερωτικού χαρακτήρα, π.χ., τρέχουσα ειδησεογραφία, κάλυψη αθλητικών γεγονότων, είναι μικρή εφόσον το περιεχόμενό τους αντικαθίσταται συνήθως γρήγορα. Αντιθέτως, η καμπύλη μεταβολής ιστοσελίδων με λιγότερο δυναμικό περιεχόμενο φθίνει σταδιακά (βλ. Σχήμα 9).



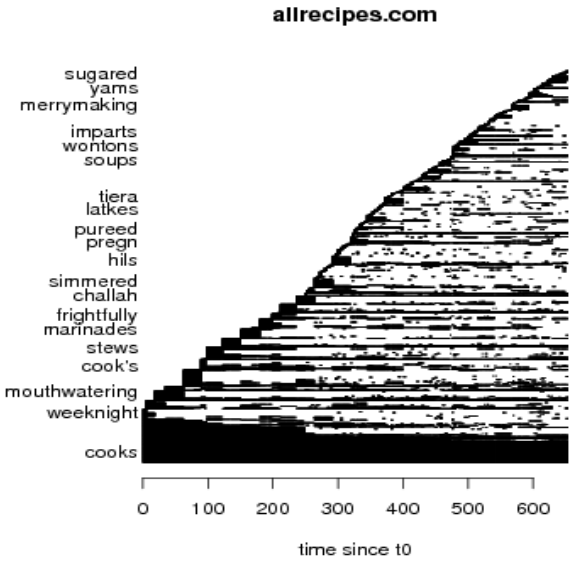
Σχήμα 9

2.4.3 Μεταβολή σε επίπεδο λέξεων

Η ανάλυση της παραγράφου «Μεταβολή του περιεχομένου της ιστοσελίδας» και οι καμπύλες μεταβολής εστιάζουν στον ρυθμό μεταβολής ολόκληρου του περιεχομένου μίας ιστοσελίδας αλλά όχι στον τρόπο μεταβολής του. Στην παρούσα παράγραφο παρουσιάζονται διάφοροι τρόποι χαρακτηρισμού της μεταβολής του περιεχομένου μίας ιστοσελίδας σε επίπεδο λέξεων. Η συγκεκριμένη προσέγγιση αποσκοπεί στην

εκτενέστερη μελέτη των μεταβολών περιεχομένου, όπως αυτές αποτυπώνονται στα διάφορα μέρη της καμπύλης μεταβολής.

Αναλυτικότερα, εξετάζεται η μεταβολή του λεξιλογίου, το οποίο χρησιμοποιείται στη σύνταξη του περιεχομένου μίας δυναμικής ιστοσελίδας. Το λεξιλόγιο συνίσταται από όρους, δηλ., λέξεις ή φράσεις. Στο Σχήμα 10 Γράφημα διάρκειας ζωής των όρων δίνεται παραδείγμα γραφήματος στο οποίο αποτυπώνεται η διάρκεια ζωής των όρων του λεξιλογίου (term lifespan plots) οι οποίοι χρησιμοποιούνται στη σύνταξη του περιεχομένου μίας ιστοσελίδας με αντικείμενο τις συνταγές μαγειρικής (www.allrecipes.com). Ο συγκεκριμένος τρόπος αναπαράστασης αποτελεί μία οπτική απεικόνιση του δυναμικού χαρακτήρα της εμφάνισης των όρων του λεξιλογίου στο κείμενο της ιστοσελίδας στη διάρκεια του χρόνου ζωής της. Ο χρόνος παρουσιάζεται κατά μήκος του άξονα X, ενώ οι οριζόντιες γραμμές στον άξονα Y αντιστοιχούν στους όρους του λεξιλογίου. Κάθε εμφάνιση ενός όρου στο κείμενο της ιστοσελίδας μία δεδομένη χρονική στιγμή t αντιστοιχει σε ένα διακριτό σημείο (στίγμα μαύρου χρώματος) στην οριζόντια γραμμή του όρου. Το γεγονός της απουσίας του όρου δεν αποτυπώνεται στην οριζόντια γραμμή. Οι όροι αρχικώς διατάσσονται με βάση την πρώτη εμφάνισή τους στο περιεχόμενο της ιστοσελίδας και κατόπιν με βάση τη διάρκεια ζωής τους.



Σχήμα 10 Γράφημα διάρκειας ζωής των όρων

Ουσιαστικά, η καμπύλη μεταβολής μίας ιστοσελίδας αποτελεί μία σύνοψη της πληροφορίας του γραφήματος που αναπαριστά τη διάρκειας ζωής των όρων του λεξιλογίου.

Γενικώς, η πυκνότητα των στιγμάτων στο γράφημα της διάρκειας ζωής των όρων είναι μεγαλύτερη στην κατώτερη περιοχή. Αυτό αποδίδεται στο γεγονός ότι οι πιο σταθεροί όροι σε μία δυναμική ιστοσελίδα τείνουν να εμφανίζονται στην αρχή της περιόδου παρατήρησης, ενώ οι πιο εφήμεροι όροι εμφανίζονται συνήθως αργότερα. Διαισθητικά, υπάρχουν δύο ενδεχόμενες λειτουργίες των όρων οι οποίοι βρίσκονται στην κατώτερη περιοχή του γραφήματος. Είτε εκφράζουν το κεντρικό νόημα του περιεχομένου είτε αντιστοιχούν σε στοιχεία πλοήγησης. Οσον αφορά τους όρους στην άνω περιοχή του γραφήματος έχουν είτε προσωρινό χαρακτήρα είτε μικρή συσχέτιση με το κεντρικό νόημα της ιστοσελίδας.

Εστω D_t συμβολίζει το περιεχόμενο της υπο μελέτη ιστοσελίδας τη χρονική στιγμή t και w ένας όρος του λεξιλογίου. Προκειμένου να προσδιοριστεί η πιθανότητα εμφάνισης του όρου w στη διάρκεια του χρόνου, ορίζεται το μέτρο της *δύναμης παραμονής (staying power measure)* $\sigma(w, D)$ του όρου w στο περιεχόμενο D της ιστοσελίδας

$$\sigma(w, D) = \sum_{t=0}^{T-1} \sum_{a=1}^{T-t} \frac{1}{T(T-a)} \cdot I\left(w \in D_t \bigwedge w \in D_{t+a}\right)$$

όπου $t \in [0 \dots T]$ είναι το χρονικό διάστημα παρατήρησης, ενώ με a συμβολίζεται το διάστημα μεταξύ δύο χρονικών παρατηρήσεων. $I(\cdot)$ είναι μία συνάρτηση δείκτης. Το μέτρο της δύναμης παραμονής εκφράζεται συναρτήσει πιθανοθεωρητικών μεγεθών ως ακολούθως

$$\sigma(w, D) \approx P(t)P(a)P(w \mid D_t, D_{t+a})$$

Σύμφωνα με την προηγούμενη σχέση, το μέτρο της δύναμης παραμονής του όρου w ισούται προσεγγιστικά με την πιθανότητα εμφάνισης του συγκεκριμένου όρου στο κείμενο της ιστοσελίδας τις χρονικές στιγμές t και $t + a$ αντίστοιχα. Οροι με χαμηλή τιμή του μέτρου της δύναμης παραμονής εμφανίζονται προσωρινά στο κείμενο της ιστοσελίδας.

Αντιθέτως, οροι με υψηλή τιμή του μέτρου είναι πιθανό να εμφανίζονται σε πολλαπλά στιγμιότυπα της ιστοσελίδας. Οι συγκεκριμένοι όροι ταξινομούνται σε τρεις κατηγορίες

- Οροι οι οποίοι είναι δηλωτικοί του περιεχομένου της ιστοσελίδας.
- Οροι οι οποίοι αντιστοιχούν σε ευρέως χρησιμοποιούμενες λέξεις, π.χ., λειτουργικές λέξεις, άρθρα, σύνδεσμοι.
- Οροι οι οποίοι διευκολύνουν την πλοήγηση του χρήστη στις ιστοσελίδες του ιστοτόπου.

Προκειμένου να προσδιοριστούν οι όροι της πρώτης κατηγορίας εξετάζεται το μέτρο της απόκλισης ή συνάφειας (divergence measure) των συγκεκριμένων όρων σε σχέση με το σύνολο των όρων του λεξιλογίου. Στη βιβλιογραφία το μέτρο της απόκλισης χρησιμοποιείται για την μέτρηση του βαθμού διαχωρισμού ενός υποσυνόλου ιστοσελίδων συγκεκριμένης θεματολογίας από την συλλογή όλων των ιστοσελίδων. Εφαρμόζοντας το μέτρο της απόκλισης σε επίπεδο όρου είναι δυνατός ο εντοπισμός των όρων οι οποίοι είναι χαρακτηριστικοί του περιεχομένου μίας ιστοσελίδας. Αναλυτικότερα, προσδιορίζονται οι όροι οι οποίοι διακρίνουν διαχρονικά τη γλώσσα, που χρησιμοποιήθηκε στη σύνταξη του περιεχομένου της ιστοσελίδας από αυτή των υπολοίπων ιστοσελίδων της συλλογής.

Ο μαθηματικός ορισμός του μέτρου της απόκλισης του όρου w είναι ο ακόλουθος

$$Div(w, D) = P(w \mid D) \log \frac{P(w \mid D)}{P(w \mid C)}$$

Οι τιμές των πιθανοθεωρητικών μεγεθών $P(w \mid D)$ και $P(w \mid C)$ δίνονται από τις παρακάτω σχέσεις

$$P(w \mid D) = \frac{1}{T} \sum_{t=0}^{T} \frac{tf_{w;\, D_t}}{|D_t|} \qquad P(w \mid C) = \frac{1}{T} \sum_{t=0}^{T} \frac{ctf_{w;C_t}}{|C_t|}$$

όπου τα μεγέθη tf και ctf αναφέρονται στη συχνότητα εμφάνισης του όρου στο κείμενο της ιστοσελίδας (document term frequency) και στη συλλογή όλων των ιστοσελίδων (collection term frequency) αντίστοιχα. $|D|$ και $|C|$ είναι το μήκος του κειμένου και το μέγεθος της συλλογής αντίστοιχα. Το μετρήσιμο μέγεθος $P(w \mid D)$ εκφράζει την πιθανότητα εμφάνισης του όρου w στις διάφορες εκδόσεις της δυναμικής ιστοσελίδας,

ενώ το μέγεθος $P(w \mid C)$ αναφέρεται στην πιθανότητα εμφάνισης του συγκεκριμένου όρου στη συλλογή όλων των εκδόσεων όλων των ιστοσελίδων. Εξ' ορισμού, όροι με υψηλή πιθανότητα εμφάνισης στο περιεχόμενο μίας ιστοσελίδας και χαμηλή πιθανότητα εμφάνισης στο σύνολο των ιστοσελίδων δίνουν υψηλή τιμή του μέτρου της απόκλισης.

Επανερχόμενοι στην ιστοσελίδα των συνταγών μαγειρικής, στην αριστερό μέρος του Πίνακας 2 Κατηγοριοποίηση των όρων που εμφανίζονται στον ιστότοπο www.allrecipes.com δίνονται οι όροι οι οποίοι είναι σχετικοί με το περιεχόμενο της ιστοσελίδας και εμφανίζονται κατά την διάρκεια της περιόδου παρατήρησης. Το μέτρο της δύναμης παραμονής των όρων με τους τονισμένους χαρακτήρες είναι μικρό ($\sigma < 0.2$). Οι συγκεκριμένοι όροι του λεξιλογίου αντιστοιχούν σε λέξεις οι οποίες έχουν προσωρινή παρουσία στην ιστοσελίδα και συνεπώς δεν είναι δηλωτικοί του περιεχομένου της. Στο δεξιό μέρος του πίνακα παρατίθενται οι όροι των οποίων το μέτρο της δύναμης παραμονής είναι υψηλό ($\sigma > 0.8$).

www.allrecipes.com	
Οροι	$\sigma > 0.8$
recipes cooking recipe	*Οροι με υψηλή τιμή του μέτρου απόκλισης*
advice	
more	cooks
salads	cookbook
cookbook	ingredient
tips	desserts
sandwiches	digest
cooks	trusted
easy	

pork	
survey	
menus	*Οροι με χαμηλή τιμή του μέτρου απόκλισης*
bbq	
widgets	
cheese	
cool	home
search	you
free	search
your	free
with	your
rachael	com
digest	with
newsroom	all
delivered	
cooks	
ingredient	
desserts	
this	

Πίνακας 2 Κατηγοριοποίηση των όρων που εμφανίζονται στον ιστότοπο www.allrecipes.com

Επιπρόσθετα, οι όροι με $\sigma > 0.8$ ταξινομούνται σε δύο κύριες υποκατηγορίες με κρίτηριο την τιμή του μέτρου απόκλισης, όροι με υψηλή και χαμηλή τιμή του μέτρου απόκλισης. Το μέτρο απόκλισης διαχωρίζει τις λειτουργικές λέξεις οι οποίες είναι σχεδόν

πάντα παρούσες στις ιστοσελίδες (χαμηλή τιμή του μέτρου απόκλισης) από του όρους οι οποίοι είναι δηλωτικοί του περιεχομένου (υψηλή τιμή του μέτρου απόκλισης).

Συνεπώς η χρήση των μέτρων της δύναμης παραμονής και απόκλισης είναι χρήσιμοι για τον εντοπισμό των λέξεων οι οποίοι χαρακτηρίζουν και διαχωρίζουν το περιεχόμενο μίας ιστοσελίδας από αυτό των υπόλοιπων ιστοσελίδων. Η παρούσα ανάλυση μπορεί να χρησιμοποιηθεί για τον προσδιορισμό των φράσεων κλειδιών μίας δυναμικής ιστοσελίδας, η γνώση των οποίων συμβάλλει στην επίτευξη καλύτερης θέση στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης. Για περισσότερες λεπτομέρειες σχετικώς με τη βέλτιστη χρήση των φράσεων κλειδιών στη σύνταξη του περιεχομένου μίας ιστοσελίδας για την καλύτερη κατάταξή της στους καταλόγους αποτελεσμάτων ο αναγνώστης παραπέμπεται στα κεφάλαια 5, 6 και 7.

2.5 Μεταβολή της δομής της ιστοσελίδας

Η μεταβολή του περιεχομένου (content change) και της δομής (structural change) προσδιορίζουν τον δυναμικό χαρακτήρα των ιστοσελίδων. Στην παρούσα παράγραφο θα εξεταστούν οι αλλαγές στο τρόπο παρουσίασης του περιεχομένου μίας ιστοσελίδας. Εκδόσεις μίας ιστοσελίδας, οι οποίες αντιστοιχούν σε διαφορετικές χρονικές στιγμές, ενδέχεται να έχουν το ίδιο περιεχόμενο κειμένου, το οποίο όμως οργανώνεται και παρουσιάζεται με διαφορετικό τρόπο. Πειραματικές μελέτες αναφέρουν ότι συνήθως η δομή των ιστοσελίδων υπόκειται σε τακτικές αλλαγές μικρής έκτασης. Δραστική αλλαγή στη δομή των ιστοσελίδων λαμβάνει χώρα ανά δίμηνο.

2.5.1 Προσδιορισμός του βαθμού μεταβολής

Η ανάλυση των δομικών αλλαγών των ιστοσελίδων πραγματοποιείται με την εξέταση του μοντέλου των αντικειμένων του εγγράφου. Οπως αναφέρθηκε το συγκεκριμένο μοντέλο είναι μία προγραμματιστική διεπαφή κατάλληλη για την αξιοποίηση πληροφοριών, οι οποίες αφορούν το περιεχόμενο, την δομή και την μορφοποίηση των ιστοσελίδων. Ο εντοπισμός των αλλαγών στη δομή μίας ιστοσελίδας περιλαμβάνει την εξέταση του κώδικα HTML της ιστοσελίδας και ειδικότερα των τύπων των ετικετών (tags). Η γλώσσα HTML περιλαμβάνει πενήντα διαφορετικούς τύπους ετικετών, μερικοί εκ των οποίων αφορούν την μορφοποίηση του κειμένου της ιστοσελίδας, π.χ., `<b>` και `<h1>`, ενώ άλλοι σχετίζονται με τη δομή της. Οι ετικέτες καθορισμού των τμημάτων

( ) και των πινάκων (<table>) αποτελούν τις ετικέτες που προσδιορίζουν τον τρόπο οργάνωσης και παρουσιάσης του περιεχομένου μίας ιστοσελίδας.

Η μελέτη του βαθμού μεταβολής της δομής προϋποθέτει την απομάκρυση των κόμβων από το δένδρο DOM, οι οποίοι αντιστοιχούν σε ετικέτες που δεν σχετίζονται με τον τρόπο παρουσίασης του περιεχομένου της ιστοσελίδας, π.χ., ετικέτες μορφοποίησης περιεχομένου (<font>, <basefont>, <b>, <i>, <h1>,...,<h6>, , <style>, , <u> και <s>). Επίσης απομακρύνονται όλοι οι κόμβοι <script> οι οποίοι εισάγουν κώδικα γλώσσας σεναρίου, π.χ., JavaScipt[1], καθώς και οι κόμβοι γραφικών <img>.

Ο βαθμός μεταβολής της δομής μίας ιστοσελίδας προσδιορίζεται με τη σύγκριση των τροποποιημένων δένδρων DOM των διαφορετικών εκδόσεών της, όπως αυτά προκύπτουν μετά την απομάκρυνση των προαναφερθέντων κόμβων. Συγκεκριμένα μετράται ο αριθμός των διαφορετικών κόμβων στη τροποποιημένη δενδρική δομή DOM. Ο αλγόριθμος σύγκρισης δέχεται στην είσοδό του δύο δενδρικές δομές τις οποίες προσπελαύνει με αναδρομικό τρόπο. Σε κάθε κόμβο του δένδρου, ο αλγόριθμος συγκρίνει την ετικέτα του κόμβου και το πλήθος των απογόνων του. Στην περίπτωση κατά την οποία η ετικέτα του κόμβου και ο αριθμός των απογόνων του είναι ίδιοι στα δύο υπο εξέταση δένδρα, ο αλγόριθμος συνεχίζει αναδρομικά. Στην περίπτωση ύπαρξης αναντιστοιχίας στον αριθμό των απογόνων ενός κόμβου υπάρχουν οι δύο ακόλουθες αλγοριθμικές επιλογές

- Ο αλγόριθμος θεωρεί ότι οι απόγονοι καθώς και τα υπόδενδρα τους είναι ανόμοιοι.

[1] Αναφέρεται ο μεγάλος βαθμός διείσδυσης της τεχνολογίας AJAX στην ανάπτυξη των νέων ιστοσελίδων, η οποία έχει ως κύριο στόχο την μεγιστοποίηση της εμπειρίας του χρήστη. Είναι σύνηθης η πρακτική της συμπερίληψης μεγάλων τμημάτων κώδικα JavaScript στο αρχείο HTML της ιστοσελιδας. Η λειτουργικότητα του κώδικα JavaScript προβλέπει συνήθως αλλαγές στη δομή της ιστοσελίδας. Ωστόσο, στην βιβλιογραφία δεν αναφέρεται σχετική εργασία για την μεταβολή της δομής των ιστοσελίδων δεδομένης της τεχνολογίας AJAX.

- Ο αλγόριθμος εξετάζει επιμέρους τους απόγονους του κόμβου για την ύπαρξη ομοιότητας. Στην περίπτωση δύο όμοιων απογόνων πραγματοποιείται αναδρομική κλήση του αλγορίθμου στα υπόδενδρα με ρίζα τον όμοιο κόμβο.

Σημειώνεται ότι ο συγκεκριμένος αλγόριθμος σύγκρισης αποδίδει μεγάλο αριθμό ανόμοιων κόμβων όταν οι δομικές αλλαγές λαμβάνουν χώρα σε κόμβους οι οποίοι βρίσκονται εγγύτερα της ρίζας του δένδρου DOM.

Η θέση του κόμβου στο τροποποιημένο δένδρο DOM, στον οποίο παρατηρήθηκε η αλλαγή δομής, προσδιορίζεται από δύο μετρικές: Αφενός την απόστασή του από τη ρίζα του δένδρου και αφετέρου την μέγιστη απόστασή του από τα φύλλα του δένδρου.

3 Αρχιτεκτονική των Μηχανών Αναζήτησης

Η αναφορά στην αρχιτεκτονική των μηχανών αναζήτησης και στα στοιχεία που τη συνθέτουν κρίνεται απαραίτητη για την κατανόηση της διαδικασίας βαθμολόγησης των ιστοσελίδων.

Η κατηγοριοποίηση των στοιχείων της μηχανής αναζήτησης γίνεται με κρίτηριο την εξάρτηση των λειτουργιών τους από το ερώτημα του χρήστη (user's query). Στις πρώτες τέσσερις υποενότητες γίνεται αναφορά στα στοιχεία των οποίων η λειτουργικότητα δεν επηρεάζεται από το ερώτημα του χρήστη. Κατόπιν στις δύο ακόλουθες υποενότητες παρατίθενται πληροφορίες σχετικές με την λειτουργεία των στοιχείων που εξαρτώνται της ερώτησης. Τέλος, δίνεται εποπτικό διάγραμμα στο οποίο παρουσιάζονται τα στοιχεία μίας μηχανής αναζήτησης καθώς και η αλληλεπίδρασή τους.

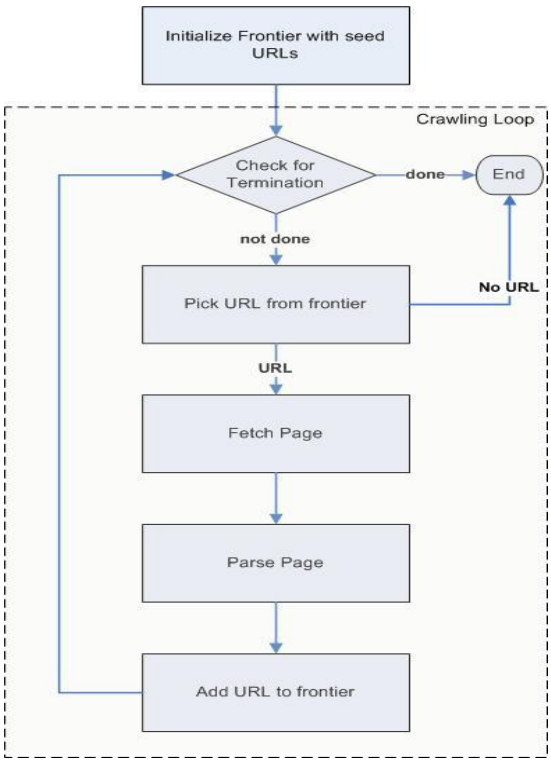
3.1 Ιχνηλάτηση του Παγκόσμιου Ιστού

Οι μηχανές αναζήτησης ιχνηλατούν τον Παγκόσμιο Ιστό προκειμένου να συλλέξουν τις ιστοσελίδες και να τις κατηγοριοποιήσουν. Ως ιχνηλάτης διαδικτύου (web crawler, web spider) ορίζεται το λογισμικό του οποίου η λειτουργικότητα αφενός είναι η περιήγηση στο Παγκόσμιο Ιστό με ένα μεθοδικό και αφετέρου η ανάκτηση ιστοσελίδων από αυτόν προς περαιτέρω επεξεργασία. Η περιήγηση μοντελοποιείται ως η επιλογή κόμβων σε ένα γράφο στον οποίο οι κόμβοι και οι ακμές αντιστοιχούν στις ιστοσελίδες και τους υπερσυνδέσμους μεταξύ των ιστοσελίδων[2] αντίστοιχα. Η επιλογή των κόμβων καθορίζεται με αλγοριθμικό τρόπο στο πρόγραμμα ιχνηλάτησης. Οι ιχνηλάτες δημιουργούν αντίγραφα όλων των επισκεπτόμενων ιστοσελίδων, τα οποία αποθηκεύονται προς μελλοντική επεξεργασία από την μηχανή αναζήτησης.

3.1.1 Η δομή ενός Προγράμματος Ιχνηλάτησης

Στο ακόλουθο σχήμα παρουσιάζεται το διάγραμμα ροής ενός προγράμματος ιχνηλάτησης.

[2] Ο αναγνώστης παραπέμπεται στην παράγραφο 4.1 για περισσότερες λεπτομέρειες σχετικώς με την θεώρηση του παγκόσμιου ιστού ως γράφου.



Σχήμα 11 Ο βασικός αλγόριθμος εκτέλεσης ενός προγράμματος ιχνηλάτησης

Το πρόγραμμα διατηρεί μία δομή δεδομένων λίστας στην οποία αποθηκεύονται οι διευθύνσεις των ιστοσελίδων οι οποίες πρόκειται να ανακτηθούν κατά την διαδικασία της ιχνηλάτησης. Η λίστα αρχικοποιείται με ένα σύνολο διευθύνσεων οι οποίες ονομάζονται αρχικές διευθύνσεις (seed urls). Το σύνολο των αρχικών διευθύνσεων καθορίζεται είτε αλγοριθμικά είτε από τον προγραμματιστή. Κατά την εκτέλεση του προγράμματος ιχνηλάτησης διακρίνονται τα ακόλουθα βασικά βήματα

1. Επιλογή μίας διεύθυνσης ιστοσελίδας από τη λίστα των διευθύνσεων.
2. Ανάκτηση του περιεχομένου της ιστοσελίδας η οποία αντιστοιχεί στην επιλεγμένη διεύθυνση του προηγούμενου βήματος. Το περιεχόμενο λαμβάνεται από τον διακομιστή στον οποίο φιλοξενείται η συγκεκριμένη ιστοσελίδα μέσω του πρωτοκόλλου HTTP.
3. Το περιεχόμενο της ανακτώμενης ιστοσελίδας, το οποίο εκφράζεται σε γλώσσα υπερκειμένου HTML, αναλύεται συντακτικά και απομονώνονται οι υπερσύνδεσμοι. Σημειώνεται ότι στη γλώσσα HTML το γνώρισμα `href` της ετικέτας `<a>` λαμβάνει ως τιμή ένα συρμό χαρακτήρων ειδικής μορφοποίησης, ο οποίος αντιστοιχεί στη γενική

περίπτωση σε μία έγκυρη διεύθυνση μίας ιστοσελίδας του διαδικτύου[3]. Η διεύθυνση μίας ιστοσελίδας ως τιμή του γνωρίσματος `href` ορίζει έναν υπερσύνδεσμο μεταξύ δύο ιστοσελίδων, της *ιστοσελίδας πηγής* και της *ιστοσελίδας στόχου*[4].

4. Από το σύνολο των υπερσυνδέσμων του προηγούμενου βήματος επιλέγονται οι διευθύνσεις των ιστοσελίδων οι οποίες δεν έχουν ανακτηθεί προηγουμένως από το πρόγραμμα ιχνηλάτησης. Οι συγκεκριμένες διευθύνσεις εισάγονται στη λίστα (frontier list).

Τα προηγούμενα βήματα επαναλαμβάνονται για κάθε ανακτώμενη ιστοσελίδα και συνθέτουν τον βρόχο ιχνηλάτησης (crawling loop). Η επαναληπτική διαδικασία των βημάτων 1-4 τερματίζεται στην περίπτωση ικανοποίησης τουλάχιστον ενός εκ των δύο ακολούθων κριτηριών.

i. Εχει ανακτηθεί από τον Παγκόσμιο Ιστό ενας προκαθορισμένος αριθμός ιστοσελίδων

ii. Η λίστα των διευθύνσεων των ιστοσελίδων προς ανάκτηση είναι κενή.

Στο βήμα 1 η επιλογή της επόμενης διεύθυνσης καθορίζεται από τον τρόπο υλοποίησης της λίστας. Αναλυτικότερα, η λίστα των διευθύνσεων μπορεί να υλοποιηθεί ως

- *Λίστα `FIFO`*: Η συγκεκριμένη δομή χρησιμοποίειται όταν ο Παγκόσμιος Ιστός «ιχνηλατείται» με την μεθοδολογία της αναζήτησης κατά βάθος ενός γράφου. Η επόμενη διεύθυνση της ιστοσελίδας προς ανάκτηση επιλέγεται από την κεφαλίδα

[3] π.χ., `<a href=”http://www.ibm.com”>Web Site of IBM</a>.`

[4] Ως «ιστοσελίδα πηγή» (source webpage) νοείται η ιστοσελίδα στο περιεχόμενο της οποίας αναγράφεται η τιμή του γνωρίσματος `href`. Ο όρος «ιστοσελίδα στόχος» (target webpage) αναφέρεται στην ιστοσελίδα της οποίας η διεύθυνση αποτελεί την τιμή του γνωρίσματος `href` στην ιστοσελίδα πηγή. Σημειώνεται ότι από μία ιστοσελίδα πηγή ενδέχεται να εκκινούν πολλοί υπερσύνδεσμοι προς ένα σύνολο ιστοσελίδων στόχων. Το σχήμα ιστοσελίδα στόχος-ιστοσελίδα πηγή έχει αμφίσημο χαρακτήρα, εφόσον στη γενική περίπτωση κάθε ιστοσελίδα είναι συγχρόνως ιστοσελίδα πηγή και ιστοσελίδα στόχος.

της λίστας ενώ οι διευθύνσεις των υπερσυνδέσμων, οι οποίες προκύπτουν από την συντακτική ανάλυση της ανακτώμενης ιστοσελίδας, εισάγονται στην ουρά της λίστας.

- Ουρά προτεραιότητας (priority queue): Η ουρά προτεραιότητας υλοποιείται ως ένας δυναμικός πίνακας ο οποίος είναι ταξινομημένος βάσει ενός βαθμού σημαντικότητας των ιστοσελίδων που δεν έχουν ανακτηθεί ακόμα. Η ιστοσελίδα με τον υψηλότερο βαθμό σημαντικότητας ανακτάται πρώτη.

Στο άρθο (4) περιγράφεται με αναλυτικό τρόπο η αρχιτεκτονική ενός εργαλείου εξερεύνησης του Παγκόσμιου Ιστού το οποίο αναπτύχθηκε για ακαδημαϊκούς σκοπούς στο Πανεπιστήμιο του Stanford. Η συγκεκριμένη υλοποίηση προβλέπει ότι στην περίπτωση κατά την οποία δεν είναι δυνατή η ανάκτηση μίας ιστοσελίδας λόγω ύπαρξης προβλημάτων στο δίκτυο ή στο διακομιστή, η διεύθυνση επαναεισάγεται στη λίστα των διευθύνσεων προς ανάκτηση. Μετά από ένα συγκεκριμένο αριθμό ανεπιτυχών προσπαθειών ανάκτησης μίας ιστοσελίδας προβλέπεται η διαγραφή της διεύθυνσής της από τη λίστα

Πέρα της λίστας των διευθύνσεων προς ανάκτηση, το πρόγραμμα ιχνηλάτησης διατηρεί μία επιπλέον λίστα στην οποία εισάγονται οι διευθύνσεις των ιστοσελίδες οι οποίες έχουν ανακτηθεί επιτυχώς. Σε κάθε στοιχείο της λίστας εκτός της διεύθυνσης της ανακτώμενης ιστοσελίδας κρατείται επιπλέον πληροφορία η οποία αφορά την ακριβή χρονική στιγμή ανάκτησης. Η συγκεκριμένη δομή δεδομένων αποθηκεύεται κατά προτίμηση στη κύρια μνήμη ώστε να είναι πιο αποδοτικός ο έλεγχος της ανάκτησης μίας ιστοσελίδας σε προηγούμενο βρόχο ιχνηλάτησης. Ο συγκεκριμένος έλεγχος έχει καθοριστική σημασία στη χρονική επίδοση του προγράμματος ιχνηλάτησης επειδή αποφεύγεται η πολλαπλή ανάκτηση της ίδιας ιστοσελίδας. Επίσης, μετά τον τερματισμό του προγράμματος η επεξεργασία της λίστας (π.χ., ταξινόμηση των στοιχείων της με κριτήριο το χρόνο ανάκτησης) δίνει πληροφορίες για την αξιολόγηση του αλγορίθμου ιχνηλάτησης.

3.1.2 Συντακτική Ανάλυση μίας Ιστοσελίδας

Η λειτουργικότητα των συντακτικών αναλυτών HTML έγκειται στον εντοπισμό ετικετών και των συσχετισμένων με αυτών χαρακτηριστικών. Στο πλαίσιο του προγράμματος της ιχνηλατησης μία ιστοσελίδα αναλύεται συντακτικά για την εξαγωγή των διευθύνσεων

των υπερσυνδέσμων, οι οποίες δίνονται ως τιμή στο γνώρισμα `href` των ετικετών `<a>`. Εκτός της εξαγωγής των διευθύνσεων η συντακτική ανάλυση περιλαμβάνει την κανονικοποίηση των διευθύνσεων, καθώς και την απομάκρυνση κατηγοριών λέξεων με μικρή σημασιολογική αξία όπως άρθρα, σύνδεσμοι κ.α.

Κανονικοποίηση των Διευθύνσεων

Η κανονικοποίηση (canonicalization) των διευθύνσεων ορίζεται ως η διαδικασία κατά την οποία διευθύνσεις οι οποίες εκ πρώτης όψεως είναι διαφορετικές και αντιστοιχούν στην ίδια ιστοσελίδα απεικονίζονται στην ίδια κανονική μορφή. Με την διαδικασία της κανονικοποίησης αποφεύγεται η πολλαπλή ανάκτηση από το παγκόσμιο ιστό της ίδιας ιστοσελίδας. Κατόπιν παρουσιάζονται με συνοπτικό τρόπο ορισμένες συνήθεις ενέργειες κανονικοποίησης των διευθύνσεων διαδικτύου.

- Μετατροπή των κεφαλαίων γραμμάτων σε μικρά, π.χ., η διεύθυνση `HTTP://www.IBM.com` μετατρέπεται σε `http://www.ibm.com`.
- Διαγραφή του τμήματος της διεύθυνσης το οποίο αναφέρεται σε ένα συγκεκριμένο σημείο της ιστοσελίδας, π.χ., η διεύθυνση `http://www.ibm.com/faq.html#what` μετατρέπεται σε `http://www.ibm.com/faq.html`.
- Κωδικοποίηση των ειδικών χαρακτήρων, όπως το κενό, '~', κ.α. Παραδειγματος χάρη οι διευθύνσεις `http://www.ibm.com/~pant/` και `http://www.ibm.com/%7pant/` απεικονίζονται στην ίδια ιστοσελίδα, εφόσον ο ειδικό χαρακτήρας '~' κωδικοποιείται ως '%7'.
- Οπου επιτρέπεται προστίθεται ο χαρακτήρας '/' στο τέλος της διεύθυνσης. Π.χ., η διεύθυνση `http://www.ibm.com` μετατρέπεται σε `http://www.ibm.com/`.
- Διαγραφή του συνδυασμού των χαρακτήρων '..' και του γονικού καταλόγου που προηγείται, π.χ., η διεύθυνση `http://www.ibm.com/%7pant/BizIntel/Seeds/../ODPSeeds.dat` μετατρέπεται στη διεύθυνση `http://www.ibm.com/%7pant/BizIntel/ODPSeeds.dat`.

Απομάκρυνση Λέξεων με Μικρή Εννοιολογική Αξία

Το σύστημα Dialog (`http://www.dialog.com`) αναγνωρίζει και απομακρύνει από το κείμενο της ανακτώμενης ιστοσελίδας εννέα λέξεις με μικρή σημασιολογική αξία

(stopwords), π.χ., “an”, “and”, “by”, “for”, “from”, “of”, “the”, “to” και “with”. Επίσης παρέχει τη δυνατότητα αντιστοίχισης παράγωγων λέξεων στη κοινή ρίζα τους. Π.χ., οι λέξεις “connect”, “connected” και “connection” αντικαθίστανται από τη λέξη “connect”.

3.1.3 Παραλληλοποίηση

Υπάρχουν σημαντικά περιθώρια βελτίωσης του χρόνου εκτέλεσης του βρόχου ιχνηλάτησης και κατ’ επέκταση της διαδικασίας ιχνηλάτησης του Παγκόσμιου Ιστού όπως αυτή παρουσιάστηκε στην παράγραφο 3.1.1. Συγκεκριμένα, η κεντρική μονάδα επεξεργασίας, η οποία εκτελεί τις εντολές του προγράμματος ιχνηλάτησης παραμένει ανένεργη, κατά τη διάρκεια ανάκτησης μίας συγκεκριμένης ιστοσελίδας από το διαδίκτυο. Αντίστοιχα, η διεπαφή του προγράμματος ιχνηλάτησης με τον δίκτυο παραμένει ανενεργή κατά την επεξεργασία του περιεχομένου μίας ανακτώμενης ιστοσελίδας. Στην παρούσα ενότητα αναλύονται οι ακόλουθοι δύο μέθοδοι παραλληλισμού της διαδικασίας ιχνηλάτησης.

1. Υιοθέτηση αρχών πολυνηματικού προγραμματισμού.
2. Ορισμός του ιστοτόπου ως μονάδας παραλληλισμού.

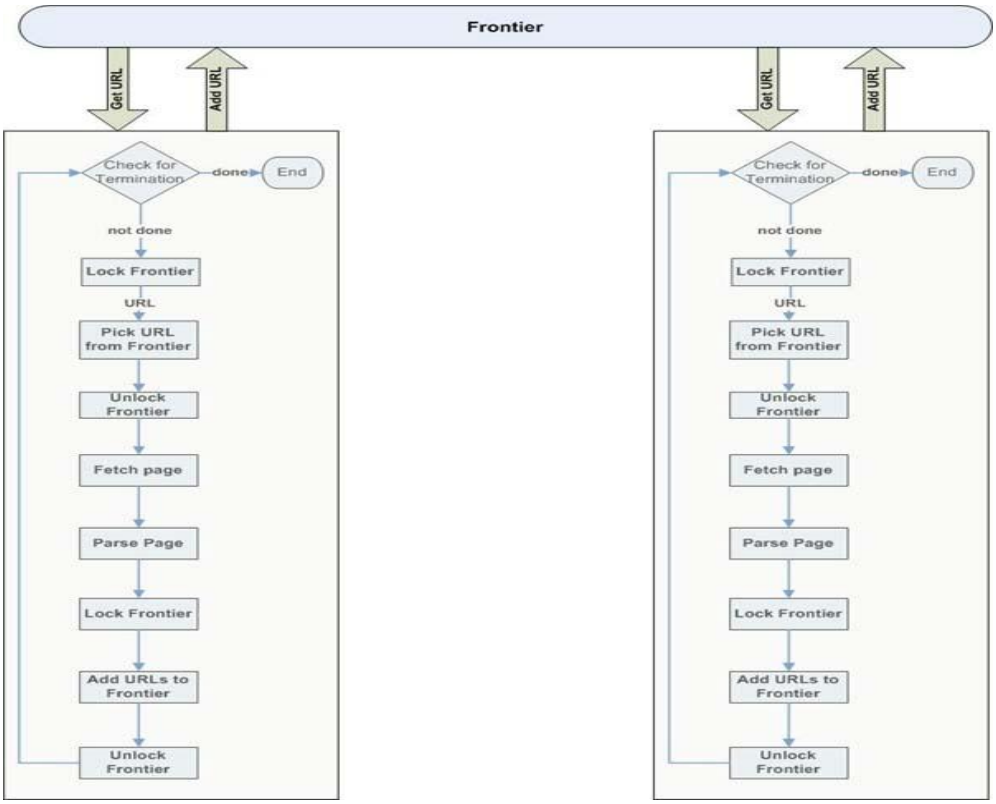
3.1.3.1 Πολυνηματική εκτέλεση (multithread execution) του βρόχου ιχνηλάτησης

Κάθε νήμα, το οποίο σε υπολογιστικό επίπεδο παριστά μία διεργασία, εκτελεί ένα βρόχο ιχνηλάτησης. Μία κοινή περιοχή μνήμης την οποία καταλαμβάνει η λίστα διευθύνσεων των ιστοσελίδων πρός ανάκτηση διαμοιράζεται σε όλα τα νήματα. Κάθε νήμα-διεργασία επιλέγει μία διεύθυνση από τη λίστα των διευθύνσεων. Κατά την ανάγνωση μίας διεύθυνσης από μία διεργασία, αποκλείεται η χρήση της λίστας από τα υπόλοιπα νήματα. Αντιστοίχως, αποκλείεται η πρόσβαση των υπολοίπων διεργασίων στη λίστα κατά την εγγραφή νέων διευθύνσεων στη λίστα από μία διεργασία. Η μη ταυτόχρονη προσπέλαση της λίστας από διαφορετικές διεργασίες εγγυάται τον συγχρονισμό. Η συγκεκριμένη μέθοδος παραλληλισμού προβλέπει τη διατήρηση μίας δεύτερης κοινής λίστας στην οποία εισάγονται οι διευθύνσεις των ιστοσελίδων οι οποίες έχουν ανακτηθεί και επεξεργαστεί επιτυχώς από το σύνολο των διεργασιών.

Η πολυπλοκότητα της συνθήκης τερματισμού της διαδικασίας ιχνηλάτησης με τη συγκεκριμένη μέθοδο παραλληλισμου αποδίδεται στην ύπαρξη πολλών διεργασίων οι

οποίες προσπελαύνουν την κοινή λίστα διευθύνσεων. Αναλυτικότερα, αν κατά την ανάγνωση της λίστας των διευθύνσεων από μία διεργασία διαπιστωθεί ότι είναι κενή, δεν συνεπάγεται ο τερματισμός της διαδικασίας ιχνηλάτησης, εφόσον είναι δύνατο να προστεθούν νέοι υπερσύνδεσμοι από τις υπόλοιπες διεργασίες. Ενας προτεινόμενος τρόπος αντιμετώπισης προβλέπει την μετάβαση της διεργασίας σε κατάσταση αδράνειας στην περίπτωση προσπάθειας ανάγνωσης μίας κενής λίστας. Περιοδικά, η συγκεκριμένη διεργασία επαναδραστηριοποιείται και αποπειράται να προσπελάσει εκ νέου τη λίστα των διευθύνσεων. Σε κεντρικό επίπεδο διατηρούνται πληροφορίες κατάστασης για όλες τις διεργασίες. Η ιχνηλάτηση του παγκόσμιου ιστού τερματίζει όταν όλες οι διεργασίες βρίσκονται σε κατάσταση αδράνειας.

Στο ακόλουθο σχήμα παρουσιάζεται η συγκεκριμένη μέθοδος παραλληλισμού.



Σχήμα 12 Πολυνηματική εκτέλεση του βρόχου ιχνηλάτησης

3.1.3.2 Ο ιστότοπος ως μονάδα παραλληλισμού

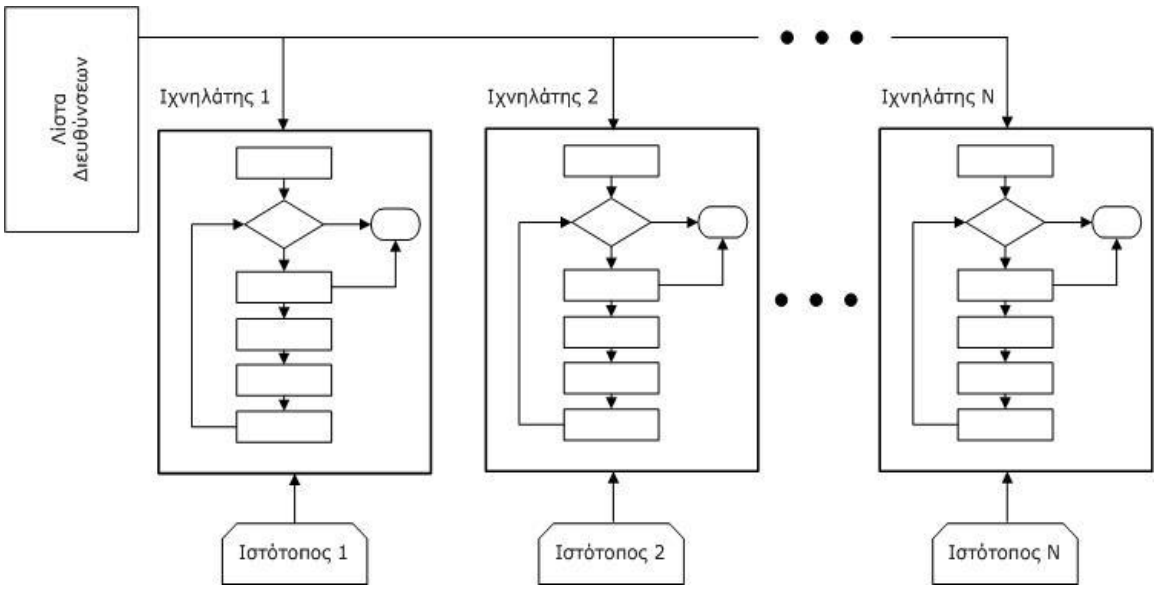
Κύριο μειονέκτημα της πολυνηματικής εκτέλεσης του βρόχου ιχνηλάτησης αποτελεί η ύπαρξη μίας διαμοιραζόμενης περιοχής μνήμης η οποία προσπελαύνεται από όλες τις

διεργασίες. Η συγκεκριμένη αρχιτεκτονική περιορίζει τον βαθμό παραλληλισμού, εφόσον δεν επιτρέπεται η ταυτόχρονη χρήση της λίστας. Στο άρθρο (4) προτείνεται μία μέθοδος παραλληλισμού στην οποία κάθε προγράμματα ιχνηλάτησης ανακτά ιστοσελίδες αποκλειστικά από έναν συγκεκριμένο ιστότοπο. Σημειώνεται ότι ένας ιστότοπος ορίζεται ως το σύνολο των ιστοσελίδων οι οποίες φέρουν το ίδιο συμβολικό όνομα εξυπηρετητή[5]. Στη γενική περίπτωση, κάθε πρόγραμμα ιχνηλάτησης εκτελείται σε διαφορετικό υπολογιστικό κόμβο[6].

Αρχικώς, δημιουργείται λίστα της οποίας τα στοιχεία αντιστοιχούν σε διευθύνσεις οι οποίες φέρουν αποκλειστικά συμβολικά ονόματα εξυπηρετητών. Η λίστα διευθύνσεων του προγράμματος ιχνηλάτησης αρχικοποιείται με την εισαγωγή μίας διεύθυνσης από την πρώτη λίστα. Κατόπιν κάθε πρόγραμμα ιχνηλάτησης εκτελείται εισάγοντας στη δική του λίστα διευθύνσεών του διευθύνσεις ιστοσελίδων οι οποίες ανήκουν στον ίδιο ιστότοπο και δεν έχουν ανακτηθεί ακόμα. Σχηματικά,

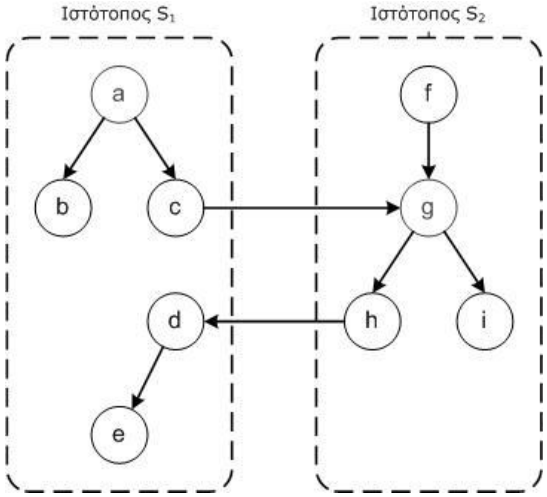
[5] Π.χ., Οι ιστοσελίδες `htpp://www.whitehouse.gov/privacy.html` και `htpp://www.whitehouse.gov/kids` έχουν το ίδιο συμβολικό όνομα εξυπηρετητή `htpp://www.whitehouse.gov`.

[6] Ενας υπολογιστικός κόμβος αποτελείται από την κεντρική μονάδα επεξεργασίας και τις διατάξεις κύριας και δευτερεύουσας μνήμης.



Σχήμα 13 Παραλληλοποίηση της διαδικασίας ιχνηλάτησης

Εφόσον οι υπερσύνδεσμοι οι οποίοι οδηγούν σε ιστοσελίδες εκτός του ιστοτόπου δεν εισάγονται στη λίστα διευθύνσεων των προγραμμάτων ιχνηλάτησης, ενδέχεται μερικές ιστοσελίδες αν μην ανακτηθούν ποτέ. Στο ακόλουθο σχήμα αποτυπώνεται η συγκεκριμένη περίπτωση.



Σχήμα 14

Οι ιστοσελίδες d και e οι οποίες ανήκουν στον ιστότοπο S_1 δεν είναι προσπελάσιμες από καμία άλλη ιστοσελίδα του συγκεκριμένου ιστοτόπου, εκτός της ιστοσελίδας h, η οποία ανήκει στον ιστότοπο S_2. Η ανάκτηση των συγκεκριμένων ιστοσελίδων γίνεται με την ανάλση του στιγμιοτύπου του Παγκόσμιου Ιστού, όπως έχει προκύψει από

προηγούμενη διαδικασία ιχνηλάτησης. Ειδικότερα, αν από την ανάλυση προκύψει η ύπαρξη υπερσυνδέσμου προς ιστοσελίδα, η οποία δεν έχει ανακτηθεί, τότε η διεύθυνση της συγκεκριμένης ιστοσελίδας εισάγεται στη αρχική λίστα διευθύνσεων της επόμενης διαδικασίας ιχνηλάτησης.

Τα κύρια πλεονεκτήματα της συγκεκριμένης μεθόδου παραλληλισμού συνοψίζονται στα ακόλουθα σημεία

- *Αποδοτικότερη διαχείριση της μνήμης*: Η διαδικασία ιχνηλάτησης περιορίζεται σε ένα συγκεκριμένο ιστότοπο αντί σε ολόκληρο τον παγκόσμιο ιστό. Συνεπώς, το μέγεθος των δομών δεδομένων είναι μικρό, εφόσον ο αριθμός των ιστοσελίδων ενός συγκεκριμένου ιστοτόπου είναι τάξεις μεγέθους μικρότερος του συνολικού αριθμού των ιστοσελίδων στο παγκόσμιο ιστό. Λόγω του μικρού μεγέθους των δομών δεδομένων αποφεύγεται η αποθήκευση τους στη δευτερεύουσα μνήμη, η οποία χαρακτηρίζεται από αργούς ρυθμούς αναζήτησης και επεξεργασίας.
- *Ανεξαρτησία των προγραμμάτων ιχνηλάτησης*: Ο μικρός απαιτούμενος βαθμός συντονισμού μεταξύ των προγραμμάτων ιχνηλάτησης επιτρέπει την κλιμάκωση του παραλληλισμού σε πολλούς υπολογιστικούς κόμβους.
- *Ευελιξία στη διαδικασία ιχνηλάτησης*: Η ικανοποίηση ενός αιτήματος ανάκτησης μίας ιστοσελίδας (το οποίο προέρχεται από το πρόγραμμα ιχνηλάτησης) συνεπάγεται αφενός επιπρόσθετο υπολογιστικό φόρτο για τον διακοσμιστή του ιστοτόπου και αφετέρου διάθεση μέρους του εύρους ζώνης της σύνδεσης. Οι διαχειριστές ιστοτόπων επιβάλλουν κανόνες στα προγράμματα ιχνηλάτησης προκειμένου να μην επηρεαστεί η δυνατότητα επισκεψιμότητας των ιστοσελίδων από τους φυσικούς χρήστες του διαδικτύου. Χαρακτηριστικό παράδειγμα κανόνα αποτελεί η εφαρμογή του πρωτοκόλλου αποκλεισμού (exclusion protocol). Το συγκεκριμένο πρωτόκολλο προβλέπει την αναζήτηση του αρχείου `robots.txt` σε συγκεκριμένη θέση στο σύστημα αρχείων του διακομιστή από το πρόγραμμα ιχνηλάτησης πριν την έναρξη της διαδικασίας ιχνηλάτησης των ιστοσελίδων του ιστοτόπου. Στο προαναφερθέν αρχείο παρατίθενται προθέματα διευθύνσεων, τα οποία αντιστοιχούν σε ομάδες ιστοσελίδων του ιστοτόπου οι οποίες δεν πρέπει να ανακτηθούν από τα προγράμματα ιχνηλάτησης. Ενα «φιλικό» πρόγραμμα ιχνηλάτησης λαμβάνει υπόψη τις εγγραφές του αρχείου `robots.txt`. Στην παρούσα μέθοδο παραλληλισμού, το

πρόγραμμα ιχνηλάτησης προσπελαύνει μία φορά το αρχείο `robots.txt` και βάσει του περιεχομένου του τροποποιεί με δυναμικό τρόπο τη διαδικασία ιχνηλάτησης.

3.1.4 Αξιολόγηση των προγραμμάτων Ιχνηλάτησης

Ενα πρόγραμμα ιχνηλάτησης αξιολογείται βάσει της ικανότητάς του να συλλέγει «καλές» ιστοσελίδες από τον Παγκόσμιο Ιστό. Η αξιολόγηση απαιτεί τον εκ των προτέρων ορισμό της έννοιας της «καλής» ιστοσελίδας. Ο ορισμός δίνεται στα πλαίσια του στοχευμένου προγραμμάτων ιχνηλάτησης των οποίων η είσοδος είναι η λίστα των αρχικών διευθύνσεων και ένα ερώτημα. Η διαδικασία της ιχνηλάτησης αξιολογείται με κριτήριο το βαθμό σχετικότητας των ανακτώμενων ιστοσελίδων με το ερώτημα (web page relevancy). Ο συγκεκριμένος τύπος προγράμματος ιχνηλάτησης χρησιμοποιείται στη ταξινόμηση των ιστοσελίδων σε θεματικές κατηγορίες. Η διαδικασία αξιολόγησης προβλέπει τον

i. ποσοτικό προσδιορισμό του βαθμού συσχέτισης του περιεχομένου μίας ιστοσελίδας με το τιθέμενο ερώτημα.
ii. προσδιορισμό της συνολικής απόδοσης βάσει του συνόλου των ανακτώμενων ιστοσελίδων.

Κατόπιν, παρατίθενται ορισμένοι μέθοδοι οι οποίοι έχουν προταθεί για τον ποσοτικό προσδιορισμό του βαθμού συσχέτισης του περιεχομένου μίας ιστοσελίδας με το ερώτημα.

1. Συχνότητα εμφάνισης των όρων του ερωτήματος στο κείμενο της ιστοσελίδας: Μία ιστοσελίδα θεωρείται σχετική αν εμφανίζονται σε αυτή ένα υποσύνολο ή το σύνολο των όρων-λέξεων του ερωτήματος. Για τον i-οστό όρο του ερωτήματος προσδιορίζεται η συχνότητα εμφάνισής του στο κείμενο της ιστοσελίδας ο οποίος δίνεται από το λόγο

$$\mathrm{tf}_i = \frac{n_i}{\sum_k n_k}$$

όπου ο αριθμητής του κλασματος ισούται με τον αριθμό των εμφανίσεων του συγκεκριμένου όρου στο κείμενο, ενώ ο παρονομαστής είναι το άθροισμα των εμφανίσεων όλων των όρων του ερωτήματος στο κείμενο.

2. *Μέτρηση του βαθμού ομοιότητας μεταξύ του κειμένου της ιστοσελίδας* p *και του ερωτήματος* q: Ο βαθμός ομοιότητας προσδιορίζεται με την μέθοδο του συνημιτόνου σύμφωνα με τη σχέση

$$\cos(q,p) = \frac{\mathbf{v_q} \cdot \mathbf{v_p}}{\|\mathbf{v_q}\| \cdot \|\mathbf{v_q}\|}$$

όπου $\mathbf{v_p}$ ($\mathbf{v_q}$) είναι διανυσματικό μεγέθος του οποίου η τιμή της συνιστώσας i ισούται με τη συχνότητα εμφάνισης του όρου i, tf_i, στο κείμενο της ιστοσελίδας (ερωτήματος, αντίστοιχα).

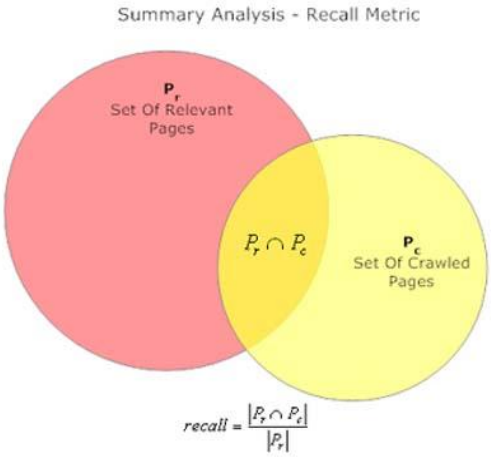
3. Μέτρηση του βαθμού ομοιότητας μεταξύ του κειμένου της ανακτώμενης ιστοσελίδας και των ιστοσελίδων, των οποίων οι διευθύνσεις έχουν εισαχθεί αρχικά στη λίστα διευθύνσεων του προγράμματος ιχνηλάτησης. Ο βαθμός ομοιότητας προσδιορίζεται με την μέθοδο του συνημιτόνου.

Δοθέντος του βαθμού συσχέτισης μίας ιστοσελίδας είναι δυνατή η αξιολόγηση της επίδοσης του στοχευμένου προγράμματος ιχνηλάτησης βάσει του συνόλου των ανακτώμενων ιστοσελίδων. Η αξιολόγηση περιλαμβάνει τον προσδιορισμό των μέτρων *ανάκλησης* και *ακρίβειας*.

Ως μέτρο ανάκλησης (recall metric) ορίζεται το κλάσμα των σχετικών ιστοσελίδων, οι οποίες ανακτήθηκαν τελικώς από το πρόγραμμα ιχνηλάτησης. Εστώ P_r το σύνολο των ιστοσελίδων με περιεχόμενο σχετικό πρός το ερώτημα του χρήστη και P_c το σύνολο των ιστοσελίδων οι οποίες ανακτήθηκαν τελικώς από το πρόγραμμα ιχνηλάτησης, τότε το μέτρο ανάκλησης δίνεται από το λόγο

$$\text{μετρο ανάκλησης} = \frac{|P_r \cap P_c|}{|P_c|}$$

όπου ∩ είναι το σύμβολο της τομής δύο συνόλων ενώ $|\cdot|$ δίνει την πληθυκότητα ενός συνόλου. Σχηματικά

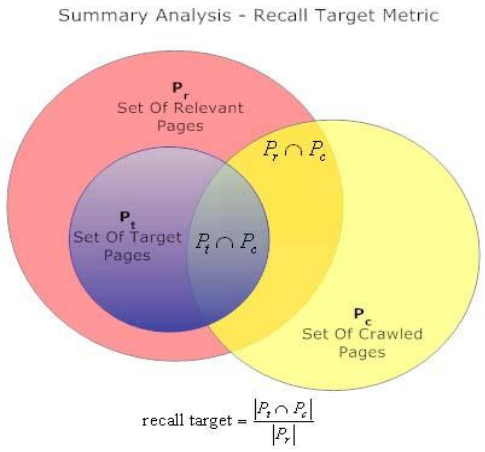


Σχήμα 15 Αναπαράσταση με διάγραμμα Venn

Η δυσκολία στον υπολογισμό του μέτρου ανάκλησης εντοπίζεται στο γεγονός το σύνολο των σχετικών ιστοσελίδων δεν είναι εκ των προτέρων γνωστό, συνέπεια του δυναμικού χαρακτήρα του διαδικτύου. Προκειμένου να δοθεί μία αριθμητική προσέγγιση του μέτρου της ανάκλησης, το άγνωστο σύνολο P_r αντικαθίσταται από το γνωστό υποσύνολό του P_t. Πλεόν η αποδοτικότητα του στοχευμένου προγράμματος ιχνηλάτησης καθορίζεται από τη δυνατότητά του να ανακτά σχετικές ιστοσελίδες οι οποίες ανήκουν στο γνωστό σύνολο P_t. Το εκτιμώμενο μέτρο ανάκλησης (recall target metric) δίνεται από το λόγο

$$\text{εκτιμώμενο μετρο ανάκλησης} = \frac{|P_t \cap P_c|}{|P_t|}$$

Διαγραμματικά,



Σχήμα 16 Αναπαράσταση με διάγραμμα Venn

Το μέτρο ακρίβειας ενός προγράμματος ιχνηλάτησης ορίζεται ως ο λόγος του πλήθους των ανακτώμενων ιστοσελίδων οι οποίες είναι σχετικές με το ερώτημα προς τον συνολικό αριθμό των ανακτώμενων ιστοσελίδων. Ενδεικτικά αναφέρονται οι ακόλουθες δύο μέθοδοι υπολογισμού του μέτρου ακρίβειας.

1. *Μέθοδος του ποσοστού συγκομιδής (harvest rate relevance score)*: Μία ιστοσελίδα είναι ή δεν είναι σχετική με το ερώτημα, δηλ., δεν ορίζονται διαβαθμίσεις σχετικότητας. Οπότε στην περίπτωση κατά την οποία q το πλήθος ιστοσελίδες είναι σχετικές σε σύνολο των p ανακτωμένων ιστοσελίδών, όπου $q \leq p$, τότε το ποσοστό συγκομιδής ισούται με q/p.
2. *Μέθοδος της ενδιάμεσης σχετικότητας (average relevance):* Ορίζονται βαθμοί συσχέτισης μεταξύ του ερωτήματος και του συνόλου των ανακτώμενων ιστοσελίδων. Το μέτρο ακρίβειας υπολογίζεται με χρήση του κριτηρίου ομοιότητας του συνημιτόνου. Συγκεκριμένα

$$\text{μέτρο ακρίβειας} = \frac{\sum_{p \in P_c} \cos(p, \boldsymbol{V})}{|P_c|}$$

όπου $\boldsymbol{V}$ είναι η διανυσματική αναπαράσταση του ερωτήματος και P_c είναι το σύνολο των ιστοσελίδων που έχουν ανακτηθεί από το πρόγραμμα ιχνηλάτησης.

3.2 Δεξαμενή Ιστοσελίδων

Η ανακτώμενη σελίδα από το πρόγραμμα ιχνηλάτησης αποθηκεύεται προσωρινά σε μία δεξαμενή ιστοσελίδων (page repository) εως ότου κατευθυνθεί ως είσοδος στο λογισμικό στο οποίο επιτελείται η επεξεργασία του περιεχομένου της. Στην πιο απλή περίπτωση κάθε ιστοσελίδα αποθηκεύεται σε ένα αρχείο με μοναδικό όνομα. Το όνομα του αρχείου προκύπτει με την εφαρμογή μίας συνάρτησης κατακερματισμού στη διεύθυνση της ιστοσελίδας. Συναρτήσεις κατακερματισμού, όπως `MD5` και `SHA1`, διασφαλίζουν με μεγάλη πιθανότητα την μοναδικότητα του ονόματος.

3.3 Επεξεργασία Ιστοσελίδας

Η επεξεργασία του περιεχομένου κάθε ιστοσελίδας συνίσταται στη δημιουργία μίας συμπιεσμένης μορφής των περιεχομένων της. Συγκεκριμένα, το λογισμικό το οποίο εκτελεί τη διαδικασία επεξεργασίας ανακτά μόνο την κρίσιμη πληροφορία από κάθε ιστοσελίδα δημιουργώντας με αυτό τον τρόπο μία συμπιεσμένη μορφή η οποία

αποθηκεύεται σε δομές δεικτών. Μετά την επεξεργασία της η ιστοσελίδα συνήθως απομακρύνεται από την δεξαμενή ιστοσελίδων. Ωστόσο ορισμένες δημοφιλείς ιστοσελίδες με ποιοτικό περιεχόμενο παραμένουν στη δεξαμενή ιστοσελίδων και μετά το τέλος της επεξεργασίας τους.

3.4 Ευρετήρια

Τα ευρετήρια (index) ορίζονται ως δομές δεδομένων οι οποίοι επιταχύνουν την αναζήτηση. Για τις ανάγκες των μηχανών αναζήτησης χρησιμοποιούνται οι ακόλουθοι τρείς τύποι δομών ευρετηρίου

- *Ευρετήριο Περιεχομένου (content index)*: Στη συγκεκριμένη δομή αποθηκεύεται σε συμπιεσμένη μορφή πληροφορίες για το περιεχόμενο της ιστοσελίδας, όπως φράσεις κλειδιά, ο τίτλος της κ.α. Για την υλοποίηση της συγκεκριμένης δομής χρησιμοποιείται η τεχνική των αντιστρόφων αρχείων. Η τεχνική των αντιστρόφων αρχείων προβλέπει τη δημιουργία των δομών του λεξιλογίου των λιστών, στις οποίες αποθηκεύονται πληροφορίες για την εμφάνιση κάθε όρου του λεξιλογίου. Σε κάθε όρο του λεξιλογίου αντιστοιχεί μία λίστα της οποίας κάθε στοιχείο φέρει πληροφορίες για τη εμφάνιση του συγκεκριμένου όρου σε μία ιστοσελίδα. Η ακρίβεια των αποτελεσμάτων μίας μηχανής αναζήτησης συναρτάται σε μεγάλο βαθμό από το πλήθος των αποθηκευμένων πληροφοριών σε κάθε στοιχείο της λίστας, όπως ο ακριβής αριθμός εμφανίσεων του όρου στο κείμενο της ιστοσελίδας, οι θέσεις των στιγμιοτύπων του όρου, κ.α.
- *Ευρετήριο δομής (structure index)*: Στη συγκεκριμένη δομή αποθηκεύεται σε συμπιεσμένη μορφή οι πληροφορίες σχετικώς με τη μορφή της δομής του γράφου του διαδικτύου, ο οποίος αντιστοιχεί στο μέρος του παγκόσμιου ιστού όπως έχει ανακτηθεί από το λογισμικό ιχνηλάτησης της μηχανής αναζήτησης. Η δομή του γράφου καθορίζεται από την ύπαρξη των υπερσυνδέσμων μεταξύ των ιστοσελίδων. Στο γράφο διαδικτύου $G(V,E)$, κάθε κόμβος $v \in V$ αντιστοιχεί σε μία ανακτώμενη ιστοσελίδα, ενώ κάθε ακμή $e \in E$ αντιπροσωπεύει έναν υπερσύνδεσμο από μία ιστοσελίδα πηγή σε μία ιστοσελίδα στόχο.

- *Ευρετήρια ειδικού σκοπού (special purpose index)*: Τα συγκεκριμένα χρησιμοποιούνται για την αναζήτηση αρχείων εικόνων, π.χ., αρχεία `jpg`, καθώς και αρχείων κειμένου ειδικής μορφοποίησης, π.χ., αρχεία με κατάληξη `doc` ή `pdf`.

3.5 Επεξεργασία του ερωτήματος του χρήστη

Η επεξεργασία του ερωτήματος του χρήστη (query processing module) περιλαμβάνει την μετατροπή της διατύπωσης από το επίπεδο της φυσικής γλώσσας, με το οποίο εκφράζεται ο χρήστης, σε μία γλώσσα η οποία είναι κατανοητή από την μηχανή αναζήτησης.

Στο άρθρο (5) ορίζεται μία σημασιολογική κατηγοριοποίηση των ερωτημάτων που υποβάλλονται στις μηχανές αναζήτησης από τους χρήστες. Στα παραδοσιακά συστήματα ανάκτησης πληροφοριών οι ερωτήσεις των χρηστών αφορούν αποκλειστικά την εύρεση αποθηκευμένων πληροφοριών ενημερωτικού χαρακτήρα, π.χ., τα στοιχεία επικοινωνίας ενός δημόσιου οργανισμού ή ενός φυσικού προσώπου. Ωστόσο, τα ερωτήματα προς τις μηχανές αναζήτησης δεν περιορίζονται μόνο στην αναζήτηση απλών πληροφοριών αλλά περιλαμβάνουν επίσης την εύρεση πληροφοριών πλοήγησης, π.χ., τη διεύθυνση μίας ιστοσελίδας, καθώς και ιστοσελίδων στις οποίες απαιτείται η συμπλήρωση στοιχείων από το χρήστη, π.χ., συμμετοχή σε ηλεκτρονική ψηφοφορία, διεκπεραίωση συναλλαγών ηλεκτρονικού εμπορίου, κ.α. Ακολουθεί συνοπτική περιγραφή των τριών κατηγοριών με χαρακτηριστικά παραδείγματα.

i. *Ερώτηση Ενημερωτικού χαρακτήρα (informational query)*: Ερωτήματα της συγκεκριμένης κατηγορίας υποβάλλονται στα παραδοσιακά συστήματα ανάκτησης πληροφοριών. Αφορούν την εύρεση πληροφοριών οι οποίες βρίσκονται σε στατική μορφή στο διαδίκτυο. Συγκεκριμένα, οι ιστοσελίδες, οι οποίες συνιστούν την απάντηση, δεν δημιουργούνται δυναμικά κατά την επεξεργασία του ερωτήματος. Καμμία επιπλέον ενέργεια δεν απαιτείται από τον χρήστη εκτός από την ανάγνωση της στατικής πληροφορίας.

ii. *Ερώτηση για την εύρεση της διεύθυνσης διαδικτύου συγκεκριμένης ιστοσελίδας (navigational query)*: Η απάντηση συνίσταται στην εύρεση της διεύθυνσης μίας συγκεκριμένης ιστοσελίδας, την οποία ο χρήστης του διαδικτύου είχε επισκεφθει στο παρελθόν ή της οποίας η ύπαρξη θεωρείται δεδομένη αλλά δεν είναι γνωστή η

διεύθυνσή της. Ακολουθούν χαρακτηριστικά παραδείγματα ερωτήσεων της συγκεκριμένης κατηγορίας καθώς και πιθανές απαντήσεις.

Ερώτημα	Πιθανή Απάντηση
Donald Knuth	http://www-cs-faculty.stanford.edu/~knuth/
Ενοικίαση Αυτοκινήτου	http://www.rent-a-car-in-greece.com/greek.htm
ΤΑ ΝΕΑ	http://www.tanea.gr/

Η επεξεργασία των ερωτημάτων της συγκεκριμένης κατηγορίας περιλαμβάνει μία «σωστή» απάντηση μη συμπεριλαμβανομένων των πιθανών συντακτικών ψευδωνύμων. Π.χ., η απάντηση στην ερώτηση «Καθημερινή» περιλαμβάνει μία από τις ακόλουθες διευθύνσεις, www.kathimerini.gr/ (δικτυακή πύλη ενημέρωσης) και http://www.ekathimerini.gr/ (ηλεκτρονική έκδοση καθημερινής εφημερίδας ευρείας κυκλοφορίας).

iii. *Ερώτηση για τη διεκπεραίωση συναλλαγών στο διαδίκτυο (transactional query):* Η απάντηση συνίσταται στην εύρεση της διεύθυνσης μίας ιστοσελίδας, με την οποία ο χρήστης του διαδικτύου αλληλεπιδρά για την ολοκλήρωση μίας συναλλαγής. Κύριες υποκατηγορίες είναι ερωτήσεις, οι οποίες αφορούν συναλλαγές ηλεκτρονικού εμπορίου, προσπέλαση βάσεων δεδομένων και διακομιστών ειδικού σκοπού (π.χ., εξυπηρετητές παιγνίων), εύρεση διαμεσολαβητικών υπηρεσιών του παγκόσμιου ιστού, κ.α..

Από την ανάλυση ενός τυχαίου επιλεγμένου συνόλου χιλίων (1000) ερωτήσεων προς τη μηχανή αναζήτηση AltaVista προκύπτει η κατανομή του ακόλουθου πίνακα. Σημειώνεται ότι κατά τη διαδικασία της κατηγοριοποίησης αγνοήθηκαν ερωτήσεις άσεμνου περιεχόμενου.

Κατηγορία Ερώτησης	Ποσοστό επί του συνόλου των κατηγοριοποιηθέντων ερωτήσεων (%)
Απλή Ενημέρωση	20

Πλοήγηση στο Διαδίκτυο	48
Διεκπεραίωση Συναλλαγών στο Διαδίκτυο	30

Το λογισμικό επεξεργασίας της ερώτησης ανατρέχει στο ευρετήριο περιεχομένου της μηχανής αναζήτησης για την απάντηση του ερωτήματος του χρήστη. Ενδεικτικά αναφέρεται η προσπέλαση της δομής του ευρετηρίου περιεχομένου για την ανάκτηση των ιστοσελίδων οι οποίες περιέχουν τους όρους του ερωτήματος. Οι ανακτώμενες ιστοσελίδες από το λογισμικό επεξεργασίας της ερώτησης καλούνται *σχετικές ιστοσελίδες* ως προς το ερώτημα και διοχετεύονται στο λογισμικό βαθμολόγησης.

3.6 Βαθμολόγηση-Ταξινόμηση Ιστοσελίδων

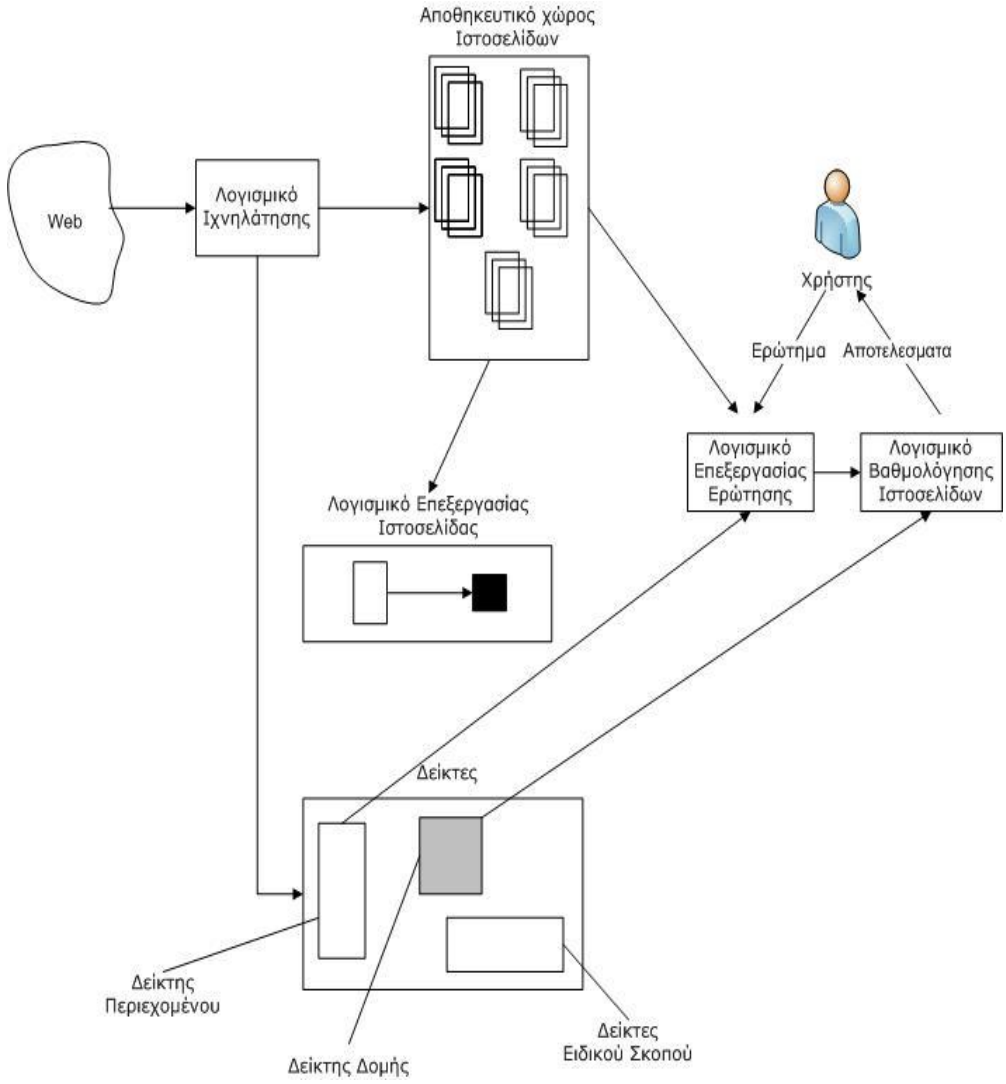
Η ταξινόμηση ως προς την σπουδαιότητα των σχετικών ιστοσελίδων πραγματοποιείται βάσει ορισμένων κριτηρίων. Η έξοδος του λογισμικού βαθμολόγησης (ranking module) είναι μία ταξινομημένη λίστα ιστοσελίδων, της οποίας τα στοιχεία, τα οποία βρίσκονται στην κορυφή της, έχουν την μεγαλύτερη πιθανότητα να ικανοποιούν της πληροφοριακές αναγκες του χρήστη, όπως αυτές εκφράστηκαν με το ερώτημά του προς την μηχανή αναζήτησης. Τις κατώτερες θέσεις της λίστας καταλαμβάνουν ιστοσελίδες οι οποίες χαρακτηρίζονται ως λιγότερο σχετικές προς το ερώτημα. Δεδομένου του όγκου πληροφορίας στον Παγκόσμιο Ιστό το μέγεθος του καταλόγου των αποτελεσμάτων ενδέχεται να είναι αρκετά μεγάλο ώστε να μην είναι αξιοποίησιμο από τον χρήστη εντός λογικών χρονικών πλαισίων. Συνεπώς γίνεται κατανοητή η κρισιμότητά του ρόλου του λογισμικού βαθμολόγησης στη διεκπεραίωση των ερωτημάτων του χρήστη.

Σε γενικές γραμμές, ο *συνολικός βαθμός (overall score)* ο οποίος αποδίδεται σε κάθε σχετική ιστοσελίδα προέρχεται από τον συνδυασμό δύο επιμέρους βαθμών, του *βαθμού περιεχομένου (content score)* και του *βαθμού δημοτικότητας (popularity score)*. Οι μηχανές αναζήτησης στα αρχικά στάδια ανάπτυξής τους απέδιδαν μεγαλύτερο βαθμό περιεχομένου στις σχετικές ιστοσελίδες, στις οποίες ο όρος του ερωτήματος εμφανίζοταν στην ετικέτα τίτλου ή στην μετα-ετικέτα του περιγραφικού κείμενο. Οσον αφορά τον βαθμό δημοτικότητας μίας ιστοσελίδας αυτός προκύπτει από την ανάλυση της δομής του γράφου $G(V,E)$ του Παγκόσμιου Ιστού. Ο βαθμός δημοτικότητας συνδυάζεται με το βαθμό περιεχόμενου και δίνεται ένας τελικός βαθμός σε κάθε σχετική

ιστοσελίδα. Τέλος, οι σχετικές ιστοσελίδες ταξινομούνται βάσει του τελικού βαθμού (από το υψηλότερο προς τον χαμηλότερο) και παρουσιάζονται στο χρήστη.

Η συνέπεια και η ορθότητα των αποτελεσμάτων διαμορφώνει ένα επίπεδο εμπιστοσύνης μεταξύ του χρήστη και της μηχανής αναζήτησης, η οποία καθορίζει σε σημαντικό βαθμό την κερδοφορία της τελευταίας. Συνεπώς, τεχνικές λεπτομέρειες, οι οποίες άπτονται της υλοποίησης των αλγορίθμων βαθμολόγησης ιστοσελίδων, περιβάλλονται από πλέγμα εμπιστευτικότητας.

Στο Σχήμα 17 παρουσιάζονται τα δομικά στοιχεία της μηχανής αναζήτησης.



Σχήμα 17 Τα δομικά στοιχεία μίας μηχανής αναζήτησης

4 Αλγόριθμοι Καθορισμού του Βαθμού Δημοτικότητας μίας Ιστοσελίδας

Στο παρόν κεφάλαιο θα εξεταστούν οι αλγόριθμοι PageRank και HITS, οι οποίοι αποτελούν τους κυριότερους αλγόριθμους καθορισμού του βαθμού δημοτικότητας των ιστοσελίδων σύμφωνα με την υπάρχουσα βιβλιογραφία. Ο βαθμός δημοτικότητας σε συνδυασμό με το βαθμό περιεχομένου μίας ιστοσελίδας καθορίζει την σειρά των ιστοσελίδων στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης. Η επιτυχία των συγκεκριμένων αλγορίθμων συνίσταται στο μεγάλο ποσοστό ικανοποίησης των πληροφοριακών αναγκών των χρηστών του διαδικτύου, όπως αυτές εκφράζονται με τις ερωτήσεις των τελευταίων προς τις μηχανές αναζήτησης.

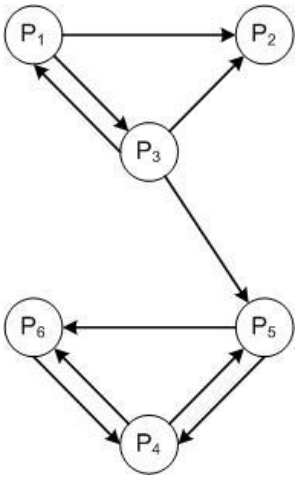
4.1 Ο γράφος διαδικτύου και η δομή του

Κοινό γνώρισμα των αλγορίθμων PageRank και HITS είναι η θεώρηση του Παγκόσμιου Ιστού ως κατευθυνόμενου γράφου. Ο *γράφος διαδικτύου (web graph)* $G(V, E)$ ορίζεται ως ακολούθως

- V είναι οι κόμβοι του γράφου, οι οποίοι αντιστοιχούν στο σύνολο των ιστοσελίδων.
- E είναι το σύνολο των ακμών του γράφου. Σε κάθε ακμή αντιστοιχεί ένας υπερ-σύνδεσμος από μία ιστοσελίδα σε μία άλλη. Κάθε ακμή περιγράφεται από το ζεύγος (u, v) όπου
  - u είναι ο κόμβος εκκίνησης της ακμής,
  - v ο κόμβος στον οποίο προσπίπτει η ακμή.

  Η ακμή με σημείο εκκίνησης τον κόμβο u καλείται *εξερχόμενος σύνδεσμος* ως προς τον συγκεκριμένο κόμβο. Αντιστοίχως, η ακμή, η οποία προσπίπτει στον κόμβο v καλείται *εισερχόμενος σύνδεσμος* για το συγκεκριμένο κόμβο.

Για την κατανόηση της μικροσκοπικής δομής του γράφου διαδικτύου δίνεται το ακόλουθο απλοϊκό σχήμα στο οποίο υποτίθεται ότι το πρόγραμμα ιχνηλάτησης ανέκτησε έξι ιστοσελίδες. Μεταξύ των ιστοσελίδων υπάρχουν σύνδεσμοι οι οποίοι μοντελοποιούνται από τις ακμές του γράφου. Σημειώνεται ότι η πληροφορία σχετικώς με τη δομή του γράφου είναι αποθηκευμένη στο ευρετήριο δομής της μηχανής αναζήτησης.



Σχήμα 18 Χαρακτηριστικό παράδειγμα γράφου διαδικτύου με 6 ιστοσελίδες

Ο γράφος του διαδικτύου G περιγράφεται από το ζεύγος (V, E) όπου

- $V = \{1,2,3,4,5,6\}$ και
- $E = \{(1,2), (1,3), (3,1), (3,5), (5,6), (4,6), (6,4), (4,5), (5,4)\}$

Ο παγκόσμιος ιστός παρομοιάζεται ως ένας περίπλοκος βιολογικός οργανισμός στον οποίο η μακροσκοπική δομή διαφέρει της δομής σε μικροσκοπική κλίμακα. Η κατανόηση της δομής του γράφου σε μακροσκοπικό επίπεδο προσφέρει πολύτιμες ιδέες για την αποδοτικότερη υλοποίηση των αλγορίθμων ιχνηλάτησης, κατηγοριοποίησης και βαθμολόγησης των ιστοσελίδων.

Προς την κατεύθυνση της διερεύνησης των μακροσκοπικών ιδιοτήτων του γράφου διαδικτύου πραγματοποιήθηκαν πειραματικές μελέτες σε σύνολο 200 εκατομμυρίων ανακτημένων ιστοσελίδων συνδεδεμένων με 1.5 δισεκατομμύρια υπερσυνδέσμους (6). Ανάλογες μελέτες μικρότερης έκτασης παρουσιάζεται στα άρθρα. Από την ανάλυση των πειραμάτων προκύπτει ότι το 90% των περίπου 200 εκτομμυρίων ιστοσελίδων σχηματίζουν μία *συνδεδεμένη συνιστώσα (connected component)* του γράφου διαδικτύου αν οι υπερσύνδεσμοι αναπαρασταθούν με ακμές χωρίς κατεύθυνση (μη κατευθυνόμενος γράφος διαδικτύου). Ως συνδεδεμένη συνιστώσα ενός μη κατευθυνόμενου γράφου ορίζεται το σύνολο των κόμβων για το οποίο για κάθε ζεύγος κόμβων u και v υπάρχει μονοπάτι (ακολουθία κόμβων) από το u στο v. Στη βιβλιογραφία οι συνδεδεμένες συνιστώσες του μη κατευθυνόμενου γράφου, ο οποίος έχει προκύψει από δεδομένο κατευθυνόμενο γράφο αγνοώντας την κατεύθυνση των

ακμών, ονομάζονται *αδύναμες συνδεδεμένες συνιστώσες (weakly connected component)*. Συνεπώς δύο ιστοσελίδες ενδέχεται να ανήκουν στην ίδια αδύναμη συνδεδεμένη συνιστώσα ακόμα κι αν δεν υπάρχει κατευθυνόμενο μονοπάτι που να τις συνδέει.

Επίσης εξετάστηκε η παρουσία και το μέγεθος των συνδεδεμένων συνιστωσών στο γράφο του διαδικτύου λαμβάνοντας υπόψη την κατεύθυνση των υπερσυνδέσμων μεταξύ των ιστοσελίδων (ισχυρές συνδεδεμένες συνιστώσες – strongly connencted components). Διαπιστώθηκε η ύπαρξη μία κύριας ισχυρής συνδεδεμένης συνιστώσας αποτελούμενη από 56 εκατομμύρια ιστοσελίδες. Το μέγεθος των υπολοίπων ισχυρών συνδεδεμένων συνιστωσών ήταν τάξεις μεγέθους μικρότερο. Για την εκτίμηση της μακροσκοπικής θέσης των υπολοίπων ιστοσελίδων οι οποίες ανήκουν στην γιγάντια αδύναμη συνδεδεμένη συνιστώσα αλλά όχι στην κύρια ισχυρή συνδεδεμένη συνιστώσα χρησιμοποιήθηκε ο αλγόριθμος της αναζήτησης κατά βάθος (first depth search)[7]. Αναλύοντας τα αποτελέσματα διακρίνονται τα ακόλουθα πέντε κύρια μέρη – σύνολα - στη μακροσκοπική δομή του γράφου. Σε κάθε μέρος δίνεται ένα χαρακτηριστικό όνομα.

- **SCC**: Αποτελεί τον κεντρικό πυρήνα του διαδικτύου. Αντιστοιχεί σε μία κύρια ισχυρή συνδεδεμένη συνιστώσα, η οποία είναι η «καρδιά» του Παγκόσμιου Ιστού. Ολες οι

[7] Σύμφωνα με το αλγόριθμο αναζήτησης κατά βάθος, η αναζήτηση ξεκινά από ένα κόμβο $u \in V$ του κατευθυνόμενου γράφου και προχωρά με τη δημιουργία συνόλων κόμβων οι οποίοι είναι προσπελάσιμοι από τον u. Κάθε σύνολο κόμβων ορίζει ένα επίπεδο αναζήτησης το οποίο χαρακτηρίζεται από το μήκος του μονοπατιού. Παραδείγματος χάριν το πρώτο επίπεδο (επίπεδο 1) αποτελείται από όλους τους κόμβους στους οποίους προσπίπτει ακμή με σημείο εκκίνησης τον κόμβο u (μονοπάτι μήκους 1). Το επίπεδο k αποτελείται από όλους τους κόμβους στους οποίους προσπίπτει ακμή η οποία εκκινά από κάποιον κόμβο του επιπέδου $k-1$. Εκ κατασκευής του αλγορίθμου, οι κόμβοι του επιπέδου k δεν ανήκουν σε κανένα προηγούμενο επίπεδο. Συνεπώς η έξοδος του αλγορίθμου συνιστά μία διαμέριση σε επίπεδα του συνόλου των κόμβων οι οποίοι είναι προσπελάσιμοι από τον κόμβο u.

ιστοσελίδες του συγκεκριμένου μέρους μπορούν να προσπελάσουν η μία την άλλη ακολουθώντας τους υπερσυνδέσμους.

- **IN**: Κάθε ιστοσελίδα προσπελαύνει μία ιστοσελίδα, η οποία ανήκει στο SCC μέρος του διαδικύου, μέσω μίας ακολουθίας υπερσυνδέσμων. Ωστόσο, δεν είναι εφικτό το αντίστροφο. Πρόκειται συνήθως για νέους ιστότοπους, των οποίων η ύπαρξη δεν έχει γνωστοποιηθεί στους δημιουργούς περιεχομένου των ιστοσελίδων που ανήκουν στον πυρήνα του διαδικτύου ώστε να προσθέσουν υπερσύνδεσμους προς αυτούς.
- **OUT**: Αποτελείται από ιστοσελίδες οι οποίες είναι προσιτές από το SCC μέρος του διαδικτύου. Ωστόσο, δεν είναι εφικτό το αντίστροφο. Χαρακτηριστικό παράδειγμα αποτελούν οι ιστοσελίδες εταιρικών ιστοτόπων, οι οποίοι περιέχουν εσωτερικούς υπερσυνδέσμους (υπερσυνδέσμους μεταξύ ιστοσελίδων που ανήκουν στον ίδιο ιστότοπο).
- **TENDRILS**: Στο συγκεκριμένο μέρος του διαδικτύου ανήκουν ιστοσελίδες οι οποίες είναι προσιτές από ιστοσελίδες του συνόλου IN, ενώ καμία από τις ενδιάμεσες ιστοσελίδες δεν ανήκουν στο σύνολο SCC. Επίσης, στο σύνολο TENDRILS ανήκουν ιστοσελίδες οι οποίες αποτελούν σημεία έναρξης κατευθυνόμενων μονοπατιών των οποίων οι κόμβοι τερματισμού αντιστοιχούν σε ιστοσελίδες του συνόλου OUT. Στην τελευταία περίπτωση κανένας από τους ενδιάμεσους κόμβους των μονοπατιών δεν ανήκει στην κύρια ισχυρή συνδεδεμένη συνιστώσα SCC.

  Σύμφωνα με τα προηγούμενα, είναι δυνατή η ύπαρξη «καναλιών» (tubes) μεταξύ ιστοσελίδων που ανήκουν στα σύνολα IN και OUT, ενώ καμία από τις ενδιάμεσες ιστοσελίδες δεν ανήκει στο σύνολο SCC

- **DISCONNECTED**: Ασύνδετα μέρη στο γράφο του διαδικτύου, των οποίων οι ιστοσελίδες δεν προσπελαύνονται από καμία ιστοσελίδα των συνόλων IN, OUT και SCC.

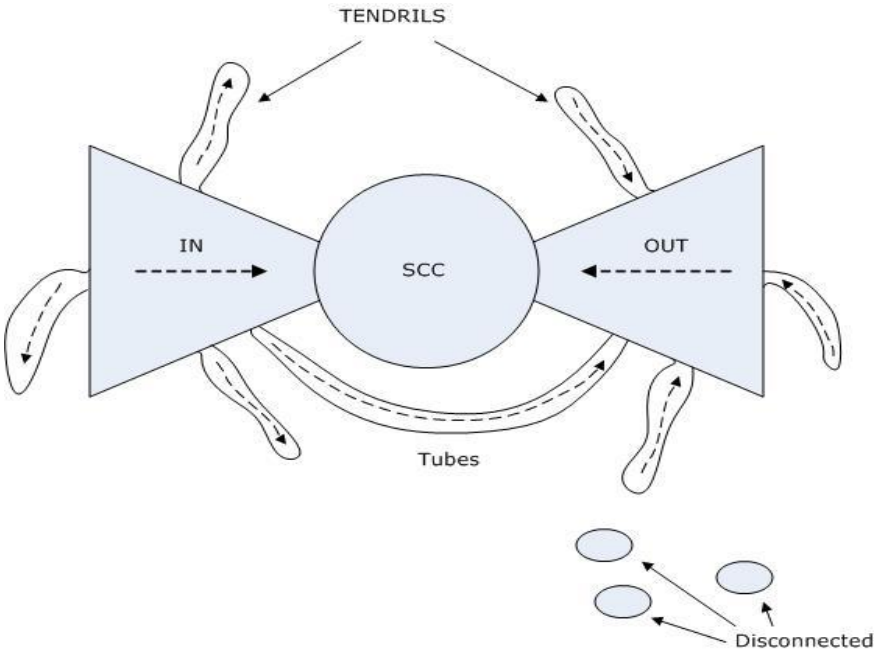
Το μέγεθος των προαναφερθέντων μερών του διαδικτύου για το σύνολο των 200 εκατομμύριων ανακτημένων ιστοσελίδων δίνεται στον ακόλουθο πίνακα

Μέρος - Σύνολο	**Μέγεθος** $(\times 10^6)$
SCC	56.46

Μέρος - Σύνολο	**Μέγεθος** $(\times 10^6)$
IN	43,34
OUT	43,17
TENDRILS	43,80
DISCONNECTED	16,78

Πίνακας 3

Η μακροσκοπική δομή του γράφου του διαδικτύου παρουσιάζεται στο ακόλουθο σχήμα.



Σχήμα 19 Μακροσκοπική δομή του γράφου διαδικτύου

4.2 Ο αλγόριθμος PageRank

Ο αλγόριθμος PageRank [39] προτάθηκε το 1998 από τους Sergei Brin και Larry Page. Εκτοτε, ο αλγόριθμος PageRank καθιερώθηκε ως ο καλύτερος αλγόριθμος καθορισμού του βαθμού δημοτικότητας μίας ιστοσελίδας. Χρησιμοποιείται από την μηχανή αναζήτησης της Google Η κεντρική ιδέα στην οποία βασίζεται ο συγκεκριμένος αλγόριθμος συνοψίζεται στην παρατήρηση ότι μία ιστοσελίδα κρίνεται ως σημαντική αν καταλήγουν σε αυτή σύνδεσμοι, οι οποίοι εκκινούν από ιστοσελίδες οι οποίες είναι

επίσης σημαντικές. Συγκεκριμένα, στη δημοσίευση των συγγραφέων αναφέρεται "*PageRank's thesis is that a webpage is important if it is pointed to by other important pages*".

Σε δύο σημεία συνοψίζονται τα κύρια χαρακτηριστικό του αλγορίθμου Pagerank. Πρωτίστως, μία προσεκτικότερη ανάλυση της κεντρικής ιδέας καταδεικνύει την ύπαρξη αναδρομής στον υπολογισμό του αλγορίθμου. Το δεύτερο χαρακτηριστικό αφορά τον καθολικό χαρακτήρα της βαθμολογίας του αλγορίθμου PageRank στις ιστοσελίδες του Παγκόσμιου Ιστού. Ο αποδιδόμενος βαθμός δημοτικότητας σε κάθε ιστοσελίδα είναι ανεξάρτητος του ερωτήματος του χρήστη και παραμένει σταθερός μέχρι την νέα εκτέλεση του αλγορίθμου. Συνεπώς, κατά τον χρόνο επεξεργασίας του ερωτήματος του χρήστη από την μηχανή αναζήτησης δεν δαπανάται υπολογιστικός χρόνος στον υπολογισμό του βαθμού δημοτικότητας της κάθε ιστοσελίδας. Σημειώνεται ότι η ανεξαρτησία των βαθμών δημοτικότητας των ιστοσελίδων από το ερώτημα του χρηστών δεν ισχύει στην περίπτωση του αλγορίθμου HITS (βλ. παράγραφο 4.3).

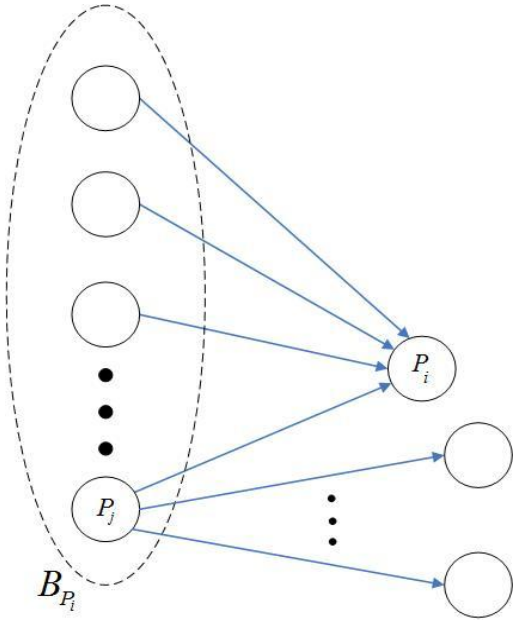
4.2.1 Μαθηματική Θεμελίωση

Στην ανάλυση που ακολουθεί υιοθετείται ο ακόλουθος συμβολισμός

- P_i: Μία ιστοσελίδα του Παγκόσμιου Ιστού,
- $r(P_i)$: Ο βαθμός δημοτικότητας της ιστοσελίδας P_i όπως αυτή δίνεται από τον αλγόριθμο,
- B_{P_i}: Το σύνολο των ιστοσελίδων οι οποίοι έχουν εξερχόμενους υπερσυνδέσμους (outlinks) προς την ιστοσελίδα P_i, και
- $|P_j|$: Ο αριθμός των εξερχομένων συνδέσμων της ιστοσελίδας P_j.

Σύμφωνα με την κεντρική ιδέα του αλγορίθμου PageRank ο βαθμός δημοτικότητας μίας ιστοσελίδας εξαρτάται από το βαθμό δημοτικότητας των ιστοσελίδων στο περιεχόμενο των οποίων υπάρχουν εξερχόμενοι υπερσύνδεσμοι προς αυτή. Υποθέτουμε ότι στο κείμενο της ιστοσελίδας P_j υπάρχουν $|P_j|$ το πλήθος εξερχόμενοι υπερσύνδεσμοι, ένας εκ των οποίων «δείχνει» προς την ιστοσελίδα P_i (βλ. Σχήμα 20). Η βασική εκδοχή του αλγορίθμου PageRank θεωρεί ότι ο βαθμός δημοτικότητας $r(P_j)$ της ιστοσελίδας P_j διαμοιράζεται με ομοιόμορφο τρόπο στους $|P_j|$ υπερ-συνδέσμους της. Συνεπώς, η

ύπαρξη του υπερ-συνδέσμου από την ιστοσελίδα P_j στην P_i αυξάνει την δημοτικότητα της τελευταίας κατά την ποσότητα $r(P_j)/|P_j|$. Τα ακριβώς ανάλογα ισχύουν για όλες τις ιστοσελίδες του συνόλου B_{P_i}.



Σχήμα 20

Η μαθηματική σύνοψη της προηγούμενης ανάλυσης δίνει την ακόλουθη σχέση για το βαθμό δημοτικότητας της ιστοσελίδας P_i

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|} \qquad 1$$

Ο υπολογισμός της ποσότητας $r(P_i)$ προϋποθέτει τη γνώση των ποσοτήτων $r(P_j)$, οι οποίες είναι επίσης ζητούμενες. Για να αντιμετωπιστεί το συγκεκριμένο πρόβλημα, οι ποσότητες $r(P_i)$ υπολογίζονται με την χρήση της επαναληπτικής μεθόδου. Αναλυτικότερα, ο βαθμός δημοτικότητας της ιστοσελίδας P_i στην $(k+1)$-οστή επανάληψη του αλγορίθμου, συμβολικά $r_{k+1}(P_i)$, υπολογίζεται βάσει των υπολογισθέντων τιμών της k-οστής επανάληψης (ακριβώς προηγούμενη επανάληψη), δηλαδή των ποσοτήτων $r_k(P_j)$. Η προηγούμενη σχέση τροποποιείται ελαφρώς για την εισαγωγή της επαναληπτικής διαδικασίας.

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \qquad 2$$

Η εκτέλεση μίας επαναληπτικής διαδικασία προϋποθέτει αφενός τον ορισμό αρχικών τιμών και αφετέρου την διατύπωση μίας συνθήκης τερματισμού. Η διαδικασία αρχικοποιείται με την απόδοση της τιμής $1/n$ στο βαθμό δημοτικότητας κάθε ιστοσελίδας P_i, δηλαδή $r_0(P_i) = 1/n$, όπου n είναι το πλήθος των ιστοσελίδων που μετέχουν στον υπολογισμό. Ο επαναληπτικός αλγόριθμος τερματίζει όταν ικανοποιούνται κριτήρια σύγκλισης (convergence criterion) των τιμών $r_k(P_j)$. Αν και θα υπάρξει εκτενής αναφορά σε επόμενη ενότητα στη σύγκλιση του αλγορίθμου PageRank και στις προϋποθέσεις που την εξασφαλίζουν, σημειώνεται ότι η επιλογή των αρχικών τιμών επηρεάζει μόνο τη ταχύτητα σύγκλισης[8] και όχι τα τελικά αποτελέσματα.

Η σχέση 2 υπολογίζει το βαθμό δημοτικότητας μίας ιστοσελίδας στην $(k+1)$-οστή επανάληψη του αλγορίθμου. Εισάγοντας συμβολισμό από τη θεωρία πινάκων είναι δυνατή η έκφραση του συνόλου των σχέσεων της μορφής 2 σε μία πιο συνεπτυγμένη μορφή. Συγκεκριμένα, εισάγεται το διάνυσμα $\boldsymbol{\pi}_k$, διάστασης $n \times 1$, όπου n είναι το πλήθος των ιστοσελίδων. Το στοιχείο i του διανύσματος $\boldsymbol{\pi}_k$, συμβολικά $\boldsymbol{\pi}_k(i)$, αντιστοιχεί στο βαθμό δημοτικότητας της ιστοσελίδας P_i στην k-οστή επανάληψη του αλγορίθμου. Η άθροιση των ποσοτήτων $r_k(P_j)/|P_j|$ στη σχέση 2 παραπέμπει στο εσωτερικό γινόμενο δύο διανυσμάτων. Αναλυτικότερα[9],

$$\boldsymbol{\pi}_{k+1}(i) = \sum \frac{r_k(P_j)}{|P_j|} = \sum r_k(P_j) \cdot \frac{1}{|P_j|} = \boldsymbol{\pi}_k^T \cdot \boldsymbol{h}_i$$

όπου $\boldsymbol{h}_i$ είναι διάνυσμα διάστασης $n \times 1$ και $\boldsymbol{h}_i(j) = 1/|P_j|$. Υπενθυμίζεται ότι στη σελίδα P_j υπάρχει εξερχόμενος υπερσύνδεσμος προς την σελίδα P_i, ενώ $|P_j|$ είναι ο αριθμός

[8] Ως ταχύτητα σύγκλισης ορίζεται το πλήθος των επαναλήψεων k του αλγορίθμου που απαιτούνται για την ικανοποίηση των κριτηρίων σύγκλισης.

[9] $\boldsymbol{A}^T$ συμβολίζει τον ανάστροφο του πίνακα $\boldsymbol{A}$.

των υπερ-συνδέσμων στη σελίδα P_j. Η τοποθέτηση του διανύσματος $\boldsymbol{h}_i$ στην i στήλη ενός πίνακα οδηγεί στο σχηματισμό του τετραγωνικού πίνακα $\mathbf{H}$ διάστασης $n \times n$. Εκ κατασκευής κάθε γραμμή του πίνακα $\mathbf{H}$, η οποία αντιστοιχεί σε ιστοσελίδα με ένα τουλάχιστον εξερχόμενο σύνδεσμο, είναι κανονικοποιημένη (row normalized hyperlink matrix)[10], εφόσον

$$\mathbf{H}(i,j) = \begin{cases} \dfrac{1}{|P_i|}, & \text{Αν υπάρχει υπερ-σύνδεσμος από τη σελίδα } P_i \text{ στη σελίδα } P_j \\ 0, & \text{Σε κάθε άλλη περίπτωση} \end{cases}$$

Οι θετικές συνιστώσες ενός διανύσματος-γραμμής i του πίνακα $\mathbf{H}$ αντιστοιχούν στους εξερχόμενους υπερ-συνδέσμους της ιστοσελίδας P_i. Αντιστοίχως, οι θετικές συνιστώσες ενός διανύσματος-στήλης j του πίνακα $\mathbf{H}$ αντιστοιχούν στους εισερχόμενους υπερ-συνδέσμους (inlinks) στη σελίδα P_j. Σημειώνεται ότι είναι δυνατή η ύπαρξη μη κανονικοποιημένων γραμμών στον πίνακα $\mathbf{H}$, των οποίων όλα τα στοιχεία είναι μηδενικά. Οι μηδενικές γραμμές αντιστοιχούν σε ιστοσελίδες, στις οποίες δεν υπάρχουν εξερχόμενοι σύνδεσμοι. Συνεπώς στη γενική περίπτωση ο πίνακας $\mathbf{H}$ είναι υποστοχαστικός (substochastic matrix). Στην επόμενη ενότητα μελετάται ο κρίσιμος ρόλος της υποστοχαστικής ιδιότητας (substochastic property) του πίνακα $\mathbf{H}$ στη σύγκλιση του αλγορίθμου.

Από δομική απόψη, ο πίνακας $\mathbf{H}$ είναι ίδιος με τον πίνακα γειτνίασης (adjacency matrix) ενός γράφου με την μόνη διαφορά να εντοπίζεται στις τιμές των μη μηδενικών στοιχείων[11]. Οι βαθμοί δημοτικότητας όλων των ιστοσελίδων υπολογίζονται από την ακόλουθη επαναληπτική σχέση

---

[10] Το άθροισμα των στοιχείων της γραμμής ισούται με 1.

[11] Οι τιμές των στοιχείων του πίνακα γειτνίασης $\boldsymbol{A}$ ενός γράφου $G(V,E)$ καθορίζονται ως εξής

$$\boldsymbol{A}(i,j) = \begin{cases} 1, & \text{Αν υπάρχει ακμή από τον κόμβο } i \text{ στον κόμβο } j \\ 0, & \text{Σε κάθε άλλη περίπτωση} \end{cases}$$

$$\boldsymbol{\pi}_{k+1}^{T} = \boldsymbol{\pi}_{k}^{T} \cdot \mathbf{H} \qquad 3$$

Ο πίνακας **H** έχει την ακόλουθη μορφή για τον γράφο του σχήματος Σχήμα 18

$$\mathbf{H} = \begin{array}{c} \\ \mathrm{P}_1 \\ \mathrm{P}_2 \\ \mathrm{P}_3 \\ \mathrm{P}_4 \\ \mathrm{P}_5 \\ \mathrm{P}_6 \end{array} \begin{array}{c} \begin{array}{cccccc} \mathrm{P}_1 & \mathrm{P}_2 & \mathrm{P}_3 & \mathrm{P}_4 & \mathrm{P}_5 & \mathrm{P}_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Σε κάθε επανάληψη του αλγορίθμου απαιτείται ο υπολογισμός του γινομένου διανύσματος-πίνακα $\boldsymbol{\pi}_{k}^{T} \cdot \mathbf{H}$. Στατιστικές μελέτες για τις ιδιότητες των ιστοσελίδων αναφέρουν ότι κατά μέσο όρο σε κάθε ιστοσελίδα υπάρχουν από επτά (7) ως δέκα (10) εξερχόμενοι σύνδεσμοι. Η συγκεκριμένη στατιστική ιδιότητα των ιστοσελίδων καθιστά τον τετραγωνικό πίνακα **H** αραιό (sparse matrix) με άμεση συνέπεια την εκτέλεση του πολλαπλασιασμού $\boldsymbol{\pi}_{k}^{T} \cdot \mathbf{H}$ σε αλγοριθμικό χρόνο μικρότερο από τον χρόνο που απαιτείται στην αντίστοιχη περίπτωση ενός πυκνού τετραγωνικού πίνακα (dense matrix). Με χρήση κατάλληλων αλγορίθμων ο πολλαπλασιασμός διανύσματος με αραιό πίνακα **H** απαιτεί $O\big(nnz(\mathbf{H})\big)$ υπολογισμούς, όπου $nnz(\mathbf{H})$ είναι ο αριθμός των μη μηδενικών στοιχείων του πίνακα **H**. Στη γενική περίπτωση του πολλαπλασιασμού απαιτούνται $O(n^2)$ υπολογισμοί, όπου n είναι το μέγεθος του τετραγωνικού πίνακα. Επίσης, η ιδότητα του **H** ως αραιού πίνακα επιτρέπει τη χρήση αποδοτικών μεθόδων αποθήκευσης, δεδομένου του γεγονότος ότι η διάστασή του είναι αρκετά υψηλή. Χαρακτηριστικά αναφέρεται ότι το έτος 2004 ο υπολογισμός του αλγορίθμου PageRank αφορούσε προσεγγιστικά $8 \cdot 10^8$ ιστοσελίδες, οπότε η αποθήκευση όλων των στοιχειών του πίνακα **H** θα απαιτούσε $(8 \cdot 10^8)^2$ θέσεις κινητής υποδιαστολής στην κύρια ή την δευτερεύουσα μνήμη.

4.2.2 Σύγκλιση του αλγορίθμου

Ο πίνακας **H** μπορεί να θεωρηθεί ως ο πίνακας μετάβασης μίας μαρκοβιανής αλυσίδας (transition probability matrix). Ευρισκόμενος σε μία συγκεκριμένη ιστοσελίδα P_j με $|P_j|$ το

πλήθος εξερχόμενους συνδέσμους, ο χρήστης επιλέγει με πιθανότητα $1/|P_j|$ να μεταβεί σε μία εκ των $|P_j|$ ιστοσελίδων. Συνεπώς, στην αρχική έκδοση του αλγορίθμου PageRank η πιθανότητα μετάβασης σε μία ιστοσελίδα P_i από μία δεδομένη ιστοσελίδα P_j ακολουθεί την κανονική κατανομή. Τα κριτήρια σύγκλισης για την εύρεση του μοναδικού στάσιμου διανύσματος (stationary vector) μίας μαρκοβιανής διαδικασίας δίνουν απάντηση στο ερώτημα της ύπαρξης σημείου σύγκλισης της επαναληπτικής διαδικασίας του αλγορίθμου PageRank. Σημειώνεται ότι το διάνυσμα $\boldsymbol{\pi}_k$, διάστασης $n \times 1$, φέρει τη στοχαστική ιδιότητα για όλες τις τιμές του k, και το στοιχείο του $\boldsymbol{\pi}_k(i)$ εκφράζει την πιθανότητα ο χρήστης του διαδικτύου να μεταβεί στην ιστοσελίδα P_i μετά από k το πλήθος μεταβάσεις, οι οποίες ορίζουν μία συγκεκριμένη διαδρομή - μονοπάτι στο γράφο του διαδικτύου.

Η ασυμπτωτική συμπεριφορά μίας μαρκοβιανής διαδικασίας, δηλαδή η ύπαρξη του ορίου

$$\boldsymbol{\pi}^{\boldsymbol{T}} = \lim_{k \to \infty} \boldsymbol{\pi}_k^T$$

προϋποθέτει τα ακόλουθα:

- Ο πίνακας $\mathbf{H}$ πρέπει να φέρει τη στοχαστική ιδιότητα (stochastic property) αντί της υποστοχαστικής (substochastic property),
- Ολες οι καταστάσεις της μαρκοβιανής αλυσίδας πρέπει να είναι παροδικές, δηλ., μη βασικές[12].

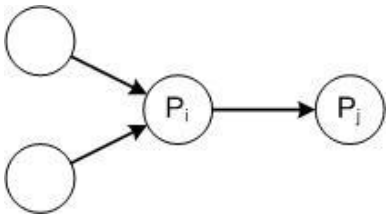
Ακολούθως, περιγράφονται οι δύο τροποποιήσεις στις οποίες υπόκειται ο πίνακας $\mathbf{H}$ για την πλήρωση των προαναφερθέντων προϋποθέσεων.

Πρώτη Τροποποίηση (stochasticity adjustment): *Μετατροπή του πίνακα* $\mathbf{H}$ *από υποστοχαστικό σε στοχαστικό*.

[12] Μια κατάσταση σε μία μαρκοβιανή αλυσίδα ονομάζεται βασική αν και μόνο αν είναι επαναληπτική ή απορρόφησης.

Οπως αναφέρθηκε προηγουμένως, στην περίπτωση του αλγορίθμου PageRank η υποστοχαστικότητα του πίνακα $\mathbf{H}$ (δηλαδή η ύπαρξη γραμμών, των οποίων το άθροισμα των στοιχείων είναι διάφορο της μονάδος) οφείλεται στην παρουσία κόμβων στο γράφο του διαδικτύου χωρίς εξερχόμενους συνδέσμους[13]. Χαρακτηριστικά παραδείγματα κόμβων απορρόφησης (dangling node) αποτελούν ιστοσελίδες, οι οποίες προβάλλουν ως μοναδικό περιεχόμενο αρχεία εικόνων, αρχεία κειμένων ειδικής μορφοποίησης, λογιστικά φύλλα κ.οκ. Η απομάκρυνση των συγκεκριμένων ιστοσελίδων δεν εγγυάται λύση, εφόσον το πρόβλημα της απουσίας συνδέσμων μετατίθεται στις ιστοσελίδες, οι οποίες έχουν εξερχόμενους συνδέσμους στις πρώτες. Το παράδειγμα, που ακολουθεί, είναι χαρακτηριστικό της ανεπάρκειας της λύσης της απομάκρυνσης των κόμβων απορρόφησης από το γράφο του διαδικτύου.

Θεωρούμε τον υπογράφο του παρακάτω σχήματος



Σχήμα 21

Η ιστοσελίδα P_j δεν έχει εξερχόμενους υπερσυνδέσμους προς καμία άλλη ιστοσελίδα του διαδικτύου. Μία ενδεχόμενη απομάκρυνσή της έχει ως επίπτωση την κατάργηση του μοναδικού υπερ-συνδέσμου της ιστοσελίδας P_i.

Η παραδοχή, η οποία προτάθηκε στην βασική εκδοχή του αλγορίθμου PageRank θεωρεί ότι αν ο χρήστης καταλήξει σε μία ιστοσελίδα χωρίς υπερ-συνδέσμους κατά την περιήγηση στο γράφο του διαδικτύου, επιλέγει με τυχαίο τρόπο να κατευθυνθεί σε μία οποιαδήποτε άλλη ιστοσελίδα του διαδικτύου εισάγοντας την διεύθυνση της τελευταίας στη γραμμή διευθύνσεων του φυλλομετρητή. Συνεπώς, κάθε μηδενικό διάνυσμα $\mathbf{0}^T$ διάστασης $1 \times n$, το οποίο αντιστοιχεί σε γραμμή με μηδενικά στοιχεία του πίνακα $\mathbf{H}$,

[13] Στη θεωρία γράφων οι κόμβοι με τη συγκεκριμένη ιδιότητα ονομάζονται κόμβοι απορρόφησης.

αντικαθίσταται από το στοχαστικό διάνυσμα $\frac{1}{n} \cdot \boldsymbol{e}^T$, όπου $\boldsymbol{e}^T$ είναι το μοναδιαίο διάνυσμα[14] διάστασης $1 \times n$.

Η κατηγοριοποίηση των ιστοσελίδων με κριτήριο την ύπαρξη ή όχι εξερχόμενων υπερσυνδέσμων στο περιεχόμενό τους επιτρέπει τον σχηματισμό διανύσματος **a** διάστασης $n \times 1$, του οποίου το πεδίο τιμών των συνιστωσών του είναι δίτιμο, $\{0,1\}$. Συγκεκριμένα, η συνιστώσα i δέχεται την τιμή 1, $\mathbf{a}(i) = 1$, αν στην ιστοσελίδα P_i δεν υπάρχει κανένας υπερ-σύνδεσμος. Σε διαφορετική περίπτωση ισχύει $\mathbf{a}(i) = 0$.

Συνδυάζοντας τα παραπάνω, ο υποστοχαστικός πίνακας **H** μετασχηματίζεται στον στοχαστικό πίνακα **S** σύμφωνα με την ακόλουθη σχέση

$$\mathbf{S} = \mathbf{H} + \mathbf{a}\left(\frac{1}{n} \cdot \boldsymbol{e}^T\right) \qquad 4$$

Ο στοχαστικός πίνακας **S** για το γράφο διαδικτύου του Σχήμα 18 έχει την ακόλουθη μορφή

$$\mathbf{S} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

<u>Δεύτερη Τροποποίηση (primitivity adjustment)</u>: *Εξάλειψη των βασικών καταστάσεων της μαρκοβιανής αλυσίδας με πίνακα μετάβασης* **S**.

Η στοχαστική ιδιότητα του πίνακα **S** δεν εγγυάται σε όλες τις περιπτώσεις την σύγκλιση του επαναληπτικού αλγορίθμου για την εύρεση του μοναδικού στάσιμου διανύσματος. Στην ανάλυση που έχει προηγηθεί, το μοντέλο περιήγησης του χρήστη στο διαδίκτυο υποθέτει αφενός την τυχαία επιλογή ενος υπερσυνδέσμου από το σύνολο των εξερχόμενων υπερσυνδέσμων μίας δεδομένης ιστοσελίδας και αφετέρου την μετάβαση με τυχαίο τρόπο σε οποιαδήποτε ιστοσελίδα του διαδικτύου στην περίπτωση απουσίας

[14] Ολα οι συνιστώσες ενός μοναδιαίου διανύσματος έχουν την τιμή 1.

υπερσυνδέσμων. Το συγκεκριμένο μοντέλο προσαυξάνεται ώστε να προσεγγίζει καλύτερα την συμπεριφορά του χρήστη κατά την περιήγησή του στο διαδίκτυο. Συγκεκριμένα, κατά την επίσκεψη σε μία ιστοσελίδα στην οποία υπάρχουν υπερ-σύνδεσμοι, διακρίνονται δύο πιθανά ενδεχόμενα.

- Επιλογή με τυχαίο τρόπο ενός εκ των εξερχομένων υπερσυνδέσμων της ιστοσελίδας όπως προβλέπει το αρχικό μοντέλο περιήγησης,
- Ο χρήστης δεν επιλέγει έναν εκ των υπερ-συνδέσμων της ιστοσελίδας. Αντ' αυτού, εισάγει μία καινούργια διεύθυνση ιστοσελίδας στο πεδίο διευθύνσεων του φυλλομετρητή. Η διαδικασία περιήγησης στο διαδίκτυο συνεχίζει από την ιστοσελίδα στην οποία θα μετέβει ο χρήστης με την εισαγωγή της καινούργιας διεύθυνσης. Επομένως, εκτός των υπαρκτών υπερ-συνδέσμων σε μία ιστοσελίδα, προστίθενται n εικονικοί υπερ-σύνδεσμοι σε όλες τις ιστοσελίδες του διαδικτύου.

Για την έκφραση με μαθηματικό τρόπο του συγκεκριμένου μοντέλου περιήγησης εισάγεται ο πίνακας $\mathbf{G}$, διάστασης $n \times n$.

$$\mathbf{G} = \alpha \cdot \mathbf{S} + (1-\alpha)\frac{1}{n} \cdot \boldsymbol{e}\boldsymbol{e}^T \qquad 5$$

όπου,

- $\boldsymbol{e}$ είναι το μοναδιαίο διάνυσμα διάστασης $n \times 1$,
- $\mathbf{S} = \mathbf{H} + \mathbf{a}\left(\frac{1}{n} \cdot \boldsymbol{e}^T\right)$, και
- η παράμετρος α είναι ένα βαθμωτό μέγεθος με πεδίο τιμών στο διάστημα $(0,1)$ και εκφράζει την πιθανότητα ο χρήστης να επιλέξει έναν εκ των εξερχομένων υπερσυνδέσμων της ιστοσελίδας την οποία επισκέπτεται για να συνεχίσει την περιήγησή του στο διαδίκτυο. Αντιθέτως, το βαθμωτό μέγεθος $1-\alpha$ εκφράζει την πιθανότητα ο χρήστης να μεταβεί με τυχαίο τρόπο σε κάποια ιστοσελίδα του διαδικτύου εισάγοντας τη διεύθυνση της τελευταίας στο πεδίο διευθύνσεων του φυλλομετρητή. Σημειώνεται ότι στη τελευταία περίπτωση η επιλογή της καινούργιας ιστοσελίδας ακολουθεί την κανονική κατανομή.

  Στις αρχικές εκτελέσεις του αλγορίθμου PageRank χρησιμοποιήθηκε η τιμή $\alpha = .85$. Η ταχύτητα σύγκλιση του αλγορίθμου καθώς και η ευαισθησία των τιμών των

συνιστωσών του στάσιμου διανύσματος της Μαρκοβιανής διαδικασίας ως προς τη μεταβολή της δομής του γράφου του διαδικτύου συναρτάται από την επιλογή της τιμής της παραμέτρου α. Συγκεκριμένα, για τιμές του α, οι οποίες τείνουν στην μονάδα[15], $\alpha \to 1$, ο υπολογιστικός χρόνος (δηλ., ο αριθμός των επαναλήψεων), ο οποίος απαιτείται για την σύγκλιση στο στάσιμο διάνυσμα, αυξάνεται δραματικά. Στον ακόλουθο πίνακα παρουσιάζεται ο αριθμός επαναλήψεων του αλγορίθμου για διάφορες τιμές της παραμέτρου α.

$\boldsymbol{\alpha}$	Αριθμός Επαναλήψεων
0.5	34
0.8	104
0.9	219
0.95	449
0.99	2,292
0.999	23,015

Η ευαισθησία-μεταβλητότητα των τιμών των συνιστωσών του στάσιμου διανύσματος ως προς την δομή του γράφου του διαδικτύου είναι υψηλή καθώς η τιμή της παραμέτρου α τείνει στην μονάδα. Συνεπώς, για $\alpha \to 1$, μικρές αλλαγές στη δομή του γράφου, όπως η προσθήκη ή η αφαίρεση ιστοσελίδων ή ιστοτόπων, επιφέρουν σημαντικές αλλαγές στις τιμές των συνιστωσών του στάσιμου διανύσματος και κατ' επέκταση στο βαθμό δημοτικότητας των τελευταίων. Δεδομένου ότι οι τροποποιήσεις στη δομή του γράφου αποτελούν τον κανόνα και όχι την εξαίρεση

[15] Δηλαδή κατά την περιήγησή του στο διαδίκτυο, ο χρήστης σπάνια μεταβαίνει σε μία ιστοσελίδα εισάγοντας την διεύθυνσή της στο πεδίο διευθ/υνσεων του φυλλομετρητή. Αντ' αυτού επιλέγει έναν εκ των εξερχομένων υπερσυνδέσμων της κάθε σελίδας την οποία επισκέπτεται για την μετάβασή του στην επόμενη ιστοσελίδα.

λόγω της δυναμικής φύσης του διαδικτύου, επιλέγεται κατάλληλη τιμή της παραμέτρου α, η οποία οδηγεί σε ασυμπτωτικό διάνυσμα του οποίου οι τιμές των συνιστωσών παραμένουν σταθερές σε μικρές μεταβολές της δομής του διαδικτύου.

Σημειώνεται ότι στη βιβλιογραφία ο πίνακας $\mathbf{G}$, διάστασης $n \times n$ αναφέρεται ως πίνακας Google (Google matrix). Ο πίνακας $\mathbf{G}$ είναι στοχαστικός εφόσον προκύπτει από τον κυρτό συνδυασμό των στοχαστικών πινάκων $\mathbf{S}$ και $\mathbf{E} = 1/\mathrm{n}\, \boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$. Με αντικατάσταση της σχέσης 3 στη σχεση 4 προκύπτει ότι

$$\mathbf{G} = \alpha \cdot \mathbf{H} + (a \cdot \mathbf{a} + (1-\alpha) \cdot \boldsymbol{e}) \cdot \frac{1}{\mathrm{n}} \cdot \boldsymbol{e}^T$$

Μετά την εφαρμογή των δύο προηγουμένων τροποποιήσεων η σχέση 3 έχει την ακόλουθη μορφή.

$$\boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}_k^T \cdot \mathbf{G} \qquad 5$$

Η ύπαρξη του στάσιμου διανύσματος $\boldsymbol{\pi}^{\mathrm{T}}$ της Μαρκοβιανής αλυσίδας ισοδυναμεί με την ύπαρξη του ορίου $\lim_{k\to\infty} \mathbf{G}^k$ εφόσον

$$\boldsymbol{\pi}^{\mathrm{T}} = \lim_{k\to\infty} \boldsymbol{\pi}_k^T \cdot \mathbf{G} = \boldsymbol{\pi}_0^T \lim_{k\to\infty} \mathbf{G}^k$$

Σε αντίθεση με τον πίνακα $\mathbf{H}$, ο τετραγωνικός πίνακας $\mathbf{G}$ είναι τελείως πυκνός εφόσον κάθε στοιχείο του (i,j) φέρει μη μηδενική τιμή. Ωστόσο, ο πίνακας $\mathbf{G}$ εξαρτάται με γραμμικό τρόπο από τον πίνακα $\mathbf{H}$, όπως προκύπτει από τον ορισμό των πινάκων $\mathbf{G}$ και $\mathbf{S}$. Συγκεκριμένα

$$\begin{aligned} \boldsymbol{\pi}_{k+1}^T &= \boldsymbol{\pi}_k^T \cdot \mathbf{G} \\ &= \alpha \cdot \boldsymbol{\pi}_k^T \mathbf{S} + \frac{1-\alpha}{n} \cdot \boldsymbol{\pi}_k^T \boldsymbol{e}\boldsymbol{e}^T \\ &= \alpha \cdot \boldsymbol{\pi}_k^T \mathbf{H} + (\alpha \cdot \boldsymbol{\pi}_k^T \mathbf{a} + 1 - \alpha) \cdot \frac{\boldsymbol{e}^T}{n} \end{aligned}$$

Σε κάθε επανάληψη του αλγορίθμου υπολογίζεται αφενός το γινόμενο $\boldsymbol{\pi}_k^T \mathbf{H}$, όπου $\mathbf{H}$ ο αραιός πίνακας διάστασης $n \times n$ και αφετέρου το εσωτερικό γινόμενο δύο διανυσμάτων διάστασης $n \times 1$, $\boldsymbol{\pi}_k^T \mathbf{a}$. Η προηγούμενη ανάλυση αποδεικνύει ότι δεν απαιτείται η αποθήκευση των στοιχείων των πυκνών πινάκων $\mathbf{G}$ και $\mathbf{S}$ κατά τον υπολογισμό του $\boldsymbol{\pi}_{k+1}^T$.

Επανερχόμενοι στον απλοϊκό γράφο διαδικτύου των έξι κόμβων του Σχήμα 18, για τιμή της παραμέτρου $\alpha = 0.9$ ο πίνακας Google είναι ο ακόλουθος

$$\mathbf{G} = 0.9 \cdot \mathbf{H} + \left(0.9 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 0.1 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right) \cdot \frac{1}{6} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\mathbf{G} = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}$$

Το ασυμπτωτικό στάσιμο διάνυσμα της μαρκοβιανής αλυσίδας με πίνακα μετάβασης $\boldsymbol{G}$ και αρχικό διάνυσμα τιμών $\boldsymbol{\pi}_0^T = (1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6)$ είναι

$$\boldsymbol{\pi}^T = (0.03721 \quad 0.05396 \quad 0.04151 \quad 0.3751 \quad 0.206 \quad 0.2862)$$

Ισχύει ότι $\|\boldsymbol{\pi}^T\| = 1$. Το μέγεθος $\boldsymbol{\pi}^T(i)$ εκφράζει την ασυμπτωτική πιθανότητα ο χρήστης να μεταβεί στην ιστοσελίδα P_i κατά την περιήγησή του στο διαδίκτυο. Σύμφωνα με τον αλγόριθμο του PageRank η κατάταξη των έξι ιστοσελίδων κατά φθίνουσα σειρά του βαθμού δημοτικότητας είναι $(P_4 \quad P_6 \quad P_5 \quad P_2 \quad P_3 \quad P_1)$.

4.2.3 Αριθμητικές Μέθοδοι Υπολογισμού του Αλγορίθμου

Σύμφωνα με τις παραδοχές του μοντέλου στο οποίο βασίστηκε ο αλγόριθμος PageRank ο προσδιορισμός του βαθμού δημοτικότητας των ιστοσελίδων ανάγεται στη μελέτη της ασυμπτωτικής συμπεριφοράς μίας Μαρκοβιανής αλυσίδας με πίνακα μετάβασης τον πίνακα Google $\mathbf{G}$, δηλαδή στην εύρεση του ορίου $\lim_{k\to\infty} \boldsymbol{\pi}_k^T$. Ο πίνακας $\mathbf{G}$ εγγυάται την ύπαρξη του συγκεκριμένου ορίου. Οπότε, ασυμπτωτικά ισχύει ότι

$$\lim_{k\to\infty} \boldsymbol{\pi}_k^T = \lim_{k\to\infty} \boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}^T$$

Τότε, η σχέση 5 παίρνει την ακόλουθη μορφή

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{G}$$

όπου $\boldsymbol{\pi}^T \boldsymbol{e} = 1$ είναι η συνθήκη κανονικοποίησης, εφόσον το $\boldsymbol{\pi}^T$ είναι στοχαστικό διάνυσμα διαστάσεων $n \times 1$.

Συνεπώς, το πρόβλημα του προσδιορισμού του βαθμού δημοτικότητας των ιστοσελίδων είναι ισοδύναμο με το πρόβλημα εύρεσης του ιδιοδιανύσματος (eigenvector) $\boldsymbol{\pi}^T$ για την ιδιοτιμή (eigenvalue) $\lambda_1(\mathbf{G}) = 1$. Σημειώνεται ότι ένας τετραγωνικός πίνακας διαστάσεων $n \times n$ έχει n το πλήθος ιδιοτιμές, ενώ η επικρατούσα ιδιοτιμή[16], λ_1, ενός στοχαστικού τετραγωνικού πίνακα ισούται με την μονάδα.

Πολλές αποδοτικές αριθμητικές μέθοδοι επίλυσης του προβλήματος εύρεσης των ιδιοτιμών και των ιδιοδιανυσμάτων ενός τετραγωνικού πίνακα έχουν προταθεί στη σχετική βιβλιογραφία. Ενδεικτικά αναφέρονται οι επαναληπτικές μέθοδοι Gauss-Seidel, Jacobi, BICGSTAB, GMRES κ.α. Ωστόσο, λόγω των ειδικών χαρακτηριστικών του προβλήματος καθορισμού του βαθμού δημοτικότητας των ιστοσελίδων, όπως το μέγεθος και η αραιή δομή του πίνακα $\boldsymbol{H}$ επικράτησε *η μέθοδος της ύψωσης σε δύναμη* (*δυναμομέθοδος - power method*). Αν και η συγκεκριμένη μέθοδος δεν είναι η πιο αποδοτική στη γενική περίπτωση, ωστόσο ενδείκνυται για τον υπολογισμό του ιδιοδιανύσματος, το οποίο αντιστοιχεί στην επικρατούσα ιδιοτιμή $\lambda_1(\mathbf{G}) = 1$ του πίνακα **G.** Κατόπιν, αναλύονται οι κυριότεροι λόγοι επιλογής της συγκεκριμένης επαναληπτικής μεθόδου.

- Η υλοποίηση σε επίπεδο γλώσσας προγραμματισμού είναι απλή και γρήγορη.
- Δεν προβλέπει τον υπολογισμό και την αποθήκευση των πυκνών πινάκων **G** και **S**. Η εκτέλεση της επανάληψης $k + 1$ του αλγορίθμου απαιτεί αποθηκευτικό χώρο για τον αραιό πίνακα **H** καθώς και για τις συνιστώσες των διανυσμάτων **a** και $\boldsymbol{\pi}_k^T$. Αποδοτικότερες επαναληπτικές μέθοδοι προβλέπουν τη δημιουργία και την αποθήκευση πολλαπλών πυκνών διανυσμάτων, το οποίο δεν ενδείκνυται στη περίπτωση προβλημάτων μεγάλων διαστάσεων. Χαρακτηριστικά αναφέρεται ότι η μέθοδος GMRES απαιτεί την αποθήκευση 10 διανυσμάτων διάστασης $n \times 1$ σε κάθε

[16] Ιδιοτιμή με την μεγαλύτερη απόλυτη τιμή.

επανάληψη, το οποίο ισούται προσεγγιστικά με τον αποθηκευτικό χώρου που καταλαμβάνει ο αραίος πίνακας $\mathbf{H}$.

- Η πολυπλοκότητα υπολογισμού του διανύσματος $\boldsymbol{\pi}_{k+1}^{T}$ σε κάθε επανάληψη του αλγορίθμου είναι $O(n)$, δηλαδή γραμμική ως προς το μέγεθος του προβλήματος.
- Ο αριθμός των επαναλήψεων, ο οποίος απαιτείται για τη σύγκλιση του αλγορίθμου, είναι σχετικά μικρός. Αποδεικνύεται ότι ο ασυμπτωτικός ρυθμός σύγκλισης της συγκεκριμένης μεθόδου εξαρτάται από το λόγο των δύο μεγαλύτερων ιδιοτιμών $\lambda_1(\mathbf{G})$ και $\lambda_2(\mathbf{G})$ του πίνακα $\mathbf{G}$. Συγκεκριμένα, ο ασυμπτωτικός ρυθμός σύγκλισης είναι ο ρυθμός στον οποίο $|\lambda_2(\mathbf{G})/\lambda_1(\mathbf{G})|^k \to 0$. Εφόσον ο πίνακας $\mathbf{G}$ είναι στοχαστικός, ισχύει ότι $\lambda_1(\mathbf{G}) = 1$, οπότε η δεύτερη μεγαλύτερη ιδιοτιμή $\lambda_2(\mathbf{G})$ καθορίζει την ταχύτητα σύγκλισης. Ειδικώς για τον πίνακα $\mathbf{G}$ αποδεικνύεται ότι $\lambda_2(\mathbf{G}) \approx \alpha$. Για αριθμό επαναλήψεων $k = 50$ και για $a = 0.85$ ισχύει ότι $a^{50} = 0.85^{50} \approx 0.000296$, το οποίο υποδηλώνει ότι στην πεντηκοστή επανάληψη του αλγορίθμου οι βαθμοί δημοτικότητας των ιστοσελίδων θα υπολογιστούν με ακρίβεια χιλιοστού.

Ενας από τους κύριους λόγους επιλογής του επαναληπτικού αλγορίθμου της δύναμης είναι ο σχετικά μικρός αριθμός επαναλήψεων σε σύγκριση με άλλους επαναληπτικούς αλγορίθμους. Ωστόσο, η ταχεία ανάπτυξη του διαδικτύου έχει ως συνέπεια τον αργό ρυθμό σύγλισης της μεθόδου της δύναμης. Η επιτάχυνση της μεθόδου είναι ουσιαστικής σημασίας εφόσον ο υπολογισμός του αλγορίθμου PageRank διαρκεί ημέρες με την εφαρμογή της κλασσικής μεθόδου της δύναμης. Γενικώς, υπάρχουν δύο προσεγγίσεις στο πρόβλημα της μείωση του υπολογιστικού χρόνου μίας επαναληπτικής διαδικασίας. Η πρώτη πρόσεγγιση προβλέπει την μείωση του υπολογιστικού χρόνου, ο οποίος απαιτείται για την εκτέλεση μίας επανάληψης. Η μείωση του αριθμού των απαιτούμενων επαναλήψεων μέχρι την ικανοποίηση των κριτηριών σύγκλισης αποτελεί τη δεύτερη προσέγγιση. Ακολουθεί μία σύντομη αναφορά σε τρείς μεθόδους επιτάχυνσης του αλγορίθμου PageRank.

4.2.3.1 Προσαρμοσμένη μέθοδο της δύναμης

Η συγκεκριμένη μέθοδος προτάθηκε από τους Sep Kamvar, Taher Haveliwala, Gene Golub και Chris Manning [102]. Σύμφωνα με την κλασσική μέθοδο της δύναμης ο επαναληπτικός αλγόριθμος τερματίζει στην επανάληψη k όταν $\|\boldsymbol{\pi}_k^T - \boldsymbol{\pi}_{k-1}^T\| < \tau$, όπου τ είναι ένα βαθμωτό μέγεθος το οποίο εκφράζει ένα αποδεκτό κριτήριο σύγκλισης. Συνεπώς, στη συγκεκριμένη συνθήκη τερματισμού λαμβάνεται υπόψη μόνο το συνολικό άθροισμα των διαφορών των συνιστωσών δύο διανυσμάτων, τα οποία αντιστοιχούν σε δυο διαδοχικές επαναλήψεις του αλγορίθμου PageRank. Σημειώνεται ότι η παράμετρος τ εκφράζει ένα ανώτατο όριο του συνολικού αθροίσματος των διαφορών των συνιστωσών.

Η προσαρμοσμένη μέθοδο της δύναμης (adaptive power method) βασίζεται στην παρατήρηση ότι υπάρχουν διαφοροποίησεις στην ταχύτητα σύγκλισης των τιμών των συνιστωσών του διανύσματος $\boldsymbol{\pi}^T$. Αναλυτικότερα, παρατηρήθηκε ότι ένας σημαντικός αριθμός συνιστωσών του διανύσματος ικανοποιούν το κριτήριο σύγκλισης μετά από ένα μικρό αριθμό επαναλήψεων του αλγορίθμου PageRank, δηλαδή $|\boldsymbol{\pi}_k^T(i) - \boldsymbol{\pi}_{k-1}^T(i)| < \epsilon$, όπου η παράμετρος ϵ είναι είναι το άνω όριο του κριτηρίου σύγκλισης για την τιμή της i-οστής συνιστώσας του διανύσματος $\boldsymbol{\pi}^T$. Οι συνιστώσες του διανύσματος οι οποίες έχουν συγκλίνει δεν επαναϋπολογίζονται σε μελλοντικές επαναλήψεις του αλγορίθμου. Προφανώς με τη συγκεκριμένη μέθοδο μειώνεται προοδευτικά ο χρόνος εκτέλεσης μίας επανάληψης.

Αν και η προτεινόμενη μέθοδος συγκλίνει στις περισσότερες των περιπτώσεων, δεν έχει αποδειχθεί μαθηματικώς η σύγκλιση. Συνεπώς η συγκεκριμένη μέθοδος ενδέχεται να αποκλίνει. Επίσης, στην περίπτωση ύπαρξης σημείου σύγκλισης (το οποίο είναι και το πιο πιθανό ενδεχόμενο), δεν είναι δεδομένη η ταύτιση της κατάταξης των ιστοσελίδων με την κατάταξη η οποία προκύπτει από την εφαρμογή της κλασσικής επαναληπτικής μεθόδου. Η συγκεκριμένη διαφορά αποδίδεται στη χρήση του μικροσκοπικού κριτηριού σύγκλισης σε επίπεδο συνιστώσας, $|\boldsymbol{\pi}_k^T(i) - \boldsymbol{\pi}_{k-1}^T(i)| < \epsilon$, αντί του μακροσκοπικού κριτηρίου σε επίπεδο διανύσματος, $\|\boldsymbol{\pi}_k^T - \boldsymbol{\pi}_{k-1}^T\| < \tau$.

4.2.3.2 Η μέθοδος της παρεμβολής

Με την εφαρμογή της συγκεκριμένης μεθόδου επιτυγχάνεται η μείωση του αριθμού των επαναλήψεων του αλγορίθμου PageRank. Γενικώς, η μέθοδος της παρεμβολής (extrapolation method) βασίζεται στην ανάπτυξη του διανύσματος $\boldsymbol{\pi}_k^T$ σε μία σειρά αθροίσματος της μορφής

$$\boldsymbol{\pi}_k^T = \boldsymbol{u}_1 + \lambda_2^k \cdot \boldsymbol{u}_2 + \cdots + \lambda_m^k \cdot \boldsymbol{u}_m + \cdots$$

όπου οι όροι $\boldsymbol{u}_m$ και λ_i είναι τα ιδιοδιανύσματα και οι ιδιοτιμές του Google πίνακα $\mathbf{G}$. Εφόσον ο πίνακας $\mathbf{G}$ είναι στοχαστικός, ισχύει ότι $\lambda_1 = 1$ ενώ το ιδιοδιανύσμα $\boldsymbol{u}_1$, το οποίο αντιστοιχεί σε αυτή την ιδιοτιμή αποτελεί το ζητούμενο διάνυσμα με τους βαθμούς δημοτικότητας των ιστοσελίδων. Οποτε ισχύει $\boldsymbol{\pi}^T = \boldsymbol{u}_1$. Οπότε, η παραπάνω σχέση παίρνει την ακόλουθη μορφή

$$\boldsymbol{\pi}_k^T - \left(\lambda_2^k \cdot \boldsymbol{u}_2 + \cdots + \lambda_m^k \cdot \boldsymbol{u}_m + \cdots\right) = \boldsymbol{\pi}^T$$

Η προσέγγιση του διανύσματος $\boldsymbol{\pi}^T$ από το διάνυσμα $\boldsymbol{\pi}_k^T$ προϋποθέτει την ελαχιστοποίηση της τιμής του αθροίσματος $\lambda_2^k \cdot \boldsymbol{u}_2 + \cdots + \lambda_m^k \cdot \boldsymbol{u}_m + \cdots$ ή ισοδύναμα την ελαχιστοποίηση των επιμέρους όρων $\lambda_m^k \cdot \boldsymbol{u}_m$. Αν υποτεθεί ότι $1 = |\lambda_1| > |\lambda_2| > \cdots > |\lambda_m|$ και η τιμή της ιδιοτιμής λ_2 είναι μεγάλη, η ελαχιστοποίηση του όρου $\lambda_2^k \cdot \boldsymbol{u}_2$ πραγματοποιείται για μεγάλες τιμές του k, δηλαδή μετά από ένα μεγάλο αριθμό επαναλήψεων του αλγορίθμου. Επομένως, η αφαίρεση του διανύσματος $\lambda_2^k \cdot \boldsymbol{u}_2$ από το διάνυσμα $\boldsymbol{\pi}_k^T$ έχει ως αποτέλεσμα την ταχύτερη σύγκλιση του αλγορίθμου στο διάνυσμα $\boldsymbol{\pi}^T$, όπως προκύπτει και από την αναδιάταξη των όρων της προηγούμενης σχέσης

$$\boldsymbol{\pi}_k^T - \lambda_2^k \cdot \boldsymbol{u}_2 = \boldsymbol{\pi}^T + \lambda_3^k \cdot \boldsymbol{u}_3 + \cdots + \lambda_m^k \cdot \boldsymbol{u}_m + \cdots$$

Ωστόσο η δυσκολία της συγκεκριμένης μεθόδου έγκειται στον υπολογισμό της ποσότητας $\lambda_2^k \cdot \boldsymbol{u}_2$. Αποδεικνύεται ότι ο υπολογισμός της προϋποθέτει την αποθήκευση των διανυσμάτων $\boldsymbol{\pi}_k^T$, $\boldsymbol{\pi}_{k-1}^T$ και $\boldsymbol{\pi}_{k-2}^T$, με αποτέλεσμα η συγκεκριμένη μέθοδος να θεωρείται μνημοβόρα. Εφόσον η μέθοδος της παρεμβολής αυξάνει το υπολογιστικό κόστος, εφαρμόζεται περιοδικά μετά από ένα συγκεκριμένο αριθμό επαναλήψεων του αλγορίθμου.

Η μέθοδος παρεμβολής του Aitken προβλέπει την αφαίρεση του όρου $\lambda_2^k \cdot \boldsymbol{u}_2$. Στην περίπτωση κατά την οποία $|\lambda_2| \approx |\lambda_3|$ η συγκεκριμένη μέθοδος δεν δίνει ικανοποιητικά αποτελέσματα εφόσον απαιτείται και η επιπλέον αφαίρεση του όρου $\lambda_3^k \cdot \boldsymbol{u}_3$. Στην μέθοδο της τετραγωνικής παρεμβολής αφαιρούνται οι όροι $\lambda_2^k \cdot \boldsymbol{u}_2$ και $\lambda_3^k \cdot \boldsymbol{u}_3$ από το διάνυσμα $\boldsymbol{\pi}_k^T$.

4.2.3.3 Η μέθοδος της συνάθροισης

Με τη μέθοδο της συνάθροισης (aggregation method) επιχειρείται η μείωση αφενός του υπολογιστικού φόρτου σε κάθε επανάληψη και αφετέρου του συνολικού αριθμού των επαναλήψεων. Σύμφωνα με την βιβλιογραφία οι επιδόσεις της συγκεκριμένης μεθόδου είναι ικανοποιητικές σε μεγάλους γράφους διαδικτύου και κατά συνέπεια προτιμάται έναντι των υπολοίπων μεθόδων.

Αναλύοντας την δομή του γράφου του διαδικτύου σε μικροσκοπικό επίπεδο, παρατηρείται ότι ο αριθμός των υπερ-συνδέσμων μεταξύ ιστοσελίδων που ανήκουν στον ίδιο ιστότοπο είναι αρκετά μεγαλύτερος του αντίστοιχου αριθμού μεταξύ ιστοσελίδων διαφορετικών ιστοτόπων. Με βάση την προηγούμενη παρατήρηση, κατασκευάζεται νέος γράφος στον οποίο σε κάθε κόμβος αντιστοιχεί ένας ιστότοπο αντί μίας ιστοσελίδας (host graph). Ο νέος γράφος έχει μικρότερη διάσταση από τον αρχικό με συνέπεια ο υπολογισμός του αλγορίθμου PageRank να είναι ταχύτερος. Ο βαθμός δημοτικότητας ενός ιστοτόπου H_i αντιπροσωπεύει τον συγκεντρωτικό, αθροιστικό βαθμό δημοτικότητας όλων των ιστοσελίδων που ανήκουν στον ιστότοπο. Δεδομένων των βαθμών δημοτικότητας των ιστοτόπων, η διαδικασία εύρεσης του βαθμού δημοτικότητας των ιστοσελίδων είναι η ακόλουθη.

1. Εκτέλεση του αλγορίθμου PageRank για τον υπολογισμό των βαθμών δημοτικότητας όλων των ιστοσελίδων P_i, που ανήκουν στον ιστοτόπου H_j. Ως γράφος διαδικτύου ορίζεται πλέον ο υπογράφος του οποίου οι κόμβοι αντιστοιχούν στις ιστοσελίδες του συγκεκριμένου ιστοτόπου ενώ οι ακμές του είναι οι υπερσύνδεσμοι μεταξύ των συγκεκριμένων ιστοσελίδων. Αγνοούνται οι υπερσύνδεσμοι οι οποίοι παραπέμπουν σε ιστοσελίδες εκτός του συγκεκριμένου ιστοτόπου. Σημειώνεται ότι ο συγκεκριμένος υπολογισμός δεν έχει ιδιαίτερο

υπολογιστικό κόστος εφόσον στη γενική περίπτωση το μέγεθος του υπογράφου είναι τάξεις μεγέθους μικρότερο από το μέγεθος του γράφου του διαδικτύου.

2. Επανάληψη του προηγούμενου βήματος για όλους τους ιστοτόπους H_i.
3. Ο καθολικός βαθμός δημοτικότητας μίας ιστοσελίδας P_i ισούται με το γινόμενο του βαθμού δημοτικότητας του ιστοτόπου H_j στην οποία ανήκει η συγκεκριμένη ιστοσελίδα επί τον βαθμό δημοτικότητάς της όπως υπολογίστηκε στο βήμα 1.

Σημειώνεται ότι η μέθοδος της συνάθροισης δίνει μία ικανοποιητική προσέγγιση των πραγματικών βαθμών δημοτικότητας των ιστοσελίδων, όπως αυτοί υπολογίζονται με την μέθοδο της δυναμομεθόδου. Ο προσεγγιστικός χαρακτήρας της μεθόδου της συνάθροισης αποδίδεται στον αποκλεισμό των υπερσυνδέσμων μεταξύ ιστοσελίδων διαφορετικών ιστοτόπων με συνέπεια την απώλεια πληροφορίας.

4.3 Ο αλγόριθμος HITS

Στην παρούσα ενότητα εξετάζεται ο εναλλακτικός αλγόριθμος HITS[17] για τον προσδιορισμό του βαθμού δημοτικότητας των ιστοσελίδων. Προτάθηκε από τον ερευνητή Jon Kleinberg (7) το έτος 1998, κατά τη διάρκεια του οποίου υλοποιήθηκε ο αλγόριθμος PageRank από τους Sergey Brin και Larry Page. Ο αλγόριθμος HITS χρησιμοποιείται για την τον υπολογισμό των βαθμών δημοτικότητας των ιστοσελίδων στην μηχανή αναζήτησης Teoma.

Η κύρια ομοιότητα μεταξύ των δύο αλγορίθμων, HITS και PageRank, αποτελεί η χρήση της δομής του γράφο διαδικτύου. Ωστόσο, ο αλγόριθμος HITS παρουσιάζει ορισμένες σημαντικές διαφοροποίησεις. Συγκεκριμένα, ενώ ο αλγόριθμος PageRank υπολογίζει ένα βαθμό δημοτικότητας για κάθε ιστοσελίδα, ο αλγόριθμος HITS αποδίδει δύο βαθμούς δημοτικότητας. Επιπλέον, τα αποτελέσματα του αλγορίθμου HITS εξαρτώνται του ερωτήματος που υποβάλλει ο χρήστης στην μηχανή αναζήτησης σε αντίθεση με τον αλγόριθμο PageRank.

Η κεντρική ιδέα του αλγορίθμου HITS συνοψίζεται στην παραδοχή ότι βάσει ενός δοθέντος ερωτήματος κάθε ιστοσελίδα θεωρείται ότι έχει διττή ιδιότητα. Αφενός μπορεί

[17] Η ονομασία HITS αποτελεί ακρωνύμιο της φράσης Hypertext Induced Topic Search.

να θεωρηθεί ως σημείο αναφοράς-πύλη (hub page) για την περαιτέρω περιήγηση του χρήστη σε ιστοσελίδες σχετικές με το ερώτημα και αφετέρου μπορεί να είναι σημείο κύριας, αυθεντικής πληροφορίας σχετικής με το ερώτημα (authority page). Σε μία ιστοσελίδα αναφοράς υπάρχουν πολλοί εξερχόμενοι σύνδεσμοι ενώ σε μία αυθεντική ιστοσελίδα καταλήγουν πολλοί εισερχόμενοι σύνδεσμοι. Οι ιστοσελίδες αναφοράς και αυθεντίας χαρακτηρίζονται ως «καλές» αν ισχύει η ακόλουθη κυκλική δήλωση: Μία ιστοσελίδα αναφοράς θεωρείται «καλή» αν στο περιεχόμενό της υπάρχουν υπερσύνδεσμοι προς «καλές» αυθεντικές ιστοσελίδες, ενώ μία αυθεντική ιστοσελίδα χαρακτηρίζεται «καλή» αν καταλήγουν σε αυτή υπερσύνδεσμοι προερχόμενοι από «καλές» ιστοσελίδες αναφοράς.

4.3.1 Μαθηματική Θεμελίωση

Το πρώτο βήμα στην μαθηματική θεμελίωση του αλγορίθμου HITS συνίσταται στην ποσοτικοποίηση των εννοιών της «αναφοράς» και «αυθεντίας» για μία ιστοσελίδα. Σε κάθε ιστοσελίδα i αποδίδεται το ζεύγος βαθμών (x_i, y_i), όπου x_i και y_i είναι οι βαθμοί αυθεντίας (authority score) και αναφοράς (hub score) αντίστοιχα. Επίσης, εστω E είναι το σύνολο των κατευθυνόμενων ακμών στο γράφο του διαδικτύου ενώ με $e_{ij} \in E$ συμβολίζεται η κατευθυνόμενη ακμή από τον κόμβο (ιστοσελίδα) i στον κόμβο (ιστοσελίδα) j. Αν υποτεθεί ότι έχουν δοθεί αρχικές τιμές $x_i^{(0)}$ και $y_i^{(0)}$ στους βαθμούς αυθεντίας και αναφοράς της ιστοσελίδας i αντίστοιχα, ο αλγόριθμος HITS υπολογίζει τους συγκεκριμένους βαθμούς με διαδοχικές επαναλήψεις σύμφωνα με τις ακόλουθες σχέσεις

$$x_i^{(k)} = \sum_{j:e_{ji} \in E} y_j^{(k)} \quad \text{και} \quad y_i^{(k)} = \sum_{j:e_{ij} \in E} x_j^{(k)} \quad \text{για } k = 1,2,3,\cdots$$

Η συγκέντρωση των παραπάνω εξισώσεων για όλες τις ιστοσελίδες που μετέχουν στον υπολογισμό και η γραφή τους σε μορφή πίνακα απαιτεί την εισαγωγή του πίνακα γειτνίασης $\mathbf{L}$ του κατευθυνόμενου γράφου του διαδικτύου. Υπενθυμίζεται ότι το πεδίο τιμών κάθε στοιχείου (i, j) του πίνακα γειτνίασης είναι δίτιμο $\{0,1\}$ και η τιμή του ορίζεται από την ακόλουθη συνάρτηση

$$\mathbf{L}_{ij} = \begin{cases} 1, & \text{αν υπάρχει ακμή από τον κόμβο } i \text{ στον κόμβο } j, \\ 0, & \text{σε κάθε άλλη περίπτωση.} \end{cases}$$

Οι επαναληπτικές σχέσεις υπολογισμού των βαθμών αναφοράς και αυθεντίας για όλες πλεόν τις ιστοσελίδες εκφράζονται στην ακόλουθη συνοπτική γραφή

$$\mathbf{x}^{(k)} = \mathbf{L}^T\mathbf{y}^{(k-1)} \qquad \text{και} \qquad \mathbf{y}^{(k)} = \mathbf{L}\mathbf{x}^{(k)} \qquad \text{για } k = 1,2,3,\cdots$$

όπου $\mathbf{x}^{(k)}$ και $\mathbf{y}^{(k)}$ είναι διανυσματικά μεγέθη διάστασης $n \times 1$ των οποίων το στοιχείο i δίνει το βαθμό αυθεντίας και αναφοράς της ιστοσελίδας i, αντίστοιχα. Ο δείκτης k δηλώνει τον αριθμό επαναλήψεων του αλγορίθμου. Ο αλγόριθμος τερματίζεται όταν ικανοποιηθούν τα κριτήρια σύγκλισης.

Συνοπτικά τα βήματα του αλγορίθμου HITS στην αρχική έκδοσή του είναι τα ακόλουθα

1. **Βήμα Αρχικοποίησης**: $\mathbf{y}^{(0)} = \mathbf{e}$, όπου $\mathbf{e}$ είναι το μοναδιαίο διάνυσμα στήλη διάστασης $n \times 1$. Υπό προϋποθέσεις, οι οποίες εξετάζονται στην ενότητα 4.3.3 το διάνυσμα αρχικοποίησης μπορεί να είναι ένα οποιοδήποτε διάνυσμα με θετικές συνιστώσες.
2. **Βήμα Επανάληψης**: Μέχρι την ικανοποίηση του κριτηρίου σύγκλισης, σε κάθε επανάληψη εκτελούνται τα ακόλουθα τέσσερα βήματα
   i. $\mathbf{x}^{(k)} = \mathbf{L}^T\mathbf{y}^{(k-1)}$
   ii. $\mathbf{y}^{(k)} = \mathbf{L}\mathbf{x}^{(k)}$
   iii. $k = k + 1$
   iv. Κανονικοποίηση των διανυσμάτων $\mathbf{x}^{(k)}$ και $\mathbf{y}^{(k)}$

   Σημειώνεται ότι το βήμα κανονικοποίησης του διανύσματος $\mathbf{x}^{(k)}$ στην k-οστή επανάληψη του βρόχου υπολογισμού προβλέπει ότι

$$\mathbf{x}^{(k)} \leftarrow \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|_1}$$

   όπου

$$\left\|\mathbf{x}^{(k)}\right\|_1 = \sum_{i=1}^{\mathrm{n}} \left|\mathbf{x}^{(k)}(\mathrm{i})\right|$$

Αντίστοιχα κανονικοποιείται το διάνυσμα $\mathbf{y}^{(k)}$.

Από το ζεύγος των διανυσματικών εξισώσεων

$$\mathbf{x}^{(k)} = \mathbf{L}^T\mathbf{y}^{(k-1)}$$

$$\mathbf{y}^{(k)} = \mathbf{L}\mathbf{x}^{(k)}$$

προκύπτει με αντικατάσταση ότι

$$\mathbf{x}^{(k)} = \mathbf{L}^T\mathbf{L}\mathbf{x}^{(k-1)}$$

$$\mathbf{y}^{(k)} = \mathbf{L}\mathbf{L}^T\mathbf{y}^{(k-1)}$$

Κάθε εξίσωση ορίζει ένα πρόβλημα εύρεσης του ιδιοδιανύσματος, που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή ($\lambda_1 = 1$), με εφαρμογή της μεθόδου της δύναμης. Η ομοιότητα με τον αλγόριθμο PageRank είναι προφανής με εξαίρεση τον διαφορετικό πίνακα συντελεστών. Στον αλγόριθμο HITS ο υπολογισμός του κυρίαρχου ιδιοδιανύσματος αφορά έναν εκ των πινάκων $\mathbf{L}^T\mathbf{L}$ ή $\mathbf{L}\mathbf{L}^T$ αντί του πίνακα Google $\mathbf{G}$ στον αλγόριθμο PageRank. Ο πίνακας $\mathbf{L}^T\mathbf{L}$ καθορίζει τους βαθμούς αυθεντίας των ιστοσελίδων και ονομάζεται *πίνακας αυθεντίας (authority matrix)*. Κατά αντιστοιχία, ο πίνακας $\mathbf{L}\mathbf{L}^T$ καθορίζει τους βαθμούς αναφοράς των ιστοσελίδων και στη βιβλιογραφία αναφέρεται ως *πίνακας αναφοράς (hub matrix)*. Οι πίνακες αναφοράς και αυθεντίας είναι συμμετρικοί, και θετικά ημιορισμένοι.

4.3.2 Υλοποίηση του αλγορίθμου

Διακρίνονται δύο κύριες φάσεις στην υλοποίηση του αλγορίθμου HITS. Αρχικώς, κατασκευάζεται ο κατευθυνόμενος γράφος γειτνίασης $G(V, E)$, ο οποίος σχετίζεται με το ερώτημα του χρήστη και στον οποίο κάθε κόμβος $v \in V$ παριστάνει μία ιστοσελίδα ενώ κάθε ακμή $e \in E$ αντιστοιχεί σε έναν εξερχόμενο υπερσύνδεσμ. Πρέπει να σημειωθεί ότι η δομή του γράφου γειτνίασης είναι διαφορετική για κάθε ερώτημα, γεγονός το οποίο οφείλεται στην εξάρτηση του αλγορίθμου από το ερώτημα του χρήστη. Στη δεύτερη φάση υπολογίζονται οι τελικοί βαθμοί αναφοράς και αυθεντίας για κάθε ιστοσελίδα του γράφου γειτνίασης $G(V, E)$. Μετά το τέλος του υπολογισμού, στο χρήστη της μηχανής αναζήτησης παρουσιάζονται δύο αντιστρόφως ταξινομημένες λίστες (από τον

μεγαλύτερο προς τον χαμηλότερο βαθμό). Στην ενότητα 4.3.1 αναφέρθηκε ο τρόπος υπολογισμού των βαθμών αναφοράς και αυθεντίας των ιστοσελίδων δεδομένου του γράφου γειτνίασης. Στην παρούσα ενότητα θα παρουσιαστεί ο τρόπος κατασκευής του γράφου.

Οι ιστοσελίδες, οι οποίες περιέχουν όρους-λέξεις του ερωτήματος, αντιστοιχούν στους αρχικούς κόμβους του γράφου γειτνίασης $G(V, E)$. Υπάρχουν διάφοροι τρόποι καθορισμού των συγκεκριμένων ιστοσελίδων. Ο πιο απλός προϋποθέτει την ανάκτηση τους από το ευρετήριο περιεχομένου, π.χ., με τη χρήση της δομής των αντιστρόφων αρχείων, η οποία μπορεί να έχει την ακόλουθη γενική μορφή για m όρους.

Ορος	Σελίδες που περιέχουν τον όρο
1ος	3, 117, 3961
2ος	8, 456, 45465, 25379,
⋮	
20ος	3, 15, 19, 101, 673, 1199
⋮	
40ος	3, 31, 909, 11114, 253791
⋮	
m-ιοστός	23487, 8956743

Για κάθε όρο παρατίθενται σε μορφή λίστας οι ιστοσελίδες οι οποίες τον περιέχουν. Κάθε όρος αντιστοιχεί σε μία λέξη, ενώ οι ιστοσελίδες ταυτοποιούνται από έναν μοναδικό ακέραιο αριθμό με την χρήση κατάλληλων συναρτήσεων κατακερματισμού. Αν υποτεθεί ότι το ερώτημα προς την μηχανή αναζήτησης περιέχει τον πρώτο και τον εικοστό όρο, τότε θα ανακτηθούν από το ευρετήριο περιεχομένου το σύνολο των ιστοσελίδων που προκύπτουν από τη συγχώνευση των δύο αντίστοιχων λιστών, δηλ., 3, 15, 19, 101, 117, 673, 1199 και 3961. Οι συγκεκριμένες ιστοσελίδες αποτελούν τους αρχικούς κόμβους στον γράφο γειτνίασης G. Κατόπιν, ο γράφος G επεκτείνεται με την

προσθήκη ιστοσελίδων οι οποίες έχουν υπερσυνδέσμους προς το αρχικό σύνολο ιστοσελίδων. Επίσης προστίθενται ιστοσελίδες στις οποίες καταλήγουν υπερσύνδεσμοι οι οποίοι εκκινούν από το αρχικό σύνολο ιστοσελίδων. Οι πληροφορίες για την ύπαρξη υπερσυνδέσμων μεταξύ ιστοσελίδων είναι αποθηκευμένες στο ευρετήριο δομής.

Ο τρόπος επέκτασης του γράφου γειτνίασης επιτρέπει τον σημασιολογικό συσχετισμό ιστοσελίδων με κοινή ή παρεμφερή θεματολογία. Αναλυτικότερα, αν το ερώτημα περιέχει τον όρο «μαγειρική», τότε με τη διαδικασία της επέκτασης θα συμπεριληφθούν στο γράφο γειτνίασης ιστοσελίδες οι οποίες περιέχουν τον όρο «κουζίνα» (με την προϋπόθεση της ύπαρξης συνδέσμων μεταξύ των ιστοσελίδων που περιέχουν τους συνώνυμους όρους). Με το συγκεκριμένο τρόπο επέκτασης του γράφου λαμβάνονται υπόψη οι συνώνυμοι όροι. Ωστόσο, ο γράφος G ενδέχεται να είναι μεγάλος μετά το τέλος της διαδικασίας της επέκτασης, εφόσον σε μία ιστοσελίδα, η οποία περιέχει έναν όρο του ερωτήματος, μπορεί να αντιστοιχούν πολλοί εισερχόμενοι και εξερχόμενοι υπερσύνδεσμοι. Πρακτικώς, για το σχηματισμό του γράφου γειτνίασης τίθενται αριθμητικοί περιορισμοί στον αριθμό των εισερχόμενων και των εξερχομένων υπερσυνδέσμων κάθε ιστοσελίδας. Παραδείγματος χάρη αν ο αριθμός των εξερχομένων υπερσυνδέσμων μίας ιστοσελίδας η οποία περιέχει έναν όρο του ερωτήματος υπερβαίνει ένα προκαθορισμένο αριθμό (έστω 100) τότε στο γράφο γειτνίασης προστίθενται ακριβώς εκατό κόμβοι.

Εφόσον έχει κατασκευαστεί ο γράφος γειτνίασης (neighborhood graph), σχηματίζεται ο πίνακας γειτνίασης $\mathbf{L}$. Η διάσταση του τετραγωνικού πίνακα $\mathbf{L}$ είναι τάξεις μεγέθους μικρότερη της διάστασης του πίνακα Google στον αλγόριθμο PageRank. Συνεπώς, το κόστος υπολογισμού των βαθμών αυθεντίας και αναφοράς των ιστοσελίδων του γράφου G με χρήση των πινάκων αυθεντίας ($\mathbf{L}^T\mathbf{L}$) και αναφοράς αντίστοιχα ($\mathbf{L}\mathbf{L}^T$) ειναι μικρότερο συγκρινόμενο με το αντίστοιχο κόστος όταν ο υπολογισμός αφορά το σύνολο των ιστοσελίδων του Παγκόσμιου Ιστού (όπως συμβαίνει στον αλγόριθμο PageRank).

Επιπλεόν, το μικρό κόστος υπολογισμού του αλγορίθμου HITS αποδίδεται στο γεγονός ότι η επίλυση ενός εκ των δύο προβλημάτων εύρεσης του επικρατέστερου ιδοδιανύσματος συνεπάγεται την άμεση λύση του δεύτερου προβλήματος. Συγκεκριμένα, αν έχει υπολογιστεί το διάνυσμα $\mathbf{x}$ με τους βαθμούς αυθεντίας των

ιστοσελίδων, τότε το διάνυσμα $\mathbf{y}$ με τους βαθμούς αναφοράς προκύπτει από την εξίσωση $\mathbf{y} = \mathbf{Lx}$.

4.3.3 Σύγκλιση του αλγορίθμου

Εφόσον ο πίνακας $\mathbf{L}^T\mathbf{L}$ ($\mathbf{L}\mathbf{L}^T$) διάστασης $n \times n$ ειναι συμμετρικός, θετικά ημιορισμένος και τα στοιχεία του έχουν θετικές τιμές, οι ιδιοτιμές λ_i είναι πραγματικοί, μη αρνητικοί αριθμοί και επομένως ορίζεται σχέση διάταξης στο σύνολο των διακριτών ιδιοτιμών $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ τέτοια ώστε $\lambda_1 > \lambda_2 > \cdots > \lambda_k \geq 0$. Συνεπώς, η επαναληπτική μέθοδο της δύναμης για τον υπολογισμό του πρωτεύοντος ιδιοδιανύσματος του πίνακα $\mathbf{L}^T\mathbf{L}$ συγκλίνει για κάθε αρχικό διάνυσμα τιμών.

Η ταχύτητα σύγκλισης της επαναληπτικής διαδικασίας καθορίζεται από τον ρυθμό σύγκλισης του λόγου $\left(\frac{\lambda_2(\mathbf{L}^T\mathbf{L})}{\lambda_1(\mathbf{L}^T\mathbf{L})}\right)^k$ στη τιμή 0, δηλ., $\left(\frac{\lambda_2(\mathbf{L}^T\mathbf{L})}{\lambda_1(\mathbf{L}^T\mathbf{L})}\right)^k \to 0$. Σε αντίθεση με τον αλγόριθμο PageRank στον οποίο η ταχύτητα σύγκλισης της μεθόδου της δύναμης προσεγγιζόταν ικανοποιητικά από την τιμή της παραμέτρου α [18], η εκτίμηση της ταχύτητας σύγκλισης του αλγορίθμου HITS δεν είναι εύκολη. Πειραματικές εκτελέσεις του αλγορίθμου αναφέρουν το τερματισμό της μεθόδου της δύναμης μετά από 10-15 επαναλήψεις. Ωστόσο, ο γρήγορος ρυθμός σύγκλισης δεν συνδυάζεται με την μοναδικότητα της λύσης για διάφορες αρχικές τιμές του διανύσματος $\mathbf{x}$. Αυτό αποδίδεται στο γεγονός της ύπαρξης επαναλαμβανομένων ριζών στο χαρακτηριστικό πολυώνυμο του πίνακα $\mathbf{L}$.

Η μοναδικότητα της λύσης για διάφορες αρχικές τιμές του διανύσματος $\mathbf{x}$ εξασφαλίζεται με την τροποποίηση του πίνακα $\mathbf{L}^T\mathbf{L}$. Συγκεκριμένα, ο κατευθυνόμενος γράφος με n το πλήθος κόμβους ο οποίος αντιστοιχεί στον τροποποιημένο πίνακα $\mathbf{L}^T\mathbf{L}$ πρέπει να είναι πλήρως συνεκτικός[19]. Δηλαδή για κάθε ζεύγος κόμβων (i, j) υπάρχει μονοπάτι από τον κόμβο i στον κόμβο j. Η προτεινόμενη τροποποίηση είναι ανάλογη αυτής στον αλγόριθμο PageRank και περιγράφεται από τη σχέση

---

[18] Βλ. ενότητα 4.2.2. Υψηλή τιμή της παραμέτρου α είχε ως συνέπεια τον αργό ρυθμό σύγκλισης.

[19] Εφαρμογή του θεωρήματος Perron-Frobenious.

$$\mathbf{M} = \xi \cdot \mathbf{L}^T\mathbf{L} + \frac{1-\xi}{n} \cdot \boldsymbol{e}\boldsymbol{e}^T$$

όπου ξ είναι πραγματικός αριθμός στο διάστημα $(0,1)$, ενώ $\boldsymbol{e}$ είναι το μοναδιαίο διάνυσμα διάστασης $n \times 1$. Η τροποποίηση για τον πίνακα αναφοράς δίνεται από τη σχέση

$$\xi \cdot \mathbf{L}\mathbf{L}^T + \frac{1-\xi}{n} \cdot \boldsymbol{e}\boldsymbol{e}^T$$

4.3.4 Μειονεκτήματα και πλεονεκτήματα του αλγορίθμου

Κύριο χαρακτηριστικό του αλγορίθμου HITS είναι η απόδοση δύο διαφορετικών βαθμών (βαθμός αναφοράς και βαθμός αυθεντίας) σε κάθε ιστοσελίδα. Δύο ταξινομημένες λίστες συνιστούν την έξοδο του αλγορίθμου. Τα στοιχεία της πρώτης λίστας αφορούν τους βαθμούς αυθεντίας των ιστοσελίδων οι οποίες είναι σχετικές με το ερώτημα του χρήστη. Ο χρήστης του διαδικτύου ενδιαφέρεται για ιστοσελίδες με υψηλό βαθμό αυθεντίας στην περίπτωση κατά την οποία η αναζήτησή του είναι στοχευμένη. Τα στοιχεία της δεύτερης λίστας αφορούν τους βαθμούς αναφοράς των ιστοσελίδων. Ιστοσελίδα με υψηλό βαθμό αναφοράς περιέχει υπερσυνδέσμους σε «καλές» με ποιοτικό περιεχόμενο ιστοσελίδες, οι οποίες σχετίζονται με το ερώτημα του χρήστη. Συνεπώς, η συγκεκριμένη ιστοσελίδα μπορεί να θεωρηθεί ως πύλη και σημείο εκκίνησης για μία ευρεία αναζήτηση στο διαδίκτυο. Η ευελιξία του αλγορίθμου του HITS έγκειται στο γεγονός της επιλογής ενός εκ των δύο βαθμολογιών αναλόγως των πληροφοριακών αναγκών του χρήστη.

Οπως αναφέρθηκε στην προηγούμενη παράγραφο, κύριο πλεονέκτημα του αλγορίθμου θεωρείται το μικρό μέγεθος του γράφου γειτνίασης G, ο οποίος στην γενική περίπτωση είναι τάξεις μεγέθους μικρότερος από το γράφο του διαδικτύου. Η διάσταση των πινάκων $\mathbf{L}^T\mathbf{L}$ και $\mathbf{L}\mathbf{L}^T$ είναι μικρή και επομένως ο υπολογιστικός φόρτος της αριθμητικής μεθόδου της δύναμης για την εύρεση των αντίστοιχων ιδιοδιανυσμάτων είναι μικρότερος του αντίστοιχου φόρτου στον αλγόριθμο PageRank. Η επικράτηση της δυναμομεθόδου στον αλγόριθμο PageRank οφείλεται στο γεγονός ότι γίνεται συντηρητικότερη διαχείριση του πόρου της κύριας μνήμης με την μη αποθήκευση πλεονάζουσας πληροφορίας μεταξύ των διαδοχικών επαναλήψεων. Στο άρθρο (7) αναφέρεται η χρήση της μεθόδου της δύναμης για τον υπολογισμό των διανυσμάτων $\mathbf{x}$ και $\mathbf{y}$, τα οποία είναι τα

επικρατέστερα ιδιοδιανύσματα των πινάκων $\mathbf{L}^T\mathbf{L}$ και $\mathbf{L}\mathbf{L}^T$, αντίστοιχα. Ωστόσο, ταχύτερες αριθμητικές μέθοδοι από τη δύναμη της μεθόδου έχουν προταθεί για τον υπολογισμό του επικρατέστερου ιδιοδιανύσματο. Κύριο γνώρισμά τους είναι η εντατική χρήση του πόρου της κύριας μνήμης, η οποία είναι επιτρεπτή εφόσον το μέγεθος του εξεταζόμενου προβλήματος είναι σχετικά μικρό. Αν και δεν είναι γνωστή η μέθοδος υπολογισμού που ακολουθείται από τις εμπορικές μηχανές αναζήτησης (π.χ., Teoma) για την βαθμολόγηση των ιστοσελίδων με τον αλγόριθμο HITS, είναι αποδεκτή η χρήση ταχύτερων επαναληπτικών μεθόδων όπως η μέθοδος του Lanczos αντί της αργής μεθόδου της δύναμης.

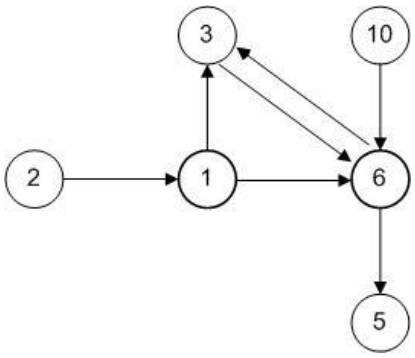
Το κυριότερο μειονέκτημα του αλγορίθμου HITS είναι η εξάρτηση των βαθμών αναφοράς και αυθεντίας από το ερώτημα του χρήστη (query-dependent). Η επεξεργασία κάθε ερωτήματος περιλαμβάνει τη δημιουργία του γράφου και του πίνακα γειτνίασης καθώς και την επίλυση ενός προβλήματος εύρεσης ιδιοδιανύσματος. Στην ενότητα 4.3.6 παρουσιάζεται μία τροποποίηση του αλγορίθμου HITS, στην οποία τα αποτελέσματα είναι ανεξάρτητα του ερωτήματος (query-independent). Συνοπτικά, προτείνεται η κατάργηση του βήματος δημιουργίας διακριτού γράφου γειτνίασης για κάθε ερώτημα. Οι βαθμοί αναφοράς και αυθεντίας υπολογίζονται λαμβάνοντας υπόψιν τον γράφο του διαδικτύου και όχι μέρος του.

4.3.5 Παράδειγμα

Στην παρούσα ενότητα παρουσιάζεται ένα απλό παράδειγμα εφαρμογής του αλγορίθμου HITS. Ο χρήστης της μηχανής αναζήτησης υποβάλλει το ερώτημά του σε φυσική γλώσσα. Με την προσπέλαση του ευρετηρίου περιεχομένου σχηματίζεται το σύνολο των ιστοσελίδων στο περιεχόμενο των οποίων εμφανίζεται οι όροι του ερωτήματος. Η απαίτηση της εμφάνισης όλων των όρων του ερωτήματος συνεπάγεται μικρό αριθμό ιστοσελίδων και κατ' επέκταση μικρό μέγεθος του γράφου γειτνίασης.

Υποτίθεται ότι στις ιστοσελίδες με τον μοναδικό αριθμό ταυτοποίησης 1 και 6 εμφανίζονται οι όροι του ερωτήματος. Στο γράφο του Σχήμα 22 οι κόμβοι που αντιστοιχούν στις συγκεκριμένες ιστοσελίδες σημειώνονται με έντονη γραμμή περιγράμματος. Με την προσπέλαση του ευρετηρίου δομής καθορίζονται οι ιστοσελίδες

με συνδέσμους από και πρός τις ιστοσελίδες 1 και 6. Μετά την επέκταση ο γράφος γειτνίασης έχει την ακόλουθη μορφή του ακόλουθου σχήματος.



Σχήμα 22 Ο γράφος γειτνίασης

Ο πίνακας γειτνίασης $\mathbf{L}$ είναι ο ακόλουθος

$$\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Οι πίνακες αυθεντίας και αναφοράς είναι οι ακόλουθοι

$$\mathbf{L}^T\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad \text{και} \quad \mathbf{L}\mathbf{L}^T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Τα ιδιοδιανύσματα $\mathbf{x}^T$ και $\mathbf{y}^T$ των οποίων οι τιμές είναι οι βαθμοί αυθεντίας και αναφοράς αντίστοιχα είναι τα ακόλουθα

$$\mathbf{x}^T = (0 \quad 0 \quad 0.3660 \quad 0.1340 \quad 0.5 \quad 0) \qquad \text{και}$$

$$\mathbf{y}^T = (0.3660 \quad 0 \quad 0.2113 \quad 0 \quad 0.2113 \quad 0.2113)$$

Δεδομένης της ερώτησης προκύπτει ότι η ιστοσελίδα με αριθμό ταυτοποίησης 6 έχει τον μεγαλύτερο βαθμό αυθεντίας ενώ στην ιστοσελίδα 1 δίνεται ο υψηλότερος βαθμός αναφοράς.

Από την ανάλυση των αποτελεσμάτων προκύπτει ότι είναι δυνατή η ύπαρξη δύο ή περισσοτέρων ιστοσελίδων με μηδενικό ή τον ίδιο βαθμό αναφοράς ή αυθεντίας. Οσον αφορά την περίπτωση μη μηδενικών ισοβαθμιών, η πιθανότητα παρουσίας τους σε γράφους μεγαλύτερων διαστάσεων από το παράδειγμα της παρούσας ενότητας είναι μικρή.

Το πρόβλημα της ύπαρξης πολλών ιστοσελίδων με μηδενικό βαθμό αυθεντίας ή αναφοράς επιλύεται με την εφαρμογή της τροποποίησης του πίνακα αυθεντίας όπως αυτή παρουσιάστηκε στην ενότητα 4.3.3. Συγκεκριμένα, για $\xi = 0.95$ και $n = 6$ ο πίνακας αυθεντίας μετασχηματίζεται στον πίνακα

$$\mathbf{M} = \xi \cdot \mathbf{L}^T\mathbf{L} + \frac{1-\xi}{n} \cdot \boldsymbol{e}\boldsymbol{e}^T$$

Με την εφαρμογή της συγκεκριμένης τροποποίησης στον πίνακα αυθεντίας ο αντίστοιχος κατευθυνόμενος γράφος είναι πλήρως συνεκτικός και εξασφαλίζεται η μοναδικότητα του επικρατέστερου ιδιοδιανύσματος $\mathbf{x}^T$ για διάφορες αρχικές τιμές των συνιστωσών του διανύσματος $\mathbf{x}_0^T$. Οι τιμές των διανυσμάτων $\mathbf{x}^T$ και $\mathbf{y}^T$ είναι

$$\mathbf{x}^T = (0.0032 \quad 0.0023 \quad 0.3634 \quad 0.1351 \quad 0.4936 \quad 0.0025) \text{ και}$$

$$\mathbf{y}^T = (0.3628 \quad 0.0032 \quad 0.2116 \quad 0.0023 \quad 0.2106 \quad 0.2106)$$

Στον πίνακα των αποτελεσμάτων οι ιστοσελίδες ταξινομούνται κατά φθίνοντα βαθμό αυθεντίας και αναφοράς.

Βαθμός Αυθεντίας	Βαθμός Αναφοράς
6	1
3	3
5	10

Βαθμός Αυθεντίας	Βαθμός Αναφοράς
1	6
10	2
2	5

Πίνακας 4 Ταξινόμηση των ιστοσελίδων κατά φθίνοντα βαθμό αυθεντίας και αναφοράς

4.3.6 Ανεξαρτησία από το ερώτημα

Ο αλγόριθμος του HITS είναι ανεξάρτητος του ερωτήματος αν υπολογιστεί ο καθολικός βαθμός αναφοράς και αυθεντίας για όλες ανεξαιρέτως τις ιστοσελίδες οι οποίες βρίσκονται στο ευρετήριο δομής της μηχανής αναζήτησης. Στην περίπτωση αυτή αποδεικνύεται ότι η επίτευξη σύγκλισης στα μοναδικά διάνυσμα $\mathbf{x}^T$ και $\mathbf{y}^T$ με τους καθολικούς βαθμούς αυθεντίας και αναφοράς, αντίστοιχα. Σημειώνεται ότι υιοθετείται εκ των προτέρων η τροποποίηση της ενότητας 4.3.2 για τους πίνακες αυθεντίας και αναφοράς εφόσον ο γράφος του διαδικτύου δεν θα είναι πλήρως συνεκτικός με μεγάλη πιθανότητα.

Τα βήματα του αλγορίθμου είναι τα ακόλουθα

1. **Βήμα Αρχικοποίησης**: $\mathbf{x}^{(0)} = \mathbf{e}/n$, όπου n και $\mathbf{e}$ είναι το πλήθος των κόμβων που αποτελούν το γράφο του διαδικτύου και το μοναδιαίο διάνυσμα στήλη διάστασης $n \times 1$, αντίστοιχα. Ως αρχική τιμή μπορει να χρησιμοποιηθεί οποιοδήποτε θετικό κανονικοποιημένο διάνυσμα δίαστασης $n \times 1$.
2. **Βήμα Επανάληψης**: Μέχρι την ικανοποίηση του κριτηρίου σύγκλισης, σε κάθε επανάληψη εκτελούνται τα ακόλουθα βήματα. Τα διανύσματα $\mathbf{x}^{(k)}$ και $\mathbf{y}^{(k)}$ κανονικοποιούνται, και συνεπώς ισχύει $\mathbf{e}^T\mathbf{x}^{(k)} = 1$ και $\mathbf{e}^T\mathbf{y}^{(k)} = 1$.
   i. $\mathbf{x}^{(k)} = \xi\mathbf{L}^T\mathbf{L}\mathbf{x}^{(k-1)} + (1-\xi)/n\,\mathbf{e}$
   ii. $\mathbf{x}^{(k)} = \mathbf{x}^{(k)}/\left\|\mathbf{x}^{(k)}\right\|_1$
   iii. $\mathbf{y}^{(k)} = \xi\mathbf{L}\mathbf{L}^T\mathbf{y}^{(k-1)} + (1-\xi)/n\,\mathbf{e}$

iv. $\mathbf{y}^{(k)} = \mathbf{y}^{(k)} / \left\| \mathbf{y}^{(k)} \right\|_1$

v. $k = k + 1$

Κατόπιν συγκρίνονται ο αλγόριθμος HITS ο οποίος είναι ανεξάρτητος του ερωτήματος με τον αλγόριθμο PageRank. Ο κύριος υπολογιστικός φόρτος σε κάθε βήμα επανάληψης του αλγορίθμου HITS εντοπίζεται στον πολλαπλασιασμό ενός πίνακα με ένα διάνυσμα διαστάσεων $n \times n$ και $n \times 1$ αντίστοιχα. Υπενθυμίζεται ότι στον αλγόριθμο HITS υπολογίζεται το γινόμενο $\mathbf{L}^T \mathbf{L}\mathbf{x}^{(k-1)}$ ενώ στον αλγόριθμο PageRank το γινόμενο $\mathbf{H}^T \mathbf{x}^{(k-1)}$. Στον παρακάτω πίνακα δίνονται με προσέγγιση οι αριθμοί των πολλαπλασιασμών και προσθέσεων, οι οποίοι πραγματοποιούνται σε κάθε βήμα επανάληψης των αλγορίθμων HITS και PageRank. Ο συμβολισμός $nnz(\mathbf{A})$ παριστά τον αριθμό των μη μηδενικών στοιχείων του πίνακα $\mathbf{A}$, ενώ n είναι το μέγεθος του προβλήματος.

Αλγόριθμος	Αριθμός Πολλαπλασιασμών	Αριθμός Προσθέσεων
HITS	0	$2\, nnz(\mathbf{L})$
HITS (Τροποποίηση Ενότητας 4.3.3)	0	$4nnz(\mathbf{L}) + 2n$
PageRank	$nnz(\mathbf{H})$	$nnz(\mathbf{H}) + n$

Πίνακας 5 Υπολογιστικός φόρτος ενός βήματος επανάληψης για τους κυριότερους αλγορίθμους βαθμολόγησης ιστοσελίδων οι οποίοι είναι ανεξάρτητοι του ερωτήματος

Στην περίπτωση εξάρτησης του αλγορίθμου HITS από το ερώτημα του χρήστη χρήστη ισχύει ότι $nnz(\mathbf{L}) \ll nnz(\mathbf{H})$, όπου $\mathbf{H}$ είναι ο πίνακας διάστασης $n \times n$ στον αλγόριθμο PageRank. Για την εκδοχή του αλγορίθμου HITS, στην οποία δεν λαμβάνεται υπόψη το ερώτημα του χρήστη ισχύει ότι $nnz(\mathbf{L}) = nnz(\mathbf{H})$. Στη τελευταία περίπτωση ο

υπολογιστικός φόρτος[20] σε κάθε βήμα επανάληψης είναι περίπου ίσος (διπλάσιος αν χρησιμοποιηθεί η τροποποίηση της παραγράφου 4.3.3) με τον αντίστοιχο φόρτο του αλγορίθμου PageRank. Εφόσον θεωρείται βέβαιη η χρήση του τροποποιημένου αλγορίθμου HITS, συνέπεια της δομής του γράφου του διαδικτύου, η επίδοση του αλγορίθμου HITS υπολείπεται αυτής του αλγορίθμου PageRank στην εκέλεση ενός βήματος επανάληψης.

Ωστόσο, η ταχύτερη εκτέλεση του βήματος επανάληψης δεν εγγυάται την καθολική υπεροχή του αλγορίθμου PageRank έναντι του αλγορίθμου HITS. Στην ενότητα 4.2.3 αναφέρθηκε ότι ο αριθμός των απαιτούμενων επαναλήψεων για τη σύγκλιση του αλγορίθμου PageRank καθορίζεται από την τιμή της παραμέτρου α. Στον τροποποιημένο αλγόριθμο HITS προσδιορίζονται ένα άνω και κάτω φράγμα του ρυθμού σύγκλισης του αλγορίθμου, ο οποίος ισούται με το λόγο γ_2/γ_1, όπου $\gamma_1 \geq \gamma_2$ είναι οι δυο μεγαλύτερες ιδιοτιμές του πίνακα $\mathbf{M}\xi \cdot \mathbf{L}^T\mathbf{L} + \frac{1-\xi}{n} \cdot \boldsymbol{e}\boldsymbol{e}^T$. Συγκεκριμένα

$$\frac{\xi\lambda_2}{\xi\lambda_1 + 1 - \xi} \leq \frac{\gamma_2}{\gamma_1} \leq \frac{\lambda_2}{\lambda_1} + \frac{(1-\xi)}{\xi\lambda_1}$$

όπου $\lambda_1 \geq \lambda_2$ είναι οι δύο μεγαλύτερες ιδιοτιμές του πίνακα αυθεντίας $\mathbf{L}^T\mathbf{L}$. Για τιμές τις παραμέτρου ξ, οι οποίες τείνουν στην μονάδα, ο ρυθμός σύγκλισης γ_2/γ_1 του τροποποιημένου αλγορίθμου HITS ισούται με λ_2/λ_1, ο οποίος είναι ο ρυθμός σύγκλισης του αρχικού αλγορίθμου HITS. Στα επιστημονικά άρθρα (8), (9) αναφέρεται ότι στη γενική περίπτωση ο ρυθμός σύγλισης του αλγορίθμου HITS είναι $\gamma_2/\gamma_1 < 0.5$. Στη βιβλιογραφία αναφέρεται η επιλογή της τιμής $\alpha = 0.85$ για την εκτέλεση του αλγορίθμου PageRank. Συνεπώς, ο τροποποιημένος αλγόριθμος HITS απαιτεί μικρότερο αριθμό επαναλήψεων μέχρι την ικανοποίηση των κριτηρίων σύγκλισης από τον αλγόριθμο PageRank.

Συνοψίζοντας, συγκρινόμενος με τον αλγόριθμο PageRank, η εκδοχή του αλγορίθμου HITS, στον οποίο η βαθμολόγηση των ιστοσελίδων δεν εξαρταται από το ερώτημα του χρήστη, απαιτεί διπλάσιο υπολογιστικό φόρτο για την εκτέλεση ενός βήματος

[20] Ο υπολογιστικός φόρτος ορίζεται ασυμπτωτικά ως το άθροισμα του αριθμού των πράξεων του πολλαπλασιασμού και της άθροισης.

επανάληψης και υποτετραπλάσιο αριθμό επαναλήψεων για την επίτευξη της σύγκλισης. Συνεπώς, στη γενική περίπτωση ο αλγόριθμος HITS είναι ταχύτερος.

5 Μέθοδοι βελτιστοποίησης του βαθμού δημοτικότητας

Από την ανάλυση του κεφαλαίου 4 η θέση ιστοσελίδας στην κατάταξη των φυσικών ή οργανικών λιστών των μηχανών αναζήτησης συναρτάται του αριθμού των εισερχομένων ή/και των εξερχομένων συνδέσμων. Η μεγιστοποίηση του αριθμού των υπερσυνδέσμων προερχομένων από ιστοσελίδες με υψηλό βαθμό δημοτικότητας σε συνδυασμό με την ελαχιστοποίηση των μη ευνοϊκών αναφορών αποτελεί μέρος των ηλεκτρονικών δημόσιων σχέσεων.

Το κτίσιμο συνδέσμων (link building) αποτελεί βασική δραστηριότητα για τη βελτιστοποίηση αποτελεσμάτων σε μηχανές αναζήτησης. Η κύρια αρχή που διέπει τη διαδικασία του κτισίματος υπερσυνδέσμων συνοψίζεται στην ακόλουθη φράση: «Δημιουργία ιστοτόπου με εξαιρετικού περιεχομένου, σύνδεσή του με εξαιρετικό περιεχόμενο και ακολούθως με ισχυρή πιθανότητα θα υπάρξει υπερσύνδεσμος από το εξαιρετικό περιεχόμενο προς τον ιστότοπο». Ωστόσο η συμπερίληψη ορισμένων υπερσυνδέσμων στις ιστοσελίδες ενός ιστοτόπου δεν αρκεί.

Οι εξερχόμενοι και οι εισερχόμενοι σύνδεσμοι ενός ιστοτόπου, οι εσωτερικοί σύνδεσμοι μεταξύ ιστοσελίδων ενός ιστοτόπου καθώς και οι υπερσύνδεσμοι οι οποίοι οδηγούν σε ιστοσελίδες χωρίς εξερχόμενους συνδέσμους αποτελούν τους τέσσερις τύπους υπερσυνδέσμων των οποίων η παρουσία καθορίζει το βαθμό δημοτικότητας της ιστοσελίδας. Ο συντελεστής βαρύτητας των τεσσάρων τύπων υπερσυνδέσμων στη διαμόρφωση του βαθμού δημοτικότητας είναι εν γένει διαφορετικός. Ως χαρακτηριστικό παράδειγμα αναφέρονται οι υπερσύνδεσμοι σε ιστοσελίδες στις οποίες δεν υπάρχουν εξερχόμενοι σύνδεσμοι. Σε αυτή την περίπτωση η παρουσία των υπερσυνδέσμων ενδέχεται να αγνοηθεί από τον αλγόριθμο βαθμολόγησης των μηχανών αναζήτησης με συνέπεια τη χαμηλή βαθμολόγηση στις λίστες αποτελεσμάτων ή ακόμα και την απομάκρυνση από αυτές.

5.1 Βασικές διαπιστώσεις

Το αποδοτικό κτίσιμο συνδέσμων είναι μία χρονοβόρα διαδικασία, η οποία στις περισσότερες των περιπτώσεων απαιτεί χρόνο πολλαπλάσιο της δημιουργίας του περιεχομένου και του παρουσιαστικού μίας ιστοσελίδας. Επιπλέον η συλλογή

υπερσυνδέσμων από ποιοτικούς ιστοτόπους και ιστοσελίδες αποτελεί μία διαδικασία συνεχής καθ' ολη τη διάρκεια ζωής της ιστοσελίδας.

Στις αρχικές εκδόσεις της ιστοσελίδας ο αριθμός των εξερχόμενων συνδέσμων συνήθως υπερτερεί του αριθμού των εισερχόμενων συνδέσμων. Ωστόσο, η συμπερίληψη στο περιεχόμενο εξερχομένων υπερσυνδέσμων σε δημοφιλείς ιστοσελίδες συμβάλλει σε βάθος χρόνου στην αύξηση του ποσοστού επισκεψιμότητας της ιστοσελίδας. Γενικώς, οι εξερχόμενοι υπερσύνδεσμοι μίας ιστοσελίδας δεν πρέπει να αντιστοιχούν αποκλειστικά σε ιστοσελίδες παραπομπής, δηλαδή σε ιστοσελίδες οι οποίες περιέχουν εξερχόμενο σύνδεσμο προς την πρώτη ιστοσελίδα. Συνίσταται η συμπερίληψη εξερχόμενων συνδέσμων και προς άλλες ιστοσελίδες υπό την προϋπόθεση ότι δεν θα έχουν αρνητικό αποτέλεσμα στην κατάταξη της ιστοσελίδας.

Η ποιότητα του περιεχομένου των ιστοσελίδων που αντιστοιχούν στους εισερχόμενους και εξερχόμενους συνδέσμους είναι σημαντικότερη της ποσότητας των υπερσυνδέσμων σε μία ιστοσελίδα. Η δημιουργία εξερχομένων υπερσυνδέσμων σε ιστοσελίδες, των οποίων το ποσοστό επισκεψιμότητας είναι υψηλότερο, είναι ευκολότερη της ύπαρξης συνδέσμων σε μία ιστοσελίδα από ποιοτικές σελίδες παραπομπής. Η κυριότερη μέθοδος αύξησης του αριθμού των παραπομπών σε μία ιστοσελίδα είναι η αποστολή αιτήματος στους διαχειριστές ιστοτόπων με ποιοτικό περιεχόμενο για την εισαγωγή υπερσυνδέσμων, οι οποίο θα «δείχνουν» στην πρώτη. Η θετική απάντηση στο αίτημα σύνδεσης (link request) προϋποθέτει τη επαρκή δικαιολόγηση, π.χ., προώθηση των προϊόντων μίας εταιρείας, από την οποία ζητείται υπερσύνδεσμος.

Τέλος, απαιτείται συνεχής έλεγχος εκ μέρους του διαχειριστή της ιστοσελίδας αφενός της δυνατότητας πρόσβασης για τις οποίες υπάρχει εξερχόμενος σύνδεσμος και αφετέρου του βαθμού συσχέτισης των περιεχομένων των ιστοσελίδων, οι οποίες αντιστοιχούν στο σύνολο των υπερσυνδέσμων, με το περιεχόμενο της συγκεκριμένης ιστοσελίδας. Σύμφωνα με τα παραπάνω, ο έλεγχος των υπερσυνδέσμων αποκτά ιδιαίτερα κρίσιμο ρόλο στην περίπτωση των δυναμικών ιστοσελίδων. Ο έλεγχος και η διαχείριση του συνόλου των υπερσυνδέσμων πραγματοποιείται συνήθως είτε από

κατάλληλο λογισμικό (λογισμικό διαχείρισης συνδέσμων), όπως το NetMap είτε από τις υπηρεσίες οι οποίες προσφέρονται από τις μηχανές αναζήτησης[21].

5.2 Τεχνικές συλλογής εισερχόμενων συνδέσμων

Οι εισερχόμενοι σύνδεσμοι μίας ιστοσελίδας θεωρούνται από τις μηχανές αναζήτησης ως αναγνώριση της σημαντικότητας του περιεχομένου της στα πλαίσια μίας διαδικτυακής κοινότητας με συγκεκριμένη θεματολογία. Συνεπώς, ένας μεγάλος αριθμός εισερχόμενων υπερσυνδέσμων προερχομένων από ιστοσελίδες με ποιοτικό περιεχόμενο συνεπάγεται καλύτερη θέση στον κατάλογο των αποτελεσμάτων αναζήτησης. Οι τεχνικές οι οποίες χρησιμοποιούνται για τη συλλογή εισερχομένων συνδέσμων συνοψίζονται στα ακόλουθα σημεία.

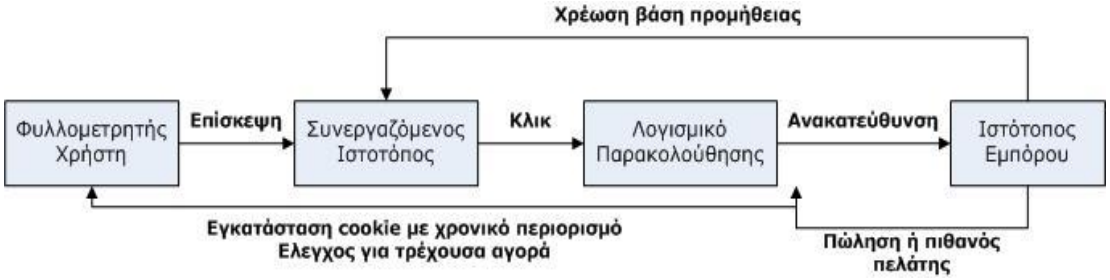
- *Αποστολή αιτήματος σύνδεσης*: Αποτελεί την πιο παραδοσιακή μέθοδο συλλογής εισερχόμενων συνδέσμων. Η συγκεκριμένη τεχνική προϋποθέτει την μελέτη της διαδικτυακής κοινότητας και τον εντοπισμό των ιστοτόπων με ποιοτικό περιεχόμενο. Κατόπιν, ακολουθεί επικοινωνία με τους διαχειριστές των συγκεκριμένων ιστοτόπων και ζητείται η δημιουργία ενός υπερσυνδέσμου προς την ιστοσελίδα. Συνήθως η απάντηση του διαχειριστή αργεί ενώ σε ορισμένες περιπτώσεις δεν αποστέλλεται ή είναι αρνητική. Συνεπώς ο αριθμός των αιτημάτων σύνδεσης ενδέχεται να είναι δυσανάλογα μεγάλος των νέων εισερχόμενων συνδέσμων.
- *Συγγραφή άρθρων*: Μία από τις αποτελεσματικότερες μεθόδους δημιουργίας παραπομπών σε μία ιστοσελίδα είναι η συγγραφή άρθρου και η δημοσίευσή του σε ιστοτόπους οργανισμών και επιχειρήσεων χωρίς καταβολή αντιτίμου υπό την προϋπόθεση ότι στο περιεχόμενό τους συμπεριλαμβάνεται μία παράγραφος με πληροφορίες σχετικά με τον συγγραφέα του άρθρου καθώς και έναν υπερσύνδεσμο στην ιστοσελίδα του τελευταίου. Σε αντίθεση με την προηγούμενη μέθοδο, η συγκεκριμένη μέθοδος είναι πιο αποτελεσματική εφόσον οι διαχειριστές αναζητούν περιεχόμενο υψηλής ποιότητας και εγκυρότητας προκειμένου να το συμπεριλάβουν

[21] Ο διαχειριστής μίας ιστοσελίδας μπορεί να χρησιμοποιήσει τη σύνταξη «`link: διεύθυνση`» στη μηχανή αναζήτησης Google προκειμένου να ενημερωθεί για τους εισερχόμενους συνδέσμους.

στις ιστοσελίδες του ιστοτόπου. Τα προηγούμενα προϋποθέτουν τη δημιουργία καλογραμμένου και ακριβούς περιεχομένου.

- *Δημοσίευση σε ιστολόγια*: Τα ιστολόγια περιγράφονται ως ηλεκτρονικά χρονογραφήματα και αποτελούν μία μέσα για τη δημοσίευση του υπερσυνδέσμου μίας ιστοσελίδας. Στα ιστολόγια περιλαμβάνονται σχόλια αναπληροφόρησης (αναδρομής) από άλλους ιστοτόπους ή από τους συντάκτες του ιστολογίου. Η συχνότητα μπορεί να είναι ωριαία, εβδομαδιαία ή και μικρότερη αλλά συνηθίζονται οι ημερήσιες ενημερώσεις. Συνήθως προηγείται επικοινωνία με το διαχειριστή του ιστολογίου στον οποίο γνωστοποιούνται πληροφορίες σχετικώς με τον οργανισμό, το προσφερόμενο προϊόν κ.α. Οι συγκεκριμένες πληροφορίες χρησιμοποιούνται κατόπιν από το διαχειριστή του ιστολογίου για τη δημοσίευση κατάλληλου άρθρου. Ωστόσο, δεν είναι δυνατός ο εποπτικός έλεγχος του περιεχομένου μίας δημοσίευσης σε ένα ιστολόγιο με συνέπεια να είναι πιθανή η καταγραφή αρνητικών σχολίων.
- *Δελτία τύπου*: Τα δελτία τύπου συνιστούν συνηθισμένη πρακτική στη διαμόρφωση ηλεκτρονικών δημοσίων σχέσεων. Η αποτλεσματικότητα της συγκεκριμένης μεθόδου υποχρεώνει πολλούς οργανισμούς να μισθώνουν επιχειρήσεις, οι οποίες διανέμουν τα δελτία τύπου όσο το δυνατόν ευρύτερα. Ιστότοποι νέων οργανισμών, εκδοτικών οίκων ακόμα και μερικά φόρουμ ενδέχεται να δημοσιεύσουν δελτία τύπου με άμεση συνέπεια την αύξηση του αριθμού των υπερσυνδέσμων στις ιστοσελίδες του ιστοτόπου. Η αποδοτικότητα της συγκεκριμένης μεθόδου προϋποθέτει την παρουσία ενός υπερσυνδέσμου στο δελτίο τύπου, ο οποίος «δείχνει» στην ιστοσελίδα της οποία ο βαθμός δημοτικότητας επιδιώκεται να βελτιωθεί.
- *Προγράμματα συνεργασιών*: Τα προγράμματα συνεργασιών συνιστούν μία κατηγορία διαδικτυακών διαφημίσεων επί πληρωμή. Συνήθως αναφέρονται και ως «διαφημίσεις μηδενικού κινδύνου». Στα συγκεκριμένα προγράμματα ο ηλεκτρονικός λιανοπωλητής πληρώνει ιστοτόπους που αναφέρονται σε αυτόν με σκοπό την μεγένθυση των πωλήσεων και δευτερευόντως την αύξηση του ποσοστού επισκεψιμότητας. Στην πρώτη περίπτωση το κόστος για τον ηλεκτρονικό λιανοπωλητή υπολογίζεται ανά απόκτηση προϊόντος ενώ στη δεύτερη περίπτωση ανά κλικ. Το κύριο πλεονέκτημα των προγραμμάτων είναι ότι οι διαφημιζόμενοι δεν πληρώνουν μέχρι να αγοραστεί το προϊόν ή να βρεθεί ένας υποψήφιος πελάτης. Τα

προγράμματα συνεργασιών είναι ιδιαιτέρως δημοφιλή στους λιανοπωλητές, επειδή η πλειονότητα αυτών πραγματοποιούν πάνω από το 20% των πωλήσεων τους στο διαδίκτυο μέσω των συνεργατών τους (οι οποίοι στη βιβλιογραφία αναφέρονται και ως «συναθροιστές», επειδή συγκεντρώνουν προσφορές από διαφορετικούς παρόχους). Το ηλεκτρονικό βιβλιοπωλείο Amazon ήταν ένας από τους πρώτους αποδέκτες των προγραμμάτων συνεργασιών και σήμερα διαθέτει χιλιάδες συνεργάτες που καθοδηγούν τους επισκέπτες στην ιστοσελίδα της Amazon μέσω υπερσυνδέσμων. Η αμοιβή καθορίζεται με προμήθεια επί των πωληθέντων προϊόντων. Προκειμένου να διαχειριστούν τη διαδικασία εύρεσης συνεργατών, διάχυσης πληροφοριών για τα προϊόντα προς διάθεση, παρακολούθησης επιλογών συνδέσμων και διεκπεραίωσης πληρωμών, πολλοί εταιρικοί οργανισμοί χρησιμοποιούν τα δίκτυα συνεργασιών. Στο ακόλουθο σχήμα δίνεται το μοντέλο λειτουργίας των προγραμμάτων συνεργασίας. Σημειώνεται ότι η διαχείριση του λογισμικού παρακολούθησης και πληρωμής της γίνεται συνήθως από το συνεργαζόμενο διαχειριστή δικτύου.



Σχήμα 23 Το μοντέλο προγραμμάτων συνεργασιών

Ωστόσο υπάρχει προβληματισμός σχετικώς με την βαρύτητα των συγκεκριμένων υπερσυνδέσμων στη διαδικασία βαθμολόγησης των ιστοσελίδων. Συγκεκριμένα, θεωρείται ότι οι ιστοσελίδες με υπερσυνδέσμους επί πληρωμή ευνοούνται έναντι των υπολοίπων ιστοσελίδων με αποτέλεσμα να αμφισβητείται η αμεροληψία των μηχανών αναζήτησης. Εντούτοις, η πλειοψηφία των μηχανών αναζήτησης θεωρούν τα προγράμματα συνεργασιών ως αποδεκτή επιχειρηματική πρακτική δίχως επίπτωση στη θέση της ιστοσελίδας στους καταλόγους αποτελεσμάτων υπό την προϋπόθεση ότι δεν αποτελούν την αποκλειστική πηγή εισερχόμενων υπερσυνδέσμων σε μία ιστοσελίδα.

- *Εισαγωγή διασταυρούμενων υπερσυνδέσμων σε ιστοτόπους με κοινό διαχειριστή*: Η πρακτική εισαγωγής υπερσυνδέσμων μεταξύ ιστοσελίδων διαφόρων ιστοτόπων με κοινό διαχειριστή είναι αποδεκτή από τις μηχανές αναζήτησης. Ωστόσο, η σκόπιμη δημιουργία ιστοτόπων των οποίων οι θα περιέχουν υπερσυνδέσμους σε τρίτους ιστοτόπους με αποκλειστικό σκοπό τη νόθευση της θέσης των ιστοσελίδων των τελευταίων στις φυσικές λίστες αποτελεσμάτων των μηχανών αναζήτησης αποτελεί απορριπτέα μέθοδο.
- *Υπερσύνδεσμοι με κόστος ανά κλικ και υπερσύνδεσμοι επί πληρωμή*: Η πρακτική της συμπερίληψης υπερσύνδεσμών με κόστος ανά κλικ είναι ανάλογη των συμβατικών διαφημίσεων. Οταν ο χρήστης μίας μηχανής αναζήτησης υποβάλλει ένα ερώτημα στη μηχανή αναζήτησης, ένας υποκατάλογος διαφημιστικών κειμένων εμφανίζεται στο δεξιό μέρος του καταλόγου των αποτελεσμάτων της μηχανής αναζήτησης. Σε αντίθεση με τη συμβατική διαφήμιση, ο διαφημιζόμενος δεν πληρώνει την προβολή της διαφήμισης αλλά την επιλογή (κλικ) του υπερσυνδέσμου του από το χρήστη της μηχανής αναζήτησης.

  Οι υπερσύνδεσμοι επί πληρωμή απαιτούν την καταβολή χρηματικού ποσού για την συμπερίληψη ενός υπερσυνδέσμου σε μία ιστοσελίδα. Ωστόσο, πρέπει να αποφεύγεται η συμπερίληψη εισερχόμενων υπερσυνδέσμων επί πληρωμή οι οποίοι προέρχονται από συστοιχίες υπερσυνδέσμων (link farm).

5.3 Τεχνικές επιλογής εξερχομένων συνδέσμων

Εμπειρικές μελέτες αναφέρουν ότι το καλύτερο σχέδιο δράσης βελτιστοποιήσης της θέσης μίας ιστοσελίδας στις καταλόγους αποτελεσμάτων των μηχανών αναζήτησης είναι η ύπαρξη ενός ισσοροπημένου μίγματος εισερχομένων και εξερχομένων υπερσυνδέσμων. Σε ορισμένες περιπτώσεις ο κύριος λόγος της επίσκεψης μίας ιστοσελίδας από το χρήστη του διαδικτύου είναι οι αναφορές της (με μορφή εξερχόμενων υπερσυνδέσμων) σε άλλες ιστοσελίδες για την εύρεση πληροφοριών. Επιπλέον, οι εξερχόμενοι σύνδεσμοι πιστοποιούν το επίπεδο γνώσης και εξειδίκευσης σε μία θεματική περιοχή. Με άλλα λόγια, όταν οι επισκέπτες μίας ιστοσελίδας μεταβαίνουν σε τρίτες ιστοσελίδες επιλέγοντας τους υπερσυνδέσμους, οι οποίοι βρίσκονται στο περιεχόμενο της πρώτης, και διαπιστώνουν ότι ικανοποιούν τις

πληροφοριακές τους ανάγκες, τότε δημιουργείται ένα επίπεδο εμπιστοσύνης μεταξύ των χρηστών και της ιστοσελίδας. Η εμπιστοσύνη των χρηστών στο περιεχόμενο μίας ιστοσελίδας καθορίζει τον βαθμό επισκεψιμότητάς της.

Κατά τη επιλογή των εξερχόμενων συνδέσμων πρέπει να λαμβάνονται υπόψη ορισμένα κριτήρια από τους διαχειριστές των ιστοτόπων. Τα κριτήρια επιλογής των εξερχομένων συνδέσμων συνοψίζονται στα ακόλουθα σημεία.

- *Λογική συσχέτιση περιεχομένου και εξερχομένων υπερσυνδέσμων*: Το συγκεκριμένο κριτήριο δεν είναι αυστηρό με την έννοια ότι επιτρέπονται υπερσύνδεσμοι σε ιστοσελίδες των οποίων το περιεχόμενο δε συσχετίζεται με το περιεχόμενο της ιστοσελίδας. Ωστόσο, η επιλογή τέτοιων συνδέσμων οδηγεί το χρήστη σε ιστοσελίδες άσχετου θεματικού περιεχομένου.
- *Αποφυγή της κατάχρησης των υπερσυνδέσμων*: Ο υπερβολικός αριθμός εξερχομένων υπερσυνδέσμων σε μία ιστοσελίδα είναι ιδαίτερα ενοχλητικός για τον επισκέπτη και αυτό λαμβάνεται υπόψη από τον αλγόριθμο βαθμολόγησης των μηχανών αναζήτησης. Ενδείκνυται η συμπερίληψη όχι περισσοτέρων των δύο εξερχομένων συνδέσμων σε κάθε ιστοσελίδα.
- *Χρήση συνοδευτικού υπερκειμένου σε κάθε υπερσύνδεσμο*: Η χρήση μίας φράσης-κλειδί της ιστοσελίδας στο υπερκείμενο ενός υπερσυνδέσμου αποτελεί τεχνική για την βελτίωση της θέσης της στις φυσικές λίστες των αποτελεσμάτων των μηχανών αναζήτησης υπό την προϋπόθεση ότι η φράση-κλειδί προσδιορίζει και το περιεχόμενο της συνδεδεμένης ιστοσελίδας.
- *Εξέταση του περιεχομένου των ιστοσελίδων παραπομπής*: Συνίσταται η αποφυγή εισαγωγής υπερσυνδέσμων, οι οποίοι «δείχνουν» σε ιστοσελίδες με περιεχόμενο χαμηλής ποιότητας. Ως παραδείγματα αναφέρονται σύνδεσμοι σε ιστοσελίδες με ιογενές περιεχόμενο (spam site) καθώς και σε συστοιχίες υπερσυνδέσμων.

  Οι μηχανές αναζήτησης επιβάλλουν «ποινή» στις ιστοσελίδες, οι οποίες περιέχουν συνδέσμους σε συστοιχίες υπερσυνδέσμων, εφόσον οι τελευταίες δεν προσφέρουν καμμία πληροφορία αξίας στους χρήστες του διαδικτύου. Συνήθως ο διαχειριστής ενός ιστοτόπου γίνεται δέκτης ηλεκτρονικών αιτημάτων απο συστοιχίες υπερσυνδέσμων για τη αμοιβαία συμπερίληψη συνδέσμων. Συνεπώς, σε

περιπτώσεις τέτοιων αιτημάτων απαιτείται αυστηρός έλεγχος του βαθμού συσχέτισης του περιεχομένου των ιστοσελίδων παραπομπής και της ιστοσελίδας αναφοράς. Σε αντίθετη περίπτωση είναι πιθανή η συμπερίληψη υπερσυνδέσμου σε ιστοσελίδες αμφιβόλου και ετερόκλητου περιεχομένου. Οι μηχανές αναζήτησης καταλογίζουν δόλο στο διαχειριστή του ιστοτόπου για την συμπερίληψη υπερσυνδέσμων σε συστοιχίες συνδέσμων και συνεπώς επιβάλλουν ποινή στη ιστοσελίδα υποβαθμίζοντάς την στους καταλόγους αποτελεσμάτων. Στα ακόλουθα σημεία συνοψίζονται τα χαρακτηριστικά των συστοιχιών υπερσυνδέσμων, τα οποία πρέπει να λαμβάνονται υπόψη από τους διαχειριστές των ιστοτόπων.

- Δεν υπάρχει σημασιολογική συσχέτιση μεταξύ του περιεχομένου που διαπραγματεύονται οι ιστοσελίδες ενός ιστοτόπου και του θέματος των ιστοσελίδων που ανήκουν σε μία συστοιχία υπερσυνδέσμων.
- Οι διευθύνσεις διαδικτύου των υπερσυνδέσμων μίας συστοιχίας είναι μακροσκελείς και πολύπλοκες.
- Οι συστοιχίες υπερσυνδέσμων αποδέχονται σχεδόν οποιαδήποτε διεύθυνση αποστέλλεται σε αυτές. Συνεπώς, είναι σχεδόν βέβαιη η δημιουργία ενός συνόλου συνδεδεμένων ιστοσελίδων με ετερόκλητο περιεχόμενο.

- *Ελεγχος της δυνατότητας πρόσβασης των ιστοσελίδων που αντιστοιχούν στους εξερχόμενους υπερσυνδέσμους*: Η μέχρι τώρα εμπειρία στις μεθόδους βελτιστοποίησης αποτελεσμάτων μηχανών αναζήτησης καταδεικνύει ότι είναι οφελιμότερη η απουσία ενός υπερσυνδέσμου από την παρουσία ενός υπερσυνδέσμου, ο οποίος παραπέμπει σε μία μη προσβάσιμη ιστοσελίδα. Ο δυναμικός χαρακτήρας του διαδικτύου επιβάλλει τον περιοδικό έλεγχο της προσβασιμότητας και του περιεχομένου των υπερσυνδέσμων παραπομπής. Η συμπερίληψη ενός υπερσυνδέσμου, του οποίου το περιεχόμενο δεν είναι προσβάσιμο για μεγάλο χρονικό, λαμβάνεται από τις μηχανές αναζήτησης ως ένδειξη ότι η ιστοσελίδα, που περιέχει το συγκεκριμένο σύνδεσμο, δεν συντηρείται με άμεση συνέπεια την υποβάθμισή της στους καταλόγους αποτελεσμάτων.

Συνοψίζοντας, ενδείκνυται η συμπερίληψη εξερχομένων υπερσυνδέσμων, οι οποίοι παραπέμπουν σε ιστοσελίδες με σχετικό και ποιοτικό περιεχόμενο το οποίο ικανοποιεί

τις πληροφοριακές ανάγκες του χρήστη. Επιπλέον, απαιτείται έλεγχος εκ μέρους του διαχειριστή του ιστοτόπου της δυνατότητας πρόσβασης στις ιστοσελίδες των εξερχομένων υπερσυνδέσμων.

5.4 Η χρήση των εσωτερικών συνδέσμων

Οι εσωτερικοί σύνδεσμοι συνδέουν ιστοσελίδες οι οποίες ανήκουν στον ίδιο ιστότοπο. Η απουσία εσωτερικών συνδέσμων συνήθως οδηγεί στην μερική ανάκτηση του συνόλου των ιστοσελίδων ενός ιστοτόπου από το πρόγραμμα ιχνηλάτησης της μηχανής αναζήτησης. Εκτός της παρεχόμενης ευκολίας προς τις μηχανές αναζήτησης, οι εσωτερικοί σύνδεσμοι διευκολύνουν τον επισκέπτη στην περιήγησή του στις ιστοσελίδες του ιστοτόπου.

Οι πιο κοινές πρακτικές για την εσωτερική σύνδεση την ιστοσελίδων ενός ιστοτόπου είναι η συμπερίληψη συνδέσμων στο κείμενο των ιστοσελίδων καθώς και η παρουσία συνδέσμων στις περιοχές του υποσέλιδου και της πλοήγησης στην ιστοσελίδα. Ενδείκνυται η παρουσία φράσεων κλειδιών στο συνοδευτικό κείμενο των εσωτερικών υπερσυνδέσμων. Οι εσωτερικοί σύνδεσμοι χρησιμοποιούνται στη σύνδεση των ιστοσελίδων ενός ιστοτόπου ενώ δεν υπάρχει αυστηρός περιορισμός στον αριθμό των επαναλήψεών τους στο περιεχόμενο των ιστοσελίδων.

6 Ο ρόλος του περιεχομένου στην κατάταξη μίας ιστοσελίδας

Δοθέντος ενός ερωτήματος, η κατάταξη μίας ιστοσελίδας στο σύνολο των αποτελεσμάτων, το οποίο επιστρέφεται από την μηχανή αναζήτησης, καθορίζεται από το συνδυασμό του βαθμού δημοτικότητας και του βαθμού περιεχομένου. Στο κεφάλαιο 4 παρουσιάστηκαν οι αλγόριθμοι HITS και PageRank. Ο αλγόριθμος PageRank αποδίδει έναν καθολικό βαθμό δημοτικότητας σε όλες τις ιστοσελίδες, οι οποίες έχουν ανακτηθεί από το πρόγραμμα ιχνηλάτησης της μηχανής αναζήτησης ενώ ο αλγόριθμος HITS αποδίδει δύο βαθμους, το βαθμό αυθεντίας και το βαθμό αναφοράς, σε κάθε ιστοσελίδα η οποία είναι σχετική με το ερώτημα. Χαρακτηριστικό γνώρισμα των δύο αλγορίθμων είναι η χρήση ολόκληρου ή μέρους του γράφου του διαδικτύου.

Στο παρόν και στο επόμενο κεφάλαιο θα εξεταστούν οι παράγοντες οι οποίοι καθορίζουν το βαθμό περιεχομένου μίας ιστοσελίδας. Πρέπει να σημειωθεί ότι σε αντίθεση με το βαθμό δημοτικότητας δεν θα παρουσιαστεί μία αυστηρή αλγοριθμική διαδικασία εφόσον κάθε μηχανή αναζήτησης διαθέτει το δικό της εξελισσόμενο αλγόριθμο, ο οποίος λαμβάνει υπόψη του μία πληθώρα συντελεστών στάθμισης για την αξιολόγηση του περιεχομένου της ιστοσελίδας. Ο εντοπισμός των παραγόντων και ο προσδιορισμός του βάρους τους στη διαμόρφωση του βαθμού περιεχομένου, όπως αυτός τελικώς δίνεται από τις εμπορικές μηχανές αναζήτησης, πραγματοποιείται με πειραματικές μελέτες. Στις συγκεκριμένες μελέτες παρατηρούνται οι μεταβολές στην κατάταξη μίας ιστοσελίδας συναρτήσει διαφόρων χαρακτηριστικών στο περιεχόμενό της, όπως η συχνή επανάληψη φράσεων κλειδιών στο κύριο κείμενό της (το οποίο οριοθετείται από τις HTML ετικέτες `<body>` και `<\body>`), η χρήση ετικετών μεταδεδομένων στον κώδικα HTML, κ.α.

Στο παρόν κεφάλαιο εξετάζονται οι φράσεις κλειδιά, των οποίων η κατάλληλη επιλογή χαρακτηρίζει τη σημασιολογία του περιεχομένου των ιστοσελίδων ενός ιστοτόπου. Παρατίθονται ευριστικές τεχνικές προσδιορισμού των φράσεων κλειδιών σε μία ιστοσελίδα, οι οποίες είναι δηλωτικές του θέματος που. Επίσης, δίνονται εμπειρικές οδηγίες (οι οποίες αφορούν τους δημιουργούς των ιστοσελίδων) σχετικώς με τη θέση των φράσεων κλειδιών στον κώδικα της ιστοσελίδας και τη συχνότητα επανάληψής τους ώστε να επιτυγχάνεται η όσο το δυνατόν υψηλότερη θέση στους καταλόγους

αποτελεσμάτων των μηχανών αναζήτησης. Στο κεφάλαιο 7 παρουσιάζονται τα αποτελέσματα πειραματικής μελέτης σχετικώς με τον ρόλο της ύπαρξης φράσεων κλειδιών εντός των μεταδεδομένων περιεχομένου μίας ιστοσελίδας στην κατάταξή της στη λίστα αποτελεσμάτων.

6.1 Η σημασία των φράσεων κλειδιών στην αναζήτηση πληροφορίας

Οι φράσεις κλειδιά προσδιορίζουν τη σημασιολογία του περιεχομένου μίας ιστοσελίδας ή ενός ιστοτόπου. Μία φράση κλειδί ορίζεται ως ένα σύνολο λέξεων-όρων, συνήθως αποτελούμενο από μία ως πέντε λέξεις. Αποτελεί την φράση βάσει της οποίας ο χρήστης του διαδικτύου θα διαμορφώσει το ερώτημά του με ισχυρή πιθανότητα, το οποίο θα υποβάλλει κατόπιν στην μηχανή αναζήτησης προκειμένου να ικανοποίησει τις πληροφοριακές του ανάγκες. Οι τελευταίες συνίστανται στην εύρεση ενός «ποιοτικού» ιστοτόπου με αξιόπιστες πληροφορίες σε μία συγκεκριμένη θεματική περιοχή. Οσο περισσότεροι είναι οι συνιστώμενοι όροι της φράσης κλειδί, που χρησιμοποιούνται στο σχηματισμό του ερωτήματος, τόσο περισσότερο στοχευμένη είναι η αναζήτηση του χρήστη.

Η σημασία των φράσεων κλειδιών στην αναζήτηση πληροφορίας στο διαδίκτυο γίνεται κατανοητή με το ακόλουθο παράδειγμα. Εστω ένας μέσος χρήστης του διαδικτύου ο οποίος αναζητά πληροφορίες για ένα εστιατόριο. Ο όρος «εστιατόριο» είναι γενικός και θεωρείται σχεδόν βέβαιη η επιστροφή ενός μεγάλου καταλόγου αποτελεσμάτων, ο οποίος είναι μη διαχειρίσιμός από το χρήστη εντός λογικών χρονικών πλαισίων. Η αναζήτηση γίνεται πιο στοχευμένη με την εισαγωγή ενός επιπρόσθετου όρου στο ερώτημα, π.χ., «κινέζικο εστιατόριο». Το απλό αυτό παράδειγμα σκιαγραφεί σε γενικές γραμμές τον ρόλο των φράσεων κλειδιών στη διαμόρφωση των λιστών αποτελεσμάτων καθώς και στη διαδικασία της βελτιστοποίησης αποτελεσμάτων των μηχανών αναζήτησης. Συνεπώς, η επιτυχής επιλογή μίας φράσης κλειδί ισορροπεί μεταξύ δύο γενικών κατευθύνσεων: αφενός, η φράση κλειδί χαρακτηρίζει το περιεχόμενο της ιστοσελίδας. Ο δημιουργός ιστοτόπων επιλέγει φράσεις κλειδιά, οι οποίες είναι δηλωτικές του περιεχομένου, π.χ., περιγραφή ενός προϊόντος. Αφετέρου, η φράση κλειδί στοχοποιεί το περιεχόμενο της ιστοσελίδας. Η επιλογή της φράσης κλειδί κρίνεται εύστοχη όταν είναι πιθανή η χρήση της από τον μέσο χρήστη της μηχανής αναζήτησης στον σχηματισμό των ερωτημάτων.

Συνεπώς, ο εντοπισμός της φράσης κλειδί και η εισαγωγή της στην κατάλληλη θέση στο κώδικα της ιστοσελίδας αποτελεί κύρια συνιστώσα της προσπάθειας βελτίωσης του ποσοστού επισκεψιμότητας στο διαδίκτυο. Η πλειοψηφία των χρηστών του διαδικτύου εντοπίζει τους ιστοτόπους με ποιοτικό περιεχόμενο μέσω των μηχανών αναζήτησης. Η υψηλή θέση στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης αποδεικνύεται ορισμένες φορές αποτελεσματικότερης της πρακτικής των διαφημίσεων επί πληρωμή για την αύξηση του ποσοστού επισκεψιμότητας

6.2 Επιλογή των κατάλληλων φράσεων κλειδιών

Η επιλογή των κατάλληλων φράσεων κλειδιών προαπαιτεί την απάντηση του βασικού ερωτήματος περί της κατηγορίας των χρηστών του διαδικτύου, οι οποίοι προσπελαύνουν ή πρόκειται να προσπελάσουν το περιεχόμενο των ιστοσελίδων του ιστοτόπου. Η απάντηση στο προηγούμενο ερώτημα είναι σημαντική για την εύρεση των φράσεων κλειδιών. Ως παράδειγμα αναφέρεται ιστότοπος μέσω του οποίου πραγματοποιείται εμπόριο σαπουνιών. Φράσεις κλειδιά, όπως «σαπούνι», «πολυτελή προϊόντα μπάνιου», «σαπούνι ειδικής σύστασης», «αρωματικά σαπούνια» κ.α αποτελούν πιθανές φράσεις κλειδιά, οι οποίες μπορούν να χρησιμοποιηθούν από τον δημιουργό περιεχομένου του ιστοτόπου. Η φράση κλειδί «σαπούνι» αποτελεί όρο με γενική σημασία σε αντίθεση με τους υπόλοιπους όρους του παραδείγματος, οι οποίοι εξειδικεύουν στο προϊόν. Επιπλεόν, ο δημιουργός του περιεχομένου πρέπει να προτείνει φράσεις κλειδιά των οποίων η παρουσία στα ερωτήματα του μέσου χρήστη προς την μηχανή αναζήτησης είναι.

6.2.1 Κατηγοριοποίηση των φράσεων κλειδιών

Διακρίνονται δύο είδη κατηγοριών φράσεων κλειδιών ως προς τη σημασιολογία τους.

- *Ταυτότητα εμπορικού σήματος (trade mark)*: Το σύνολο των συσχετισμών του εμπορικού σήματος που θα πρέπει να κοινοποιηθούν συμπεριλαμβανομένων της ονομασίας και των συμβόλων. Εξ' ορισμού η ταυτότητα της μάρκας δεν περιορίζεται στην ονομασία της εταιρείας ή του προϊόντος. Οταν ένας εταιρικός οργανισμός επαναπροσδιορίζει ή παρουσιάζει εκ νέου τη δικτυακή παρουσία του έχει τις ακόλουθες επιλογές όσον αφορά την χρήση της ταυτότητας του εμπορικού σήματος στο περιεχόμενο της ιστοσελίδας

- *Μεταφορά του εμπορικού σήματος στο διαδίκτυο*: Αποτελεί την πιο συνηθισμένη προσέγγιση. Εταιρείες με εδραιωμένα εμπορικά σήματα στον πραγματικό κόσμο τα αναπαράγουν στο διαδίκτυο.
- *Επέκταση του παραδοσιακού εμπορικού σήματος*: Περιλαμβάνει τη δημιουργία μίας ελαφρώς διαφορετικής έκδοσης της εταιρικής επωνυμίας. Ακολουθώντας τη συγκεκριμένη προσέγγιση, η εταιρεία αποσκοπεί στη διαφοροποίηση από παρόμοιες ανταγωνίστριες επιχειρήσεις.
- *Συνεργασία με υπάρχουσα ψηφιακή μάρκα*: Προσφέρει την δυνατότητα καλύτερης προώθησης των προϊόντων μίας εταιρείας σε συνεργασία με μία δυνατή ψηφιακή ή διαδικτυακή επωνυμία.
- *Δημιουργία ενός νέου ψηφιακού εμπορικού σήματος*: Είναι δυνατόν να κριθεί απαραίτητη η δημιουργία ενός νέου ψηφιακού εμπορικού σήματος στην περίπτωση κατά την οποία το υπάρχον εμπορικό σήμα δημιουργεί αρνητικούς συνειρμούς ή είναι πολύ παραδοσιακό για το διαδίκτυο.

- *Γενικές λέξεις κλειδιά*: Ορίζονται οι λέξεις κλειδιά, οι οποίες δεν συσχετίζονται άμεσα με την ταυτότητα του οργανισμού ενώ είναι δηλωτικές του περιεχομένου της ιστοσελίδας.

Εκτός της προηγούμενης κατηγοριοποίησης, οι φράσεις-κλειδιά ταξινομούνται στις ακόλουθες δύο ομάδες.

- *Φράσεις-κλειδιά επί πληρωμή*: Βασίζεται στην υψηλότερη προσφορά κόστους ανά κλικ (cost per click) για κάθε φράση κλειδί. Οι φράσεις κλειδιά της συγκεκριμένης κατηγορίας εμφανίζονται συνήθως σε διαφημίσεις κειμένων ενώ η χρήση τους δημοπρατείται.
- *Φράσεις-κλειδιά δίχως πληρωμή*: Καλούνται επίσης οργανικές φράσεις κλειδιά (organic keyword). Δεν απαιτείται η καταβολή τιμήματος για την εμφάνιση ιστοσελίδων που τις περιέχουν στους καταλόγους αποτελεσμάτων.

6.2.2 Εμπειρικές τεχνικές εύρεσης φράσεων κλειδιών

Η διαδικασία εύρεσης των κατάλληλων φράσεων κλειδιών δεν είναι τυποποιημένη. Δίνονται ορισμένες εμπειρικές τεχνικές οι οποίες ακολουθούνται για τον εντοπισμό των πιθανών φράσεων κλειδιών. Πρέπει να σημειωθεί ότι ο δυναμικός χαρακτήρας του

περιεχομένου των περισσότερων ιστοσελίδων επιβάλλει την περιοδική ανάλυση του περιεχομένου για την επιλογή του κατάλληλου συνόλου φράσεων κλειδιών. Με την πάροδο του χρόνου ορισμένες φράσεις κλειδιά ενδέχεται να μην επιφέρουν τα επιθυμητά αποτελέσματα στα πλαίσια της βελτιστοποίησης αποτελεσμάτων των μηχανών αναζήτησης.

- *Εντοπισμός των προφανών φράσεων κλειδιών*.
- *Εξαγωγή γνώσης από τα αρχεία του εξυπηρετητή, στα οποία κρατείται το ιστορικό πρόσβασης του ιστοτόπου*. Η διαδικασία εντοπισμού συνίσταται στην επεξεργασία του περιεχομένου του κειμένου αγκύρωσης σε υπερσυνδέσμους τρίτων ιστοσελίδων, οι οποίοι επελέγησαν από το χρήστη κατά την περιήγησή του στο διαδίκτυο προκειμένου να μεταβεί σε μία ιστοσελίδα του ιστοτόπου.
- *Η εξέταση του συνόλου των φράσεων κλειδιών, οι οποίοι χρησιμοποιούνται στη σύνταξη του περιεχομένου των ιστοτόπων των ανταγωνιστών*. Η γνώση των συγκεκριμένων φράσεων κλειδιών είναι δυνατή με την επισκόπηση του κώδικα HTML της ιστοσελίδας του ανταγωνιστή και ειδικότερα με την ανάκτηση του περιεχομένου του πεδίου μεταδεδομένων `<META NAME="keywords">`. Ωστόσο, είναι συνηθισμένη η απουσία του συγκεκριμένου πεδίου μεταδεδομένων από τον κώδικα των περισσοτέρων ιστοσελίδων.
- *Χρήση λογισμικού εύρεσης φράσεων κλειδιών (keyword suggestion tools)*. Η έξοδος του συγκεκριμένου λογισμικού δίνει μία λίστα προτεινόμενων φράσεων κλειδιών για έναν συγκεκριμένο όρο, ο οποίος αποτελεί την είσοδο στο λογισμικό. Επιπλέον, το λογισμικό παρέχει ορισμένες στατιστικές πληροφορίες σχετικώς με την ανταγωνιστικότητα των φράσεων κλειδιών. Η στατιστική ανάλυση παρέχει πληροφορίες για κάθε προτεινόμενη φράση κλειδί όπως το ποσοστό των χρηστών επί ημερησίας βάσης οι οποίοι τη συμπεριλαμβάνουν στο ερώτημά τους, ο αριθμός των επιτυχών αναζητήσεων δεδομένης ανά φράση κλειδί (δηλαδή ο αριθμός των αναζητήσεων οι οποίοι καταλήγουν σε μία επίσκεψη σε έναν ιστότοπο) καθώς και οι σχετικοί όροι αναζήτησης. Βάσει των στατιστικών στοιχείων υπολογίζεται ο δείκτης αποτελεσματικότητας μίας φράσης κλειδί. Ο δείκτης αποτελεσματικότητας ενός όρου ορίζεται ως το μέτρο σύγκρισης του αριθμού των χρηστών οι οποίοι υποβάλλουν

ένα ερώτημα στην μηχανή αναζήτησης, το οποίο περιλαμβάνει τον συγκεκριμένο όρο, με τον αριθμό των ιστοσελίδων στο κατάλογο για το συγκεκριμένο όρο. Τέλος, προτείνει λανθασμένες ορθογραφίες της φράσης κλειδί οι οποίες καταγράφηκαν ως ερώτηματα στη μηχανή αναζήτησης.

Η ανάλυση των αποτελεσμάτων του λογισμικού εύρεσης φράσεων κλειδιών περιλαμβάνει τον προσδιορισμό αφενός της ανταγωνιστικότητας του όρου αναζήτησης (competitiveness of a term) και αφετέρου της δημοτικότητάς του. Σε αυτή την κατεύθυνση ιδανική επιλογή φράσεων κλειδιών αποτελούν εκείνες, οι οποίες δεν είναι ιδιαίτερα ανταγωνιστικές αλλά απολύτως σχετικές με το περιεχόμενο του ιστοτόπου.

Κατόπιν παρατίθενται τα πιο γνωστά λογισμικά εύρεσης φράσεων κλειδιών καθώς και οι χαρακτηριστικότερες λειτουργίες τους.

- **Overture Keyword Selector Tool**: Δίνει τον αριθμό των αναζητήσεων σε διάστημα ενός μηνός για κάθε φράση κλειδί σχετικής με τον όρο εισόδου στο λογισμικό.
- **Wordtracker**: Δίνει τη συχνότητα αναζήτησης μίας φράσης κλειδιού καθώς και των αριθμό των ανταγωνιστικών ιστοτόπων οι οποίοι τη χρησιμοποιούν στο περιεχόμενό τους. Με τη χρήση του συγκεκριμένου λογισμικού προσδιορίζεται η «ανταγωνιστικότητα» κάθε φράσης κλειδί.
- **Trelian Keyword Discovery Tool**: Δίνει τη δυνατότητα στο χρήστη να εξακριβώσει την αξία μεριδίου αγοράς για ένα δεδομένο όρο αναζήτησης.
- **Google AdWords Keyword Tool**: Δεδομένης μίας φράσης κλεδί παρέχει ένα σύνολο προτεινόμενων φράσεων κλειδιών. Επιπλέον, δίνει τη δυνατότητα υποβολής του περιεχομένου μίας ιστοσελίδας καθώς και των ιστοσελίδων ενός ιστοτόπου και προτείνει φράσεις κλειδιά, η συμπερίληψη των οποίων ενδίκνυται για την αύξηση του ποσοστού επισκεψιμότητας.

- *Αποκλεισμός διφορούμενων φράσεων κλειδιών*. Από το αρχικό σύνολο φράσεων κλειδιών απομακρύνονται οι όροι ή λέξεις, οι οποίες ερμηνεύονται με διαφορετικό τρόπο από διαφορετικές ομάδες χρηστών. Σε πρακτικό επίπεδο είναι δύσκολος ο εντοπισμός τους.

- *Μη συμπερίληψη φράσεων κλειδιών με ευρεία σημασία*: Οι φράσεις κλειδιά με ευρεία σημασία είναι συνήθως αρκετά «ανταγωνιστικοί» όροι με άμεση συνέπεια τη δυσκολία βελτιστοποίησης της θέσης μίας ιστοσελίδας στους καταλόγους αποτελεσμάτων στα πλαίσια της χρησιμοποίησής τους.
- *Δημιουργία φράσεων κλειδιών προερχόμενων από συνδυασμό όρων.*
- *Παράθεση των λανθασμένων ορθογραφιών των φράσεων κλειδιών στο πεδίο μεταδεδομένων* <META NAME=”keywords”>.
- *Προσαρμογή των φράσεων κλειδιών και της ορολογίας στη χώρα στην οποία δημοσιεύεται το περιεχόμενο*: Σε ορισμένες θεματικές περιοχές, η ορολογία διαφέρει ανάλογα με την χώρα ή την περιοχή. Χαρακτηριστικό παράδειγμα αποτελεί η χρήση διαφορετικών φράσεων στη αγγλική γλώσσα για την περιγραφή των υπηρεσιών που προσφέρονται από τους αυτοκινητιστές. Στις Η.Π.Α το επάγγελμα των αυτοκινητιστών περιγράφεται από τη φράση κλειδί «taxi cab» ενώ στο Ηνωμένο Βασίλειο από τον όρο «car hire». Συνεπώς, ο συντάκτης του περιεχομένου μίας ιστοσελίδας οφείλει να γνωρίζει την ορολογία, η οποία χρησιμοποιείται για την περιγραφή των προϊόντων και των υπηρεσιών σε μία χώρα.

6.3 Θέση των φράσεων κλειδιών

Στην προηγούμενη παράγραφο παρουσιάστηκαν ευριστικοί κανόνες προσδιορισμού των φράσεων κλειδιών του περιεχομένου μίας ιστοσελίδας. Το επόμενο βήμα στα πλαίσια της βελτιστοποίησης αποτελεσμάτων των μηχανών αναζήτησης είναι η τοποθέτηση των φράσεων κλειδιών σε κατάλληλα σημεία στον κώδικα της ιστοσελίδας. Παρενθετικά, αναφέρεται ότι ο κατάλογος των φράσεων κλειδιών, οι οποίες είναι δηλωτικές του περιεχομένου του ιστοτόπου, ενδέχεται να είναι μακροσκελής. Η ανάπτυξη του περιεχομένου μίας ιστοσελίδας ενός πρέπει να είναι προσανατολισμένη στην ανάδειξη το πολύ δύο φράσεων κλειδιών, ονομαστικά της κύριας και της δευτερεύουσας. Η εμπειρία υποδεικνύει ότι είναι αρκετά δύσκολη η βελτιστοποίηση του περιεχομένου μίας ιστοσελίδας αριθμό φράσεων κλειδιών μεγαλύτερο του δύο. Ωστόσο, ο συγκεκριμένος περιορισμός δεν αποκλειεί τη χρήση φράσεων κλειδιών, εκτός της πρωτεύουσας και της δευτερεύουσας, στο περιεχόμενο μίας ιστοσελίδας όπου αυτό κρίνεται κατάλληλο. Εκτός της διάχυσης των φράσεων κλειδιών σε όλο το κείμενο της

ιστοσελίδας, τα σημεία του κώδικα HTML, στα οποία συνίσταται η εμφάνιση της πρωτεύουσας ή/και της δευτερεύουσας φράση κλειδί, είναι τα ακόλουθα.

- *Ετικέτα τίτλου της ιστοσελίδας (title tag)*: Οι φράσεις κλειδιά στην ετικέτα τίτλου μίας ιστοσελίδας, της οποίας το περιεχόμενο εμφανίζεται στην κορυφή του παραθύρου του φυλλομετρητή, υποδεικνύονται στον κώδικα HTML με λεκτικό <TITLE>. Αν και δεν είναι υποχρεωτικός, συνίσταται ο προσδιορισμός της ετικέτας τίτλου σε κάθε ιστοσελίδα του ιστοτόπου. Επίσης, ο τίτλος πρέπει να είναι σαφής (40 ως 60 χαρακτήρες) και να αναγράφει το όνομα της εταιρείας και του προϊόντος, της υπηρεσίας ή της προσφοράς που παρουσιάζεται στην ιστοσελίδα. Μεγαλύτερη βαρύτητα πρέπει να δίνεται σε φράσεις κλειδιά στην αριστερή πλευρά της ετικέτας τίτλου ενώ προτείνεται η επανάληψη φράσεων κλειδιών στο κείμενο του τίτλου εντός λογικών πλαισίων. Επιπρόσθετα, η ετικέτα τίτλου HTML μίας ιστοσελίδας είναι ζωτικής σημασίας για το μάρκετινγκ αναζήτησης, αφού τυπικά πρόκειται για το υπογραμμισμένο κείμενο στη σελίδα των αποτελεσμάτων που σχηματίζει τον υπερσύνδεσμο για τη συγκεκριμένη ιστοσελίδα. Εαν η ετικέτα τίτλου, η οποία εμφανίζεται στη σελίδα αποτελεσμάτων αναζήτησης, παρουσιάζει μεγαλύτερη συνάφεια, τότε είναι πολύ πιθανό να προτιμηθεί από τους χρήστες της μηχανής αναζήτησης και κατά συνέπεια να αυξηθεί το ποσοστό επισκεψιμότητάς της.
- *Εναλλακτικό κείμενο γραφικών (alt tag)*: Μια ιστοσελίδα, της οποία το περιεχόμενο περιέχει πολλά γραφικά και προσαρτήματα λογισμικού είναι λιγότερο πιθανό να καταλαμβάνει υψηλή θέση στις οργανικές λίστες των αποτελεσμάτων των μηχανών αναζήτησης. Οι μηχανές αναζήτησης εισάγουν τη συγκεκριμένη σελίδα στο ευρετήριό τους βάσει του κειμένου που σχετίζεται με κάθε γραφική απεικόνιση. Οι φράσεις κλειδιά εισάγονται στο κείμενο το οποίο αποτελεί τιμή του γνωρίσματος ALT στην ετικέτα IMG, π.χ.,

  ```
  <IMG NAME='logo' SRC='logo.gif' ALT="Electronic Marketplace Company Name">
  ```

  Σημειώνεται ότι το κείμενο εντός του γνωρίσματος ALT δεν είναι ορατό από το χρήστη εκτός της περίπτωσης στην οποία η δυνατότητα παρουσίασης γραφικών έχει απενεργοποιηθεί στο φυλλομετρητή. Τέλος, αντενδείκνυται η εισαγωγή κειμένου

εντός των γραφικών, εφόσον οι μηχανές αναζήτησης δεν επεξεργάζονται το συγκεκριμένο κείμενο, το οποίο αποτελεί μέρος του γραφικού.

- *Ετικέτες μεταδεδομένων (μετα-ετικέτες)*: Οι ετικέτες μεταδεδομένων (meta-tags) αποτελούν μέρος του κώδικα HTML της ιστοσελίδας, ενώ το περιεχόμενό τους καθορίζεται από τον δημιουργό της ιστοσελίδας. Το περιεχόμενό των μεταδεδομένων δεν είναι ορατό στην ιστοσελίδα από το χρήστη. Υπάρχουν δύο σημαντικές μεταετικέτες οι οποίες τοποθετούνται στην κορυφή του κώδικα HTML της ιστοσελίδας. Δηλώνονται με τη χρήση της δεσμευμένου λεκτικού `<META NAME="">`.
  i. Στην ετικέτα μεταδεδομένων "keywords" δηλώνονται οι φράσεις κλειδιά της ιστοσελίδας, π.χ., `<META NAME="keywords" content="e-business, e-commerce">`.
  ii. Η μετα-ετικέτα "description" περιέχει τις πληροφορίες οι οποίες θα προβληθούν στη σελίδα αποτελεσμάτων αναζήτησης. Συνεπώς είναι σημαντικό να δοθεί ακριβής περιγραφή του περιεχομένου της ιστοσελίδας, η οποία θα παροτρύνει το χρήστη να επιλέξει το σύνδεσμο του.

  Στην κοινότητα των μηχανικών ανάπτυξης ιστοσελίδων υπάρχει έντονος προβληματισμός σχετικά με το βαθμό βαρύτητας που προσδίδουν οι μηχανές αναζήτησης στη συμπερίληψη φράσεων κλειδιών στις προαναφερθείσες μεταετικέτες. Εκτενής μελέτη για τη συμπερίληψη φράσεων κλειδιών στα μεταδεδομένα ως αποδοτικής τεχνικής βελτιστοποίησης της θέσης στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης παρουσιάζεται στο κεφάλαιο 7.

- *Κείμενο αγκύρωσης σε υπερσυνδέσμους (anchor text)*: Συνίσταται η συμπερίληψη φράσεων κλειδιών στα συνοδευτικά κείμενα των εισερχόμενων και εξερχόμενων υπερσυνδέσμων.
- *Ετικέτες τίτλου κειμένου (header tag)*: Υπάρχουν έξι διαφορετικοί τύποι ετικετών τίτλου οι οποίες δηλώνονται στον κώδικα HTML με τις ετικέτες `<H1>, <H2>, ... <H6>`. Συνίσταται η χρήση ενός κύριου τίτλου (`<H1>`) ανά ιστοσελίδα. Η χρήση των υπολοίπων τίτλων αποσκοπεί στη δομημένη παρουσίαση του περιεχομένου της ιστοσελίδας και κατά συνέπεια στην εύκολη ανάγνωσή του από το χρήστη του

διαδικτύου. Συνήθως οι δευτερεύοντες τίτλοι (<H2>, ... <H6>) περιέχουν όρους οι οποίοι είναι ελαφρώς πιο συγκεκριμένοι από τις φράσεις κλειδιά, οι οποίες χρησιμοποιούνται στη κύρια επικεφαλίδα.

6.4 Συχνότητα επανάληψης των φράσεων κλειδιών

Η συχνότητα επανάληψης μίας φράσης κλειδί (keyword frequency) ορίζεται ως ο λόγος του αριθμού εμφανίσεων της στον κώδικα HTML προς το συνολικό αριθμό των λέξεων στην ιστοσελίδα. Αποτελεί βασικό συντελεστή στάθμισης, ο οποίος επηρεάζει άμεσα τη θέση της ιστοσελίδας στο κατάλογο αποτελεσμάτων των μηχανών αναζήτησης. Το περιεχόμενο πρέπει να συντάσσεται με τέτοιο τρόπο ώστε να αυξηθεί ο αριθμός επαναλήψεων της πρωτεύουσας (και της δευτερεύουσας) φράσης κλειδί. Ωστόσο, σημειώνεται ότι η σκόπιμη πολλαπλή επανάληψη της φράσης κλειδί με αποκλειστικό κίνητρο την αύξηση της πυκνότητας εμφάνισής της φέρνει τα αντίθετα από τα προσδοκώμενα αποτελέσματα εφόσον οι μηχανές αναζήτησης εντοπίζουν με αλγοριθμικό τρόπο τη συγκεκριμένη πρακτική. Επίσης, οι μηχανές αναζήτησης μπορούν να ελέγξουν αλγοριθμικά αν πολλαπλές εμφανίσεις της φράση κλειδί είναι «κρυμμένες» με τη χρήση κειμένου του ιδίου χρώματος και φόντου.

Στην επιστημονική κοινότητα διατυπώνονται ενστάσεις για την απόδοση σημαντικού βάρους στο συντελεστή της συχνότητας επανάληψης μίας φράσης κλειδί. Συγκεκριμένα, η πυκνότητα δεν παρέχει κρίσιμες πληροφορίες όπως η απόσταση μεταξύ δύο εμφανίσεων μίας φράσης κλειδί στο κείμενο της ιστοσελίδας κ.α. Το κύριο μειονέκτημα του συγκεκριμένου μεγέθους εντοπίζεται στο γεγονός ότι δεν αποτελεί αξιόπιστο μέτρο της ποιότητας του περιεχομένου μίας ιστοσελίδας εκτός ελαχίστων ακραίων περιπτώσεων, όπως όταν η συχνότητα επανάληψης είναι μεγαλύτερη του 50%.

7 Μεταδεδομένα

Στο παρόν κεφάλαιο θα διερευνηθεί ο ρόλος των μεταδεδομένων περιεχομένου τον κώδικα της ιστοσελίδας στην κατατακτήρια θέση που κατέχει αυτή στον κατάλογο των αποτελεσμάτων των μηχανών αναζήτησης. Προς αυτή την κατεύθυνση σημαντικά πειραματικά αποτελέσματα παρουσιάζονται στο επιστημονικό άρθρο (10) σύμφωνα με τα οποία η χρήση των μεταδεδομένων υπό ορισμένους προϋποθέσεις συνιστά έναν αποδοτικό μηχανισμό για την βελτίωση της κατακτήριας θέσης μίας ιστοσελίδας.

7.1 Ο ρόλος των μεταδεδομένων στην κατάταξη των αποτελεσμάτων

Ενας ευρύς ορισμός των μεταδεδομένων αναφέρει ότι είναι δεδομένα τα οποία περιγράφουν δεδομένα. Σε γενικές γραμμές τα μεταδεδομένα προσφέρουν έναν αποδοτικό μηχανισμό περιγραφής των δεδομένων. Η ανάπτυξη των μεταδεδομένων απαιτεί την ουσιαστική συνεργασία μεταξύ διαφορετικών ομάδων χρηστών. Στα πλαίσια των τεχνολογιών του παγκόσμιου ιστού το έντονο ερευνητικό ενδιαφέρον για τα μεταδεδομένα οφείλεται στο γεγονός ότι διευκολύνουν τη διαδικασία ταυτοποίησης και εξαγωγής πληροφορίας από το διαδίκτυο.

Η εισαγωγή μεταδεδομένων στον κώδικα HTML μίας ιστοσελίδας δεν είναι υποχρεωτική. Οι δημιουργοί ιστοσελίδων επιλέγουν τα μεταδεδομένα τα οποία θα ενσωματωθούν στον κώδικα της ιστοσελίδας. Από την άλλη πλευρά, οι μηχανές αναζήτησης προσπελαύνουν τα μεταδεδομένα για την εξαγωγή πληροφοριών σχετικών με το περιεχόμενο της ιστοσελίδας, π.χ., φράσεις κλειδιά κ.α. Οι πληροφορίες χρησιμοποιούνται ως όροι δεικτοδότησης στα ευρετήρια των βάσεων δεδομένων της μηχανής αναζήτησης.

Κατά τη διάρκεια των προηγούμενων ετών, έχει επισημανθεί σπουδαιότητα των μεταδεδομένων στις διαδικασίες εξόρυξης πληροφορίας. Σύμφωνα με την μελέτη (11) το ποσοστό των ιστοσελίδων οι οποίες περιέχουν μία η περισσότερες ετικέτες μεταδεδομένων περιεχομένου πλησιάζει το 70%, εκ των οποίων το 15% αφορά μεταδεδομένα περιγραφής του περιεχομένου και 17% αφορά μεταδεδομένα για τις χρησιμοποιούμενες φράσεις. Ωστόσο η απλή παρουσία των μεταδεδομένων δεν εγγυάται την υψηλή θέση στον κατάλογο αποτελεσμάτων των μηχανών αναζήτησης.

Δεν αρκεί η απλή προσθήκη ετικετών META στον κώδικα της ιστοσελίδας και η υποβολή της διεύθυνσής της στις μηχανές αναζήτησης ώστε να αυξηθεί το ποσοστό επισκεψιμότητάς της.

Τα μεταδεδομένα περιεχομένου συντάσσονται με υποκειμενικά και αντικειμενικά κριτήρια από τους δημιουργούν των ιστοσελίδων. Το περιεχόμενο των μεταδεδομένων, στα οποία παρατίθενται πληροφορίες για τον δημιουργό της ιστοσελίδας, την ημερομηνία δημοσίευσής της στον Παγκόσμιο Ιστό, τον αριθμό έκδοσης, καθορίζεται με αντικειμενικά κριτήρια. Αντιθέτως, υπάρχουν μεταδεδομένα των οποίων η σύνταξη βασίζεται σε υποκειμενικά κριτήρια. Στοιχεία μεταδεδομένων στα οποία παρατίθενται οι φράσεις κλειδιά και δίνεται μία συνοπτική περιγραφή του αντικειμένου που διαπραγματεύεται η ιστοσελίδα συντάσσονται με υποκειμενικά κριτήρια. Η επιτυχής σύνταξη των υποκειμενικών μεταδεδομένων μίας ιστοσελίδας βασίζεται στη δυνατότητα πρόσληψης και κατανοήσης του περιεχομένου της. Η καθημεριμή πρακτική στην ανάπτυξη ιστοσελίδων καταδεικνύει την χρήση των μεταδεδομένων εκτός του προαναφερθέντος πλαισίου.

Διακρίνονται οι ακόλουθες τρεις κατευθύνσεις στην ανάλυση της συμβολής των μεταδεδομένων περιεχομένου στην βελτίωση του ποσοστού επισκεψιμότητας μίας ιστοσελίδας.

1. Προσδιορισμός των παραμέτρων στα μεταδεδομένα μίας ιστοσελίδας τα οποία επηρεάζουν την θέση της στην κατάταξη των αποτελεσμάτων (search engine result list) των μηχανών αναζήτησης.
2. Συγκριτική μελέτη της επίδρασης των παραμέτρων, όπως αυτοί προσδιορίστηκαν από την υλοποίηση του πρώτου στόχου, στο πλαίσιο της επίτευξης καλύτερης θέσης στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης.
3. Τέλος, διατύπωση καθολικών στρατηγικών (ανεξαρτήτων της υλοποίησης των μηχανών αναζήτησης) στη σύνταξη των μεταδεδομένων περιεχομένου με σκοπό τη βελτίωση της θέση της ιστοσελίδας στους καταλόγους των αποτελεσμάτων.

Σημειώνεται ότι οι παράγοντες των μεταδεδομένων, των οποίων ο ρόλος θα διερευνηθεί, είναι απολύτως ελέγξιμοι και διαχειρίσιμοι από τους μηχανικούς ανάπτυξης των ιστοσελίδων.

Η εξαγωγή ορθών συμπερασμάτων για τη συμβολή των μεταδεδομένων στη βελτίωση του ποσοστού επισκεψιμότητας της ιστοσελίδας πραγματοποιείται με την εξέταση πέντε στατιστικών υποθέσεων. Γίνεται η παραδοχή ότι το ποσοστό επισκεψιμότητας μίας ιστοσελίδας (web page visibility) καθορίζεται πλήρως από τη θέση που κατέχει αυτή στην κατάταξη των αποτελεσμάτων

1. Η μεταβολή του ποσοστού επισκεψιμότητας είναι ανεξάρτητη της συμπερίληψης μεταδεδομένων στην ιστοσελίδα.
2. Δεν υπάρχει μεταβολή στο ποσοστό επισκεψιμότητας μεταξύ ιστοσελίδων με διαφορετικό αριθμό συνδυασμών στοιχείων των μεταδεδομένων.
3. Δεν υπάρχει μεταβολή στο ποσοστό επισκεψιμότητας μεταξύ των ακολούθων τριών κατηγοριών ιστοσελίδων
   - Ιστοσελίδες στις οποίες υπάρχει το στοιχείο μεταδεδομένου τίτλου (metadata title) στον κώδικα HTML.
   - Ιστοσελίδες στις οποίες υπάρχει το στοιχείο μεταδεδομένου θέματος (metadata subject) στον κώδικα HTML.
   - Ιστοσελίδες στις οποίες υπάρχει το στοιχείο μεταδεδομένου περιγραφής (metadata description) στον κώδικα HTML.
4. Δεν υπάρχει μεταβολή στο ποσοστό επισκεψιμότητας μεταξύ ιστοσελίδων με διαφορετικούς συνδυασμούς των στοιχείων μεταδεδομένων τίτλου, θέματος και περιγραφής.
5. Η πηγή προέλευσης των φράσεων κλειδιών στα μεταδεδομένα δεν παίζει ρόλο στη διαμόρφωση του ποσοστού επισκεψιμότητας της ιστοσελίδας. Υποτίθεται ότι οι φράσεις κλειδιά προέρχονται από τον τίτλο της ιστοσελίδας και το κείμενό της.

Σύμφωνα με την υπόθεση υπ. αριθμ. 1 οι ιστοσελίδες ταξινομούνται σε δύο βασικές κατηγορίες – ιστοσελίδες με μεταδεδομένα και ιστοσελίδες χωρίς μεταδεδομένα. Η εξακρίβωση της ορθότητας της συγκεκριμένης υπόθεσης απαντά στο ερώτημα αν η εισαγωγή των μεταδεδομένων σε μία ιστοσελίδα συμβάλλει στη βελτίωση της θέσης της στην κατάταξη των αποτελεσμάτων. Οι υποθέσεις υπ. αριθ. 2 ως 5 αφορούν την συμβολή των μεταδεδομένων περιεχομένου στη διαμόρφωση του ποσοστού επισκεψιμότητας της ιστοσελίδας. Συγκεκριμένα, η υπόθεση υπ. αριθμ. 5 εστιάζει στην προέλευση των φράσεων κλειδιών, οι οποίες εμφανίζονται στα μεταδεδομένα, ενω οι

υποθέσεις υπ. αριθ. 2, 3 και 4 εξετάζουν την συμβολή αφενός των στοιχείων μεταδεδομένων τίτλου, θέματος και περιγραφής μεμονωμένα, και αφετέρου των διαφόρων συνδυασμών τους. Η διαφορά μεταξύ των υποθέσεων υπ. αριθ. 2 και 4 είναι δισδιάκριτη. Στην υπόθεση υπ. αριθμ. 2 ενδιαφέρει ο αριθμός των στοιχείων μεταδεδομένων που συνιστούν ένα συνδυασμό μεταδεδομένων. Στην υπόθεση υπ. αριθμ. 4 εκτός του αριθμού των στοιχείων μεταδεδομένων σε ένα συνδυασμό μεταδεδομένων λαμβάνεται υπόψη και το περιεχόμενο του συνδυασμού.

7.2 Πειραματικά Δεδομένα

Οι παράμετροι των μεταδεδομένων διακρίνονται σε εσωτερικούς και εξωτερικούς. Οι εσωτερικοί παράμετροι αναφέρονται στη θέση και τη συχνότητα εμφάνισης μίας φράσης-κλειδί σε ένα στοιχείο μεταδεδομένου. Η φράση κλειδί είναι αντιπροσωπευτική του περιεχομένου της ιστοσελίδας. Οι εξωτερικοί παράγοντες αναφέρονται στην πηγή προέλευσης της φράσης κλειδί σε μία ιστοσελίδα. Μία φράση κλειδί μπορεί να εξαχθεί από το περιεχόμενο της ιστοσελίδας (τίτλος και σώμα κεμένου). Επίσης, ενδέχεται η πηγή προέλευσής της να είναι εκτός του περιεχομένου της ιστοσελίδας (δεν συνίσταται).

Η θέση και η συχνότητα εμφάνισης μίας φράσης κλειδί στο τίτλο και στο σώμα κειμένου της ιστοσελίδας σε συνδυασμό με τη θέση και τη συχνότητα εμφάνισής της στα μεταδεδομένα περιεχομένου καθορίζουν τον τρόπο δημιουργίας των ιστοσελίδων δοκιμής, οι οποίες θα χρησιμοποιηθούν για την εξέταση της ορθότητας των πέντε υποθέσεων. Αναλυτικότερα, δίνεται ο κατάλογος παραμέτρων, οι οποίοι λαμβάνονται υπόψη στη δημιουργία των ιστοσελίδων δοκιμής.

- *Μεταδεδομένα*: Οι πληροφορίες τίτλου, δημιουργού (metadata creator), θέματος, περιγραφής, εκδότη (metadata publisher), ημερομηνίας (metadata date), τύπου (metadata type), μορφοποίησης (metadata format), γλώσσας (metadata language), πνευματικών δικαιωμάτων (metadata rights) ορίζουν το βασικό σύνολο στοιχεία μεταδεδομένων. Εξετάζεται η συμβολή μόνο των μεταδεδομένων τίτλου, θέματος και περιγραφής στη διαμόρφωση της θέσης της ιστοσελίδας στον κατάλογο των αποτελεσμάτων, εφόσον το περιεχόμενό τους συνήθως σχετίζεται άμεσα με το θέμα το οποίο διαπραγματεύεται η ιστοσελίδα. Επίσης, το περιεχόμενο των συγκεκριμένων μεταδεδομένων είναι πιθανότερο να ανακτηθεί από τις μηχανές

αναζήτησης και συνεπώς οι φράσεις-κλειδιά, που περιέχονται σε αυτό, να χρησιμοποιηθούν ως όροι αναζήτησης στο ευρετήριο περιεχομένου της μηχανής αναζήτησης.

- *Συχνότητα εμφάνισης των φράσεων κλειδιών*: Είναι ο αριθμός εμφανίσεων μίας φράσης κλειδί στο σώμα κειμένου της ιστοσελίδας καθώς και στο περιεχόμενο των στοιχείων μεταδεδομένων τίτλου, θέματος και περιγραφής. Για την διεξαγωγή των πειραμάτων υποτίθεται ότι η μέγιστη συχνότητα εμφάνισης μίας φράσης κλειδί στο σώμα κειμένου της ιστοσελίδας καθώς και στα στοιχεία μεταδεδομένων τίτλου, θέματος και περιγραφής είναι 4, 5, 4, 4 και 4 αντίστοιχα.
- *Συνδυασμός των στοιχείων μεταδεδομένων*: Διάφοροι συνδυασμοί των μεταδεδομένων τίτλου, θέματος και περιγραφής λαμβάνονται υπόψη.
- *Θέση της φράσης κλειδί στην ιστοσελίδα (keyword position).*

7.2.1 Δημιουργία των ιστοσελίδων δοκιμής

Για τη δημιουργία των ιστοσελίδων δοκιμής, θεωρείται δεδομένη η ύπαρξη μίας αρχικής ιστοσελίδας, το περιεχόμενο της οποίας διαπραγματεύεται τη βελονοθεραπεία. Η μεθοδική τροποποίηση των προαναφερθέντων τεσσάρων παραμέτρων της αρχικής ιστοσελίδας οδηγεί στην παραγωγή περισσοτέρων από 40 ιστοσελίδων δοκιμής. Επίσης, το περιεχόμενο της αρχικής ιστοσελίδας τροποποιήθηκε ελαφρώς κατά την παραγωγή ορισμένων ιστοσελίδων, π.χ., η κατάργηση και η προσθήκη φράσεων κλειδιών στο τίτλο και στο σώμα κειμένου της ιστοσελίδας.

Οι παραγόμενες ιστοσελίδες αναρτήθηκαν κάτω από ένα δημόσιο όνομα περιοχής ώστε να είναι προσβάσιμες από τα προγράμματα ιχνηλάτησης των μηχανών αναζήτησης και από τους χρήστες του διαδικτύου. Οι διευθύνσεις διαδικτύου των παραγόμενων ιστοσελίδων γνωστοποιήθηκαν σε 19 γνωστές μηχανές αναζήτησης με τη συμπλήρωση κατάλληλων ηλεκτρονικών αιτήσεων.

Μετά από την θεματολογική ανάλυση του περιεχομένου των παραγόμενων ιστοσελίδων η φράση κλειδί «acupuncture» προσδιορίστηκε ως ο όρος βάσει του οποίου θα ανακτηθούν οι ιστοσελίδες από το ευρετήριο περιεχομένου. Οπότε, το ερώτημα του χρήστη προς τη μηχανή αναζήτησης αποτελείται από τον μοναδικό όρο «acupuncture». Δεδομένης της ερώτησης, στην περίπτωση παρουσίας μίας ιστοσελίδας στον κατάλογο

των αποτελεσμάτων μίας μηχανής αναζήτησης καταγράφεται το όνομα αρχείου της ιστοσελίδας, η κατατακτήρια θέση της καθώς και το όνομα της μηχανής αναζήτησης.

Στους πίνακες Πίνακας 6 και Πίνακας 7 κάθε εγγραφή αντιστοιχεί σε μία παραγόμενη ιστοσελίδα ενώ κάθε στήλη εκτός της τελευταίας αντιστοιχεί σε μία παράμετρο. Η τιμή στη τελευταία στήλη δηλώνει αν η παραγόμενη ιστοσελίδα εμφανίστηκε στον κατάλογο των αποτελεσμάτων τουλάχιστον μίας μηχανής αναζήτησης. Κάθε ιστοσελίδα φέρει ένα μοναδικό όνομα, π.χ., το όνομα της 20^{ης} ιστοσελίδας είναι W_P_20. Στον Πίνακας 6 παρουσιάζεται η κατανομή (πολλαπλότητα εμφάνισης) της κύριας φράσης-κλειδί «acupuncture» στις παραγόμενες ιστοσελίδες δοκιμής. Οι αριθμοί στις δευτερεύουσες στήλες αντιστοιχούν στις συχνότητες εμφάνισης της λέξης «acupuncture» στις ιστοσελίδες, π.χ., η λέξη «acupuncture» εμφανίζεται από μία φορά ως πέντε φορές στο σώμα κειμένου μίας ιστοσελίδας. Τέλος, στον Πίνακας 7 δίνεται η πληροφορία σχετικώς με τα σημεία στον κώδικα HTML της ιστοσελίδας στα οποία εμφανίζεται η συγκεκριμένη λέξη.

Ιστοσελίδα	Τίτλος				Σώμα Κειμένου					Στοιχείο Μεταδεδομένου Τίτλου				Στοιχείο Μεταδεδομένου Θέματος				Στοιχείο Μεταδεδομένου Περιγραφής				Εμφάνιση
	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2	3	4	
W_P_1	x																					●
W_P_2		x																				●
W_P_3			x																			●
W_P_4				x																		●
W_P_5					x																	●
W_P_6						x																●
W_P_7							x															●
W_P_8								x														●
W_P_9									x													●
W_P_10										x												

Ιστοσελίδα	Τίτλος				Σώμα Κειμένου					Στοιχείο Μεταδεδομένου Τίτλου				Στοιχείο Μεταδεδομένου Θέματος				Στοιχείο Μεταδεδομένου Περιγραφής				Εμφάνιση
	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2	3	4	
W_P_11											X											
W_P_12												X										
W_P_13													X									
W_P_14														X								
W_P_15															X							
W_P_16																X						
W_P_17																	X					
W_P_18																		X				
W_P_19																			X			
W_P_20																				X		●
W_P_21																					X	

Πίνακας 6

Ιστοσελίδα	Τίτλος	Σώμα Κειμένου	Στοιχείο Μεταδεδομένου Τίτλου	Στοιχείο Μεταδεδομένου Θέματος	Στοιχείο Μεταδεδομένου Περιγραφής	Εμφάνιση
W_P_22	X	X				●
W_P_23	X		X			●
W_P_24	X			X		●
W_P_25	X				X	
W_P_26		X	X			●
W_P_27		X		X		●
W_P_28		X			X	●
W_P_29			X	X		

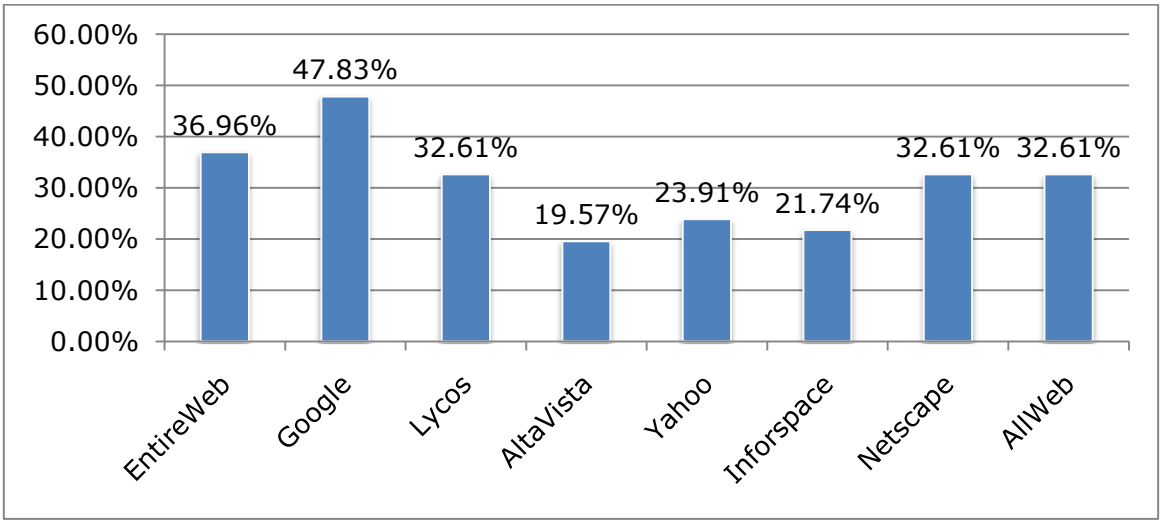
Ιστοσελίδα	Τίτλος	Σώμα Κειμένου	Στοιχείο Μεταδεδομένου Τίτλου	Στοιχείο Μεταδεδομένου Θέματος	Στοιχείο Μεταδεδομένου Περιγραφής	Εμφάνιση
W_P_30			X		X	
W_P_31				X	X	
W_P_32	X	X	X			
W_P_33	X	X		X		
W_P_34	X	X			X	
W_P_35	X		X	X		
W_P_36	X		X		X	
W_P_37	X			X	X	
W_P_38		X	X	X		
W_P_39		X	X		X	
W_P_40		X		X	X	
W_P_41			X	X	X	
W_P_42	X	X	X	X		
W_P_43	X	X	X		X	
W_P_44	X		X	X	X	
W_P_45		X	X	X	X	
W_P_46	X	X	X	X	X	

Πίνακας 7

7.2.2 Ανάλυση των πειραματικών αποτελεσμάτων

Οι διευθύνσεις των παραγόμενων ιστοσελίδων δοκιμής των πινάκων Πίνακας 6 και Πίνακας 7 υποβλήθηκαν σε 19 γνωστές μηχανές αναζήτησης. Τελικώς, οι ιστοσελίδες δοκιμής παρουσιάστηκαν στους καταλόγους αποτελεσμάτων οκτώ (8) εξ' αυτών. Οι λόγοι για την αποκλεισμό ιστοσελίδων από τα αποτελέσματα ορισμένων μηχανών αναζήτησης ποικίλλουν. Συγκεκριμένα, ορισμένες μηχανές αναζήτησης συγχωνεύθηκαν

με άλλες, η λειτουργία των εξυπηρετητών μερικών μηχανών διεκόπει, ενώ υπήρξαν και περιπτώσεις κατα τις οποίες οι μηχανές αναζήτησης δεν συμπεριέλαβαν ποτέ στο ευρετήριό τους τις ιστοσελίδες δοκιμής. Από το σύνολο των 46 ιστοσελίδων, μόνο οι διευθύνσεις 28 ιστοσελίδων εμφανίστηκαν τελικώς στον κατάλογο αποτελεσμάτων τουλάχιστον μίας μηχανής αναζήτησης. Στο ιστόγραμμα τους Σχήμα 24 δίνεται το ποσοστό εμφάνισης (appearance rate) των ιστοσελίδων δοκιμής ανά μηχανή αναζήτησης. Δεδομένης μίας μηχανής αναζήτησης, ως ποσοστό εμφάνισης ορίζεται ο λόγος του συνολικού αριθμού ιστοσελίδων δοκιμής οι οποίες εμφανίζονται στον κατάλογο αποτελεσμάτων της προς τον συνολικό αριθμό των ιστοσελίδων δοκιμής.



Σχήμα 24 Ποσοστό εμφάνισης των 8 μηχανών αναζήτησης

Το ποσοστό εμφάνισης είναι ποσοτικός δείκτης της ικανότητας μίας μηχανής αναζήτησης να ανακτά ιστοσελίδες και να τις αποθηκεύει στο ευρετήριο περιεχομένου της. Οι μηχανές αναζήτησης των εταιρειών υψηλής τεχνολογίας Google και Altavista σημειώσαν το υψηλότερο και το χαμηλότερο ποσοστό εμφάνισης των ιστοσελίδων δοκιμής, αντίστοιχα. Στον ακόλουθο πίνακα παρουσιάζεται η κατανομή των εμφανίσεων των ιστοσελίδων στις 8 μηχανές αναζήτησης.

Ιστοσελίδα	Μηχανές Αναζήτησης							
	AllWeb	EntireWeb	Google	Lycos	AltaVista	Yahoo	Infor-Space	Netscape
W_P_1	X	X	X	X		X	X	X
W_P_2	X		X		X	X		X

W_P_3	X	X	X	X	X	X	X	X
W_P_4	X	X	X	X	X		X	X
W_P_5	X	X	X	X			X	X
W_P_6	X	X	X	X		X	X	X
W_P_7	X	X	X	X		X	X	X
W_P_8	X	X	X	X	X		X	X
W_P_9	X	X	X	X	X	X	X	X
W_P_10								
W_P_11								
W_P_12								
W_P_13								
W_P_14								
W_P_15								
W_P_16								
W_P_17								
W_P_18								
W_P_19								
W_P_20			X					
W_P_21								
W_P_22	X	X	X	X	X	X	X	X
W_P_23			X	X				X
W_P_24			X					X
W_P_25								
W_P_26	X	X		X				
W_P_27			X					
W_P_28			X					
W_P_29								

W_P_30														
W_P_31														
W_P_32								X						
W_P_33	X		X			X				X		X		
W_P_34								X						
W_P_35			X											
W_P_36					X									
W_P_37			X		X					X				X
W_P_38	X		X			X								
W_P_39														
W_P_40					X									
W_P_41														
W_P_42			X		X									
W_P_43					X									X
W_P_44	X		X		X		X			X				
W_P_45														
W_P_46	X		X		X		X		X		X			X

Πίνακας 8 Κατανομή εμφανίσεων των ιστοσελίδων δοκιμής στις μηχανές αναζήτησης

Οι στοσελίδες W_P_3, W_P_9 και W_P_21 εμφανίζονται στους καταλόγους αποτελεσμάτων και των οκτώ μηχανών αναζήτησης.

Αναλύοντας τα δεδομένα του Πίνακας 6, οι ιστοσελίδες W_P_10, W_P_11, W_P_12, W_P_13, W_P_14, W_P_15, W_P_16, W_P_17, W_P_18, W_P_19 και W_P_21 δεν συμπεριλήφθηκαν στους καταλόγους αποτελεσμάτων καμίας μηχανής αναζήτησης. Η ιστοσελίδα δοκιμής W_P_20, στο κώδικα της οποίας περιλαμβάνεται το μεταδεδομένο περιγραφής, εμφανίζεται στον κατάλογο αποτελεσμάτων της μηχανής Google. Το κοινό χαρακτηριστικό των μη εμφανισθέντων ιστοσελίδων δοκιμής στους καταλόγους των αποτελεσμάτων είναι ότι οι φράσεις-κλειδία, οι οποίες εμφανίζονται στα στοιχεία μεταδεδομένων τίτλου, θέματος και περιγραφής, δεν προέρχονται από τον τίτλο ή το

σώμα κειμένου της ιστοσελίδας. Συνεπώς, αν μία λέξη ή φράση, η οποία δεν εμφανίζεται στον τίτλο ή στο σώμα κειμένου της ιστοσελίδας, χρησιμοποιηθεί ως φράση κλειδί στα στοιχεία μεταδεδομένων, η πιθανότητα παρουσίας της ιστοσελίδας στον κατάλογο αποτελεσμάτων είναι μικρή. Η εμφάνιση της ιστοσελίδας W_P_20 στα αποτελέσματα μίας μόνο μηχανής αποτελεί εξαίρεση στον προηγούμενο εμπειρικό κανόνα.

Η εξαγωγή ασφαλούς εμπειρικού κανόνα από την ανάλυση των δεδομένων του Πίνακας 7 είναι δυσκολότερη. Οι ιστοσελίδες δοκιμής W_P_25, W_P_29, W_P_30, W_P_31, W_P_39, W_P_41 και W_P_45 δεν συμπεριλήφθηκαν στα αποτελέσματα καμμίας μηχανής αναζήτησης. Η απουσία των περισσοτέρων ιστοσελίδων από τους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης εξηγείται από τον προηγούμενο εμπειρικό κανόνα. Ωστόσο, οι ιστοσελίδες W_P_25, W_P_31 και W_P_45 αποτελούν εξάιρεση. Οι φράσεις κλειδιά, οι οποίες χρησιμοποιούνται στη σύνταξη των μεταδεδομένων τίτλου, περιγραφής και θέματος, εμφανίζονται στο σώμα κειμένου της ιστοσελίδας. Συνεπώς, θα πρέπει να ήταν αναμενόμενη η εμφάνισή τους στον κατάλογο αποτελεσμάτων τουλάχιστον μίας μηχανής αναζήτησης.

7.2.2.1 Εξέταση της ορθότητας των υποθέσεων

Για την εξέταση της ορθότητας των πέντε υποθέσεων της παραγράφου 7.2 χρησιμοποίηθηκαν οι τρείς στατιστικές μέθοδοι (α) ANOVA απλής κατεύθυνσης (ANOVA one-way), (β) ANOVA διπλής κατεύθυνσης (ANOVA two-way) και (γ) T-δοκιμή ανεξαρτήτου-δείγματος (independent-sample T-test). Υποτίθεται ότι η εξαρτώμενη μεταβλητή στις τρεις μεθόδους ακολουθεί την κανονική κατανομή ενώ οι τιμές της είναι ανεξάρτητες μεταξύ τους. Ως μέγεθος μέτρησης ορίζεται η θέση της ιστοσελίδας δοκιμής στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης. Η επιλογή του συγκεκριμένου μεγέθους κρίνεται κατάλληλη εφόσον αυτό είναι μετρήσιμο και διαχειρίσιμο. Υψηλή θέση της ιστοσελίδας δοκιμής αντιστοιχεί σε χαμηλή τιμή της μεταβλητής θέσης και αντιστρόφως.

7.2.2.1.1 Πρώτη Υπόθεση

Σύμφωνα με την πρώτη υπόθεση η συμπερίληψη των μεταδεδομένων σε μία ιστοσελίδα δεν βελτιώνει τη θέση της στους καταλόγους αποτελεσμάτων. Οι ιστοσελίδες

δοκιμής ταξινομήθηκαν σε δύο κατηγορίες. Στην πρώτη κατηγορία ιστοσελίδων δοκιμής υπάρχουν στοιχεία μεταδεδομένων τίτλου, θέματος και περιγραφής. Στη δεύτερη κατηγορία απουσιάζουν τα προαναφερθέντα στοιχεία μεταδεδομένων από τον κώδικα της ιστοσελίδας ενώ οι φράσεις κλειδιά εμφανίζονται με διαφορετική συχνότητα στον τίτλο και στο σώμα κειμένου της ιστοσελίδας δοκιμής. Οι ιστοσελίδες των δύο κατηγοριών αποτελούν τις ανεξάρτητες μεταβλητές ενώ η μεταβλητή θέσης της ιστοσελίδας αποτελεί την εξαρτημένη μεταβλητή. Ο μέσος όρος της εξαρτημένης μεταβλητής για τις δύο κατηγορίες ιστοσελίδων δίνεται στον ακόλουθο πίνακα.

Κατηγορία Ιστοσελίδων	Μέσος όρος της μεταβλητής θέσης
Με στοιχεία μεταδεδομένων	6.1037
Χωρίς στοιχεία μεταδεδομένων	6.9288

Πίνακας 9 Αποτελέσματα στατιστικής ανάλυσης

Οι ιστοσελίδες με στοιχεία μεταδεδομένων (χαμηλή μέση τιμή της μεταβλητής θέσης) καταλαμβάνουν καλύτερη θέση στον κατάλογο των αποτελεσμάτων από τις ιστοσελίδες στις οποίες απουσιάζουν τα στοιχεία μεταδεδομένων (υψηλή μέση τιμή της μεταβλητής θέσης). Συνεπώς, ο ισχυρισμός της πρώτης υπόθεσης καταρρίπτεται.

7.2.2.1.2 Δεύτερη Υπόθεση

Σύμφωνα με τη δεύτερη υπόθεση δεν υπάρχει καμία διαφορά στη κατατακτήρια θέση μεταξύ των ιστοσελίδων με διαφορετικούς αριθμούς συνδυασμών στοιχείων μεταδεδομένων. Δηλαδή, εξετάζεται αν ο αριθμός των συνδυασμών των στοιχείων μεταδεδομένων σε μία ιστοσελίδα επηρεάζει το ποσοστό επισκεψιμότητάς της. Οι ανεξάρτητες μεταβλητές είναι οι ιστοσελίδες με διαφορετικό αριθμό των στοιχείων μεταδεδομένων. Η μεταβλητή τύπου TYPE στον ακόλουθο πίνακα αντιπροσωπεύει τους συνδυασμούς και με βάση την τιμή της οι ιστοσελίδες δοκιμής ταξινομούνται στις εξής τρεις κατηγορίες.

1. Η τιμή 1 (TYPE=1) ορίζει την κατηγορία στην οποία σε κάθε ιστοσελίδα υπάρχουν ακριβώς ένα εκ των στοιχείων μεταδεδομένων τίτλου, θέματος και περιγραφής.

2. Στην τιμή 2 (TYPE=2) αντιστοιχούν ιστοσελίδες στον κώδικα των οποίων εμφανίζεται ένας εκ των παρακάτω συνδυασμών
   - Στοιχεία μεταδεδομένων τίτλου και θέματος.
   - Στοιχεία μεταδεδομένων τίτλου και περιγραφής.
   - Στοιχεία μεταδεδομένων θέματος και περιγραφής.
3. Η τιμή 3 (TYPE=3) ορίζει την κατηγορία ιστοσελίδων στην οποία σε κάθε ιστοσελίδα συνυπάρχουν τα στοιχεία μεταδεδομένων τίτλου, θέματος και περιγραφής.

Η μεταβλητή θέσης της ιστοσελίδας αποτελεί την εξαρτημένη μεταβλητή. Λόγω των πολλαπλάσιων ανεξαρτήτων μεταβλητών, η τεχνική ANOVA απλής κατεύθυνσης χρησιμοποιήθηκε για την εξακρίβωση της συγκεκριμένης υπόθεσης. Τα αποτελεσμάτα της στατιστικής ανάλυση παρουσιάζονται στον ακόλουθο πίνακα

Κατηγορία Ιστοσελίδων (TYPE)	Μέση τιμή της μεταβλητής θέσης	Τυπική απόκλιση
1	6.5133	5.31022
2	7.9756	6.42062
3	4.4897	4.94695

Πίνακας 10 Αποτελέσματα στατιστικής ανάλυσης

Η μεγάλη τιμή της τυπικής απόκλισης στη δεύτερη κατηγορία δηλώνει ότι ο βαθμός συμβολής των τριών συνδυασμών μεταδεδομένων (τίτλου και θέματος, τίτλου και περιγραφής, θέματος και περιγραφής) στην διαμόρφωση της θέσης στον κατάλογο των αποτελεσμάτων δεν είναι σταθερός.

Η διαφορά των μέσων τιμών των μεταβλητών θέσεων μεταξύ των τριών κατηγοριών ιστοσελίδων παρουσιάζεται στον ακόλουθο πίνακα

Κατηγορία Ιστοσελίδων (TYPE I)	Κατηγορία Ιστοσελίδων (TYPE J)	Διαφορά Μέσης Τιμής (I-J)
1	2	-1.4623
	3	1.8236
2	1	1.4623
	3	3.2860
3	1	-1.8236
	2	-3.2860

Πίνακας 11 Συγκριτικός πίνακας μέσων τιμών

Η διαφορά της μέσης τιμής της τρίτης κατηγορίας από την αντίστοιχη τιμή της πρώτης και δεύτερης κατηγορίας είναι -3.2860 και -1.8236 αντίστοιχα. Το αρνητικό πρόσημο των δύο συγκεκριμένων διαφορών υποδηλώνει ότι η θέση των ιστοσελίδων, στις οποίες οι φράσεις κλειδιά εμφανίζονται στα τρία στοιχεία μεταδεδομένων τίτλου, περιγραφής και θέματος, είναι υψηλότερη της θέσης των ιστοσελίδων των υπολοίπων κατηγοριών. Τέλος, σε αντίθεση με το διαισθητικώς αναμενόμενο, οι ιστοσελίδες δοκιμής της πρώτης κατηγορίας καταλαμβάνουν καλύτερη θέση από τις ιστοσελίδες της δεύτερης κατηγορίας.

7.2.2.1.3 Τρίτη Υπόθεση

Η τρίτη υπόθεση εξετάζει τη διακριτή συμβολή των στοιχείων μεταδεδομένων τίτλου, περιγραφής και θέματος στην κατάταξη των αποτελεσμάτων. Οι ανεξάρτητες μεταβλητές είναι οι ιστοσελίδες με μεταδεδομένα τίτλου, οι ιστοσελίδες με μεταδεδομένα περιγραφής και οι ιστοσελίδες με μεταδεδομένα θέματος, ενώ η εξαρτημένη μεταβλητή είναι η μεταβλητή θέσης της ιστοσελίδας στον κατάλογο των αποτελεσμάτων. Για την διαπίστωση της ισχύος της υπόθεσης χρησιμοποιήθηκε ο στατιστικός έλεγχος ANOVA απλής κατεύθυνσης.

Σε αναλογία με τη διαδικασία εξέτασης της δεύτερης υπόθεσης, οι ιστοσελίδες δοκιμής ταξινομούνται σε τρείς κατηγορίες. Σε κάθε κατηγορία αποδίδεται ο μοναδικός αριθμός τύπου (TYPE).

1. Ιστοσελίδες με μεταδεδομένα τίτλου (TYPE=1).
2. Ιστοσελίδες με μεταδεδομένα θέματος (TYPE=2).
3. Ιστοσελίδες με μεταδεδομένα περιγραφής (TYPE=3).

Τα αποτελεσμάτα του στατιστικής ανάλυσης καθώς και οι διαφορές των μέσων τιμών των μεταβλητών θέσεων των τριών προηγούμενων κατηγοριών ιστοσελίδων παρουσιάζονται στους πίνακες Πίνακας 12 και Πίνακας 13.

Κατηγορία Ιστοσελίδων (TYPE)	Μέση τιμή της μεταβλητής θέσης	Τυπική απόκλιση
1	8.1316	5.23590
2	4.6094	4.31541
3	12.0000	5.31022

Πίνακας 12 Αποτελέσματα στατιστικής ανάλυσης

Κατηγορία Ιστοσελίδων (TYPE I)	Κατηγορία Ιστοσελίδων (TYPE J)	Διαφορά Μέσης Τιμής (I-J)
1	2	3.5222
	3	-3.8684
2	1	-3.5222
	3	-7.3906
3	1	3.8684
	2	7.3906

Πίνακας 13 Συγκριτικός πίνακας μέσων τιμών

Η διαφορά της μέσης τιμής της δευτερης κατηγορίας από την αντίστοιχη τιμή της πρώτης και τρίτης κατηγορίας είναι -3.5222 και -7.3906 αντίστοιχα. Το αρνητικό πρόσημο των δύο συγκεκριμένων διαφορών υποδηλώνει ότι η συμβολή των μεταδεδομένων θέματος στην κατατακτήρια θέση των ιστοσελίδων έχει μεγαλύτερο σχετικό βάρος. Συνεπώς, ο ρόλος των μεταδεδομένων θέματος είναι ο σημαντικότερος και ο σταθερότερος (χαμηλή τυπική απόκλιση) στα πλαίσια της βελτίωσης του

ποσοστού επισκεψιμότητας μίας ιστοσελίδας. Αντιθέτως, όσον αφορά τα μεταδεδομένα περιγραφής, η συμβολή τους είναι μικρή και χαρακτηρίζεται από αστάθεια (μεγάλη τιμή της τυπικής απόκλισης).

7.2.2.1.4 Τέταρτη Υπόθεση

Σύμφωνα με τη τέταρτη υπόθεση, δεν υπάρχει διαφορά στη θέση κατάταξης ιστοσελίδων με διαφορετικό συνδυασμό στοιχείων μεταδεδομένων τίτλου, θέματος και περιγραφής. Η συγκεκριμένη υπόθεση είναι συμπληρωματική της δεύτερης.

Ο αριθμός των δυνατών συνδυασμών των τριών στοιχείων μεταδεδομένων είναι επτά[22] ενώ οι ιστοσελίδες δοκιμής ομαδοποιούνται σε ισάριθμες κατηγορίες. Σε κάθε κατηγορία δίνεται μοναδικός αριθμός τύπου (TYPE). Αναλυτικότερα

Κατηγορία Ιστοσελίδας (TYPE)	Μεταδεδομένα		
	Τίτλος	Θέμα	Περιγραφή
1	X		
2		X	
3			X
4	X	X	
5	X		X
6		X	X
7	X	X	X

Πίνακας 14 Κατηγορίες Ιστοσελίδων

Οι ανεξάρτητες μεταβλητές είναι οι ιστοσελίδες με διαφορετικό συνδυασμό των τριών στοιχείων μεταδεδομένων ενώ η εξαρτημένη μεταβλητή είναι η θέση της ιστοσελίδας στον κατάλογο των αποτελεσμάτων. Τα αποτελεσμάτα του στατιστικού ελέγχου καθώς και οι διαφορές των μέσων τιμών των μεταβλητών θέσεων των τριών προηγούμενων κατηγοριών ιστοσελίδων παρουσιάζονται στους ακόλουθους δύο πίνακες.

[22] $\sum_{i=1}^{3}\binom{3}{i} = 7$.

Κατηγορία Ιστοσελίδων (TYPE)	Μέση τιμή της μεταβλητής θέσης	Τυπική απόκλιση
1	8.1316	5.23590
2	4.6094	4.31541
3	12.0000	5.31022
4	10.0000	6.78970
5	3.8571	1.95180
6	6.6087	5.80604
7	3.5926	4.12776

Πίνακας 15 Αποτελέσματα στατιστικής ανάλυσης

Κατηγορία Ιστοσελίδων (TYPE I)	Κατηγορία Ιστοσελίδων (TYPE J)	Διαφορά Μέσης Τιμής (I-J)
1	2	3.5222
	3	-3.8684
	4	-1.8684
	5	4.2744
	6	1.5229
	7	4.5390
2	1	-3.5222
	3	-7.3906
	4	-5.3906
	5	0.7522
	6	-1.9993
	7	1.0168
3	1	3.8684

Κατηγορία Ιστοσελίδων (TYPE I)	Κατηγορία Ιστοσελίδων (TYPE J)	Διαφορά Μέσης Τιμής (I-J)
	2	7.3906
	4	2.0000
	5	8.1429
	6	5.3913
	7	8.4074
4	1	1.8684
	2	5.3906
	3	-2.0000
	5	6.1429
	6	3.3913
	7	6.4074
5	1	-4.2744
	2	-0.7522
	3	-8.1429
	4	-6.1429
	6	-2.7516
	7	0.2646
6	1	-1.5229
	2	1.9993
	3	-5.3913
	4	-3.3913
	5	2.7516

Κατηγορία Ιστοσελίδων (TYPE I)	Κατηγορία Ιστοσελίδων (TYPE J)	Διαφορά Μέσης Τιμής (I-J)
	7	3.0161
7	1	-4.5390
	2	-1.0168
	3	-8.4074
	4	-6.4074
	5	-0.2646
	6	-3.0161

Πίνακας 16 Συγκριτικός πίνακας μέσων τιμών

Από την ανάλυση των δεδομένων του πίνακα προκύπτει ότι οι ιστοσελίδες, στις οποίες φράσεις κλειδιά προερχόμενες από το σώμα κειμένου εμφανίζονται στα μεταδεδομένα τίτλου, θέματος και περιγραφής (πρώτη κατηγορία), καταλαμβάνουν υψηλότερες θέσεις στους καταλόγους των αποτελεσμάτων έναντι των ιστοσελίδων των υπολοίπων κατηγοριών. Ακολουθούν οι ιστοσελίδες με φράσεις κλειδιά στα μεταδεδομένα θέματος και κατόπιν οι ιστοσελίδες με φράσεις κλειδιά στα μεταδεδομένα τίτλου και περιγραφής.

7.2.2.1.5 Πέμπτη Υπόθεση

Στην παρούσα παράγραφο εξετάζεται ο ρόλος της πηγής προέλευσης των φράσεων-κλειδιών, οι οποίες εμφανίζονται στα μεταδεδομένα μίας ιστοσελίδας, στη θέσης της τελευταίας στους καταλόγους των αποτελεσμάτων. Οι φράσεις κλειδιά, οι οποίες χρησιμοποιούνται στη σύνταξη των μεταδεδομένων, εμφανίζονται (α) στον τίτλο της ιστοσελίδας, (β) στο σώμα κειμένου της ιστοσελίδας και (γ) στον τίτλο και στο σώμα κείμενο της ιστοσελίδας. Ανάλογα με θέση των φράσεων κλειδιών στο περιεχόμενο της ιστοσελίδας δοκιμής, ορίζεται η ακόλουθη κατηγοριοποίηση των ιστοσελίδων (σε κάθε κατηγορία δίνεται μοναδικός αριθμός τύπου (TYPE)).

1. Ιστοσελίδες με πληροφορίες μεταδεδομένων στη σύνταξη των οποίων χρησιμοποιούνται φράσεις κλειδιά οι οποίες εμφανίζονται στο τίτλο της ιστοσελίδας (TYPE=1).
2. Ιστοσελίδες με πληροφορίες μεταδεδομένων στη σύνταξη των οποίων χρησιμοποιούνται φράσεις κλειδιά οι οποίες εμφανίζονται στο σώμα κειμένου της ιστοσελίδας (TYPE=2).
3. Ιστοσελίδες με πληροφορίες μεταδεδομένων στη σύνταξη των οποίων χρησιμοποιούνται φράσεις κλειδιά οι οποίες εμφανίζονται στο τίτλο και στο σώμα κειμένου της ιστοσελίδας (TYPE=3).

Οι ανεξάρτητες μεταβλητές είναι οι φράσεις κλειδιά. Εξαιρούνται φράσεις-κλειδιά, οι οποίες εμφανίζονται μόνο στα μεταδεδομένα περιεχομένου. Η εξαρτημένη μεταβλητή είναι η θέση της ιστοσελίδας στον κατάλογο των αποτελεσμάτων της μηχανής αναζήτησης. Για τη εξέταση της συγκεκριμένης υπόθεσης χρησιμοποιήθηκε η τεχνική ANOVA απλής κατεύθυνσης. Τα αποτελεσμάτα της στατιστικής ανάλυσης καθώς και οι διαφορές των μέσων τιμών των μεταβλητών θέσεων των τριών κατηγοριών ιστοσελίδων παρουσιάζονται στους παρακάτω πίνακες.

Κατηγορία Ιστοσελίδων (TYPE)	Μέση τιμή της μεταβλητής θέσης	Τυπική απόκλιση
1	3.3223	3.93110
2	12.8889	4.68505
3	3.5403	3.42017

Πίνακας 17 Αποτελέσματα στατιστικής ανάλυσης

Κατηγορία Ιστοσελίδων (TYPE I)	Κατηγορία Ιστοσελίδων (TYPE J)	Διαφορά Μέσης Τιμής (I-J)
1	2	-9.5666
	3	-0.2180
2	1	9.5666
	3	9.3486
3	1	0.2180
	2	-9.3486

Πίνακας 18 Συγκριτικός πίνακας μέσων τιμών

Η διαφορά των μέσων τιμών των μεταβλητών θέσεων των ιστοσελίδων της πρώτης και δεύτερης κατηγορίας είναι ισχυρά αρνητική (-9.5666). Η ίδια διαπίστωση ισχύει για τη διαφορά των μέσων τιμών της τρίτης και της δεύτερης κατηγορίας ιστοσελίδων (-9.3486). Η διαφορά των μέσων τιμών για την πρώτη και τη τρίτη κατηγορία είναι -0.2180. Συνεπώς, οι φράσεις κλειδιά, οι οποίες χρησιμοποιούνται στη σύνταξη των μεταδεδομένων και εμφανίζονται είτε στο τίτλο της ιστοσελίδας είτε στο τίτλο και το σώμα κειμένου της ιστοσελίδας, συμβάλλουν σε μεγαλύτερο βαθμό στην κατάληψη καλύτερης θέσης από τις φράσεις κλειδιά, οι οποίες εμφανίζονται αποκλειστικά στο σώμα κειμένου της ιστοσελίδας. Για τη βελτίωση του ποσοστού επισκεψιμότητας των ιστοσελίδων επιβάλλεται η χρήση φράσεων κλειδιών στη σύνταξη των μεταδεδομένων, οι οποίες προέρχονται είτε από τον τίτλο είτε από τον τίτλο και το σώμα κειμένου της ιστοσελίδας.

7.3 Συμπεράσματα

Στην παράγραφο 7.2.2 προτείνονται τεχνικές για την βελτιστοποίηση των θέσεων των ιστοσελίδων στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης στα πλαίσια της συμπερίληψης των μεταδεδομένων στον κώδικα HTML. Για την εξαγωγή ασφαλών συμπερασμάτων 46 ιστοσελίδες δοκιμής παρήχθησαν τροποποιώντας μια επιλεγμένη αρχική ιστοσελίδα. Οι συγκεκριμένες ιστοσελίδες αναρτήθηκαν σε ένα δημόσιο όνομα περιοχής ενώ οι διευθύνσεις τους υποβλήθηκαν και γνωστοποιήθηκαν σε 19 εμπορικές

μηχανές αναζήτησης. Χρησιμοποιήθηκε μία απλή διατυπωμένη ερώτηση (αποτελούμενη από έναν όρο) για την αναζήτηση των ιστοσελίδων δοκιμής στις 19 μηχανές αναζήτησης. Το χρονικό διάστημα των παρατηρήσεων ήταν 21 εβδομάδες κατά το οποίο 8 μηχανές αναζήτησης ανέκτησαν και ευρετηρίασαν τις ιστοσελίδες δοκιμής. 28 ιστοσελίδες εμφανίστηκαν στον κατάλογο των αποτελεσμάτων τουλάχιστον μία φορά.

Τρείς διαφορετικές στατιστικές τεχνικές χρησιμοποιήθηκαν για την εξέταση της ορθότητας πέντε υποθέσεων σχτικών με το ρόλο των μεταδεδομένων στην κατάληψη καλύτερης θέσης στους καταλόγους αποτελεσμάτων. Τα ευρήματα της εξέτασης των υποθέσεων προτείνουν διάφορες μεθόδους για την βελτιστοποίηση του ποσοστού επισκεψιμότητας των ιστοσελίδων. Συγκεκριμένα, τα συμπεράσματα της εξέτασης συνοψίζονται στα ακόλουθα σημεία

- Δεν ενδείκνυται η χρήση φράσεων κλειδιών στη σύνταξη των μεταδεδομένων περιεχομένου, οι οποίες δεν εμφανίζονται είτε στο τίτλο είτε στο σώμα κειμένου της ιστοσελίδας. Οι φράσεις κλειδιά πρέπει να εμφανίζονται στο περιεχόμενο της ιστοσελίδα (τίτλος, σώμα κειμένου).
- Με κριτήριο το ποσοστό κάλυψης του διαδικτύου (αριθμός ανακτώμενων ιστοσελίδων από το πρόγραμμα ιχνηλάτησης) η εμπορική μηχανή αναζήτησης Google παρουσιάζει την καλύτερη επίδοση έναντι των υπολοίπων μηχανών αναζήτησης.
- Ιστοσελίδες με στοιχεία μεταδεδομένων περιεχομένου καταλαμβάνουν υψηλότερη θέση στους καταλόγους των αποτελεσμάτων έναντι των ιστοσελίδων χωρίς μεταδεδομένα.
- Ιστοσελίδες στις οποίες οι φράσεις κλειδιά συνυπάρχουν στα στοιχεία μεταδεδομένων τίτλου, περιεχομένου και θέματος καταλαμβάνουν καλύτερη θέση έναντι των ιστοσελίδων με άλλους πιθανούς συνδυασμούς των στοιχείων μεταδεδομένων περιεχομένου.
- Η συμπερίληψη του μεταδεδομένου θέματος στον κώδικα της ιστοσελίδας συμβάλλει στη κατάληψη καλύτερης θέσης στους καταλόγους αποτελεσμάτων των μηχανών αναζήτησης.

- Ιστοσελίδες στις οποίες οι φράσεις κλειδιά στα μεταδεδομένα εμφανίζονται είτε στο τίτλο είτε στο τίτλο και στο σώμα κειμένου καταλαμβάνουν υψηλότερη θέση στους καταλόγους αποτελεσμάτων έναντι των ιστοσελίδων με φράσεις-κλειδιά οι οποίες εμφανίζονται αποκλειστικώς στο σώμα κειμένου.

Συνοψίζοντας, η συμπερίληψη των μεταδεδομένων περιεχομένου στον κώδικα μίας ιστοσελίδας συνιστά έναν αποδοτικό μηχανισμό για την βελτιώση του ποσοστού επισκεψιμότητάς της.

Ορολογία

adaptive power method	προσαρμοσμένη μέθοδο της δύναμης
adjacency matrix	πίνακα γειτνίασης
aggregation method	μέθοδο της συνάθροισης
alt tag	εναλλακτικό κείμενο γραφικών
anchor text	κείμενο αγκύρωσης σε υπερσυνδέσμους
ANOVA one-way	ANOVA απλής κατεύθυνσης
ANOVA two-way	ANOVA διπλής κατεύθυνσης
appearance rate	ποσοστό εμφάνισης
application server	εξυπηρετητής εφαρμογών
application web	εφαρμογή διαδικτύου
authority matrix	πίνακας αυθεντίας
authority page	ιστοσελίδα με αυθεντική πληροφορία (Αλγόριθμος HITS)
authority score	βαθμός αυθεντίας
average relevance	μέθοδος της ενδιάμεσης σχετικότητας
canonicalization	κανονικοποίηση
change curves	καμπύλη μεταβολής
checksum	έλεγχος αθροίσματος
collection term frequency	συχνότητα εμφάνισης του όρου στη συλλογή των ιστοσελίδων
competitiveness of a term	ανταγωνιστικότητας ενός όρου
connected component	συνδεδεμένη συνιστώσα
content change	μεταβολή του περιεχομένου
content index	ευρετήριο περιεχομένου
content score	βαθμός περιεχομένου
convergence criterion	κριτήριο σύγκλισης
cost per click	κόστους ανά κλικ
crawling loop	βρόχο ιχνηλάτησης
database management systems	σύστημα διαχείρισης βάσης δεδομένων

divergence measure	μέτρο της απόκλισης
Document Object Model	μοντέλο αντικειμένων εγγράφου
document term frequency	συχνότητα εμφάνισης του όρου στο κείμενο
dynamic web page	δυναμική ιστοσελίδα
edge server	εξυπηρετητής ακμής
eigenvalue	ιδιοτιμή
eigenvector	ιδιοδιάνυσμα
exclusion protocol	πρωτόκολλο αποκλεισμού
extrapolation method	μέθοδος της παρεμβολής
first depth search	αναζήτηση κατά βάθος
footer	υποσέλιδο
fragment	τμήμα ιστοσελίδας
frontier list	λίστα διευθύνσεων ιστοσελίδων προς ανάκτηση στο
Google matrix	πίνακας Google
harvest rate relevance score	μέθοδος του ποσοστού συγκομιδής
Header	Επικεφαλίδας
header tag	ετικέτα τίτλου κειμένου
hub matrix	πίνακας αναφοράς
hub page	ιστοσελίδα πύλη (Αλγόριθμος HITS)
hub score	βαθμός αναφοράς
Hyperlink	υπερσύνδεσμος
immediate fragment	τμήμα άμεσης δημοσίευσης
independent-sample T-test	Τ-δοκιμή ανεξαρτήτου-δείγματος
index	ευρετήριο
information retrieval	σύστημα ανάκτησης πληροφοριών
informational query	ερώτηση ενημερωτικού χαρακτήρα
inlink	εισερχόμενος (υπερ)σύνδεσμος
keyword	φράση κλειδί
keyword frequency	συχνότητα επανάληψης μίας φράσης κλειδί
keyword frequency	συχνότητα επανάληψης μίας φράσης κλειδί

keyword suggestion tools	λογισμικό εύρεσης φράσεων κλειδιών
knot point	σημείο καμπής
link building	κτίσιμο συνδέσμων
link request	αίτημα σύνδεσης
metadata	μεταδεδομένα
metadata description	μεταδεδομένα περιγραφής
metadata subject	μεταδεδομένα θέματος
metadata title	μεταδεδομένα τίτλου
meta-tag	μετα-ετικέτα
Model-View-Control	πρότυπο-όψη-ελεγχος
multithread execution	πολυνηματική εκτέλεση
navigational query	ερώτηση για την εύρεση της διεύθυνσης διαδικτύου
neighborhood graph	γράφος γειτνίασης
object dependence graph	γράφος εξάρτησης των τμημάτων
organic keyword	οργανική φράση κλειδί
outlink	εξερχόμενος (υπερ)σύνδεσμος
overall score	συνολικός βαθμός ιστοσελίδας
page repository	δεξαμενή ιστοσελίδων
popularity score	βαθμός δημοτικότητας
power method	δυναμομέθοδος, μέθοδος της δύναμης
primitivity adjustment	δεύτερη τροποποίηση στον πίνακα **S**
priority queue	ουρά προτεραιότητας
quality controlled fragment	τμήμα ελεχόμενης δημοσίευσης
query processing module	επεξεργασία του ερωτήματος του χρήστη
ranking module	λογισμικό βαθμολόγησης
recall metric	μέτρο ανάκληση
recall target metric	εκτιμώμενο μέτρο ανάκλησης
row normalized hyperlink	πίνακας με κανονικοποιημένες γραμμές
search engine	μηχανή αναζήτησης
search engine marketing	μάρκετινγκ των μηχανών αναζήτησης

search engine result list	κατάλογος αποτελεσμάτων μηχανής αναζήτησης
seed urls	αρχικές διευθύνσεις στο πρόγραμμα ιχνηλάτησης
service web	υπηρεσία ιστού
spam site	ιστοσελίδα με ιογενές περιεχόμενο
sparse matrix	αραιός πίνακας
special purpose index	ευρετήρια ειδικού σκοπού
static web page	ιστοσελίδα με στατικό περιεχόμενο
stationary vector	στάσιμο διάνυσμα
staying power measure	μέτρο της δύναμης παραμονής
stochastic property	στοχαστική ιδιότητα
stochasticity adjustment	πρώτη τροποποίηση στον πίνακα **H**
stopwords	λέξεις με μικρή σημασιολογική αξία
strongly connencted	ισχυρή συνδεδεμένη συνιστώσα
structural change	μεταβολή της δομής
structure index	ευρετήριο δομής
substochastic property	υποστοχαστική ιδιότητα
tag	ετικέτα
term	όρος, λέξη, φράση
term lifespan plots	καμπύλη διάρκειας ζωής
title tag	ετικέτα τίτλου της ιστοσελίδας
trade mark	εμπορικό σήμα
traffic-building campaign	εκστρατείας δημιουργίας επισκεψιμότητας
transactional query	ερώτηση για τη διεκπεραίωση συναλλαγών στο διαδίκτυο
transition probability matrix	πίνακας μετάβασης μαρκοβιανής αλυσίδας
user’s query	ερώτημα του χρήστη
weakly connected	αδύναμη συνδεδεμένη συνιστώσα
web browser	φυλλομετρητής
web crawler, web spider	ιχνηλάτης διαδικτύου
web graph	γράφος διαδικτύου
web page	ιστοσελίδα

web page relevancy	βαθμός σχετικότητας των ανακτώμενων ιστοσελίδων με το
web page visibility	ποσοστό επισκεψιμότητας ιστοσελίδας
web server	εξυπηρετητής διαδικτύου
website	ιστότοπος
World Wide Web	παγκόσμιος ιστός

Ακρωνύμια

DOM	**D**ocument **O**bject **M**odel
FIFO	**F**irst-**I**n **F**irst-**O**ut
HITS	**H**yperlink-**I**nduced **T**opic **S**earch
HTML	**H**yper**T**ext **M**arkup **L**anguage
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
MVC	**M**odel-**V**iew-**C**ontrol
SCC	**S**trongly **C**onnected **C**omponent
XML	e**X**tensible **M**arkup **L**anguage

Βιβλιογραφικές Αναφορές

1. *The Evolution of the Web and Implications for an Incremental Crawler.* **Cho, Junghoo και Garcia-Molina, Hector.** s.l. : Morgan Kaufmann, 2000. σσ. 200-209.

2. *A large-scale study of the evolution of Web pages.* **Fetterly, Dennis, και συν.** 2004 , Softw., Pract. Exper., Τόμ. 34, σσ. 213-237. 2.

3. *Automatic detection of fragments in dynamically generated web pages.* **Ramaswamy, Lakshmish, και συν.** s.l. : ACM, 2004, WWW, σσ. 443-454.

4. *Stanford WebBase components and applications.* **Cho, Junghoo, και συν.** 2006, ACM Trans. Internet Techn., Τόμ. 6, σσ. 153-186.

5. *A taxonomy of web search.* **Broder, Andrei Z.** 2002, SIGIR Forum, Τόμ. 36, σσ. 3-10.

6. *Graph structure in the Web.* **Broder, Andrei Z., και συν.** 2000, Computer Networks, Τόμ. 33, σσ. 309-320. 1-6.

7. *Authoritative sources in a hyperlinked environment.* **Kleinberg, Jon.** 1999, Journal of the ACM.

8. *PageRank, Hits and a unified framework for link analysis.* **Ding, Chris, και συν.** Tampere, Finland : s.n., August 2002. In Proceedings of the 25th ACM SIGIR Conference. σσ. 353-354.

9. **Ding, Chris, και συν.** *Link analysis: Hubs and authorities on the World Wide Web.* Lawrence Berkeley National Laboratory. 2001. Technical Report. 47847.

10. *The impact of metadata implementationon webpage visibility in search engine results (Part II).* **Zhang, Jin και Dimitroff, Alexandra.** 2005, Inf. Process. Manage., Τόμ. 41, σσ. 691-715. 3.

11. *Web Characterization Project: An Analysis of Metadata Usage on the Web.* **O'Neill, Edward T., Lavoie, Brian F. και McClain, Patrick D.** s.l. : The Haworth Press, Inc., 2001, Journal of Library Administration, Τόμ. 34, σσ. 359-374. 3-4.

12. **Kleinberg, Jon M., και συν.** The Web as a Graph: Measurements, Models, and Methods. *COCOON.* 1999.

13. *The Web as a graph: How far we are.* **Donato, Debora, και συν.** 2007, ACM Trans. Internet Techn., Τόμ. 7.

14. *Efficient URL caching for world wide web crawling.* **Broder, Andrei Z., Najork, Marc και Wiener, Janet L.** 2003. σσ. 679-689.

15. *Crawling the Web.* **Pant, Gautam, Srinivasan, Padmini και Menczer, Filippo.** s.l. : Springer, 2004. σσ. 153-178.

16. *The Web as a Graph.* **Kumar, Ravi, και συν.** 2000. σσ. 1-10.

17. **Berkhin, Pavel.** Survey: A Survey on PageRank Computing. *Internet Mathematics.* 2005, Τόμ. 2.

18. *Crawling on web graphs.* **Cooper, Colin και Frieze, Alan M.** 2002. σσ. 419-427.

19. *Syntactic Clustering of the Web.* **Broder, Andrei Z., και συν.** 1997, Computer Networks, Τόμ. 29, σσ. 1157-1166. 8-13.

20. *A fragment-based approach for efficiently creating dynamic web content.* **Challenger, Jim και Dantzig, Paul.** 2005, ACM Trans. Internet Techn., Τόμ. 5, σσ. 359-389. 2.

21. *The web changes everything: understanding the dynamics of web content.* **Adar, Eytan, και συν.** s.l. : ACM, 2009, WSDM, σσ. 282-291.

22. *What's new on the web?: the evolution of the web from a search engine perspective.* **Ntoulas, Alexandros, Cho, Junghoo και Olston, Christopher.** s.l. : ACM, 2004, WWW, σσ. 1-12.

23. *Recrawl scheduling based on information longevity.* **Olston, Christopher και Pandey, Sandeep.** s.l. : ACM, 2008, WWW, σσ. 437-446.

24. *A large-scale study of the evolution of web pages.* **Fetterly, Dennis, και συν.** 2003. σσ. 669-678.

25. *Trawling the Web for Emerging Cyber-Communities.* **Kumar, Ravi, και συν.** 1999, Computer Networks, Τόμ. 31, σσ. 1481-1493. 11-16.

26. **Chaffey, Dave.** *Ηλεκτρονικό Επιχειρείν και Ηλεκτρονικό Εμπόριο.* s.l. : Εκδόσεις Κλειδάριθμος, 2008.

27. **Ledford, Jerri L.** *SEO Search Engine Opotimization Bible.* s.l. : Wiley Publishing, Inc., 2008.

28. **Sirovich, Jaimie και Darie, Christian.** *Professional Search Engine Optimization with PHP.* s.l. : Wiley Publishing, Inc., 2007.

29. *Scaling personalized web search.* **Jeh, Glen και Widom, Jennifer.** 2003. σσ. 271-279.

30. *Extrapolation methods for accelerating PageRank computations.* **Kamvar, Sepandαρ D., και συν.** 2003. σσ. 261-270.

31. **Langville, Amy Nicole και Dean Meyer, Carl.** Survey: Deeper Inside PageRank. *Internet Mathematics.* 2003, Τόμ. 1.

32. **Langville, Amy Nicole και Meyer, Carl.** *Updating the Stationary Vector of an Irreducible Markov Chain with an Eye on Google's PageRank.* 2004.

33. **Baeza-Yates, Ricardo και Ribeiro-Neto, Berthier.** *Modern Information Retrieval.* New York : ACM Press, 2003.

34. *Finding authorities and hubs from link structures on the World Wide Web.* **Borodi, Allan, και συν.** 2001. σσ. 415-429.

35. *Introduction to Algorithms.* **Cormen, Thomas H., και συν.** 2001. Τόμ. MIT Press.

36. *PageRank: HITS and a Unified Framework for Link Analysis.* **Ding, Chris H. Q., και συν.** San Francisco : SIAM, 2003. Proceedings of the Third SIAM International Conference on Data Mining.

37. *Modifications of Kleinberg's HITS Algorithm Using Matrix Exponentiation and WebLog Records.* **Miller, Joel C., Rae, Gregory και Schaefer, Fred.** New Orleans : SIGIR, 2001. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. σσ. 444-445.

38. Google. [Ηλεκτρονικό] www.google.com.

39. Bing. [Ηλεκτρονικό] www.bing.com.