ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ (Τ.Ε.Ι.) ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ: ΤΕΧΝΙΚΕΣ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ

WEB INFORMATION RETRIEVAL: TECHNIQUES & ALGORITHMS

ΔΟΥΜΑ ΙΩΑΝΝΑ
&
ΛΟΞΑ ΓΕΩΡΓΙΑ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: κ. ΓΙΩΤΟΠΟΥΛΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

ΠΑΤΡΑ, ΑΠΡΙΛΙΟΣ 2014

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΙΕΣ

Θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή της πτυχιακής μας εργασίας τον κ. Ηλία Σταυρόπουλο για την καθοδήγηση, την βοήθεια, τις γνώσεις που μας προσέφερε και την υπομονή που έδειξε καθ' όλη τη διάρκεια της δημιουργίας της.

Ευχαριστούμε θερμά όλους τους φίλους, συμφοιτητές και συναδέλφους μας στη δουλειά για την στήριξη και την κατανόησή τους σε αυτή μας την προσπάθεια όλα αυτά τα χρόνια.

Τέλος, το μεγαλύτερο ευχαριστώ το οφείλουμε στους γονείς μας, Γεωργία και Χριστίνα, στο σύζυγο Ανδρέα, στα αδέλφια μας Στέφανο, Αντώνη, Άγγελο και Έφη, καθώς και όλους τους στενούς συγγενείς για την συμπαράσταση, την αγάπη και την εμπιστοσύνη που μας έδειξαν όλα αυτά τα χρόνια στους οποίους αφιερώνουμε την εργασία μας, όπως επίσης και στο νέο μέλος της οικογένειας της Ιωάννας, στη μικρή μπεμπούλα μας!

ΠΕΡΙΛΗΨΗ

Καθημερινά δημιουργούνται περίπου ένα εκατομμύριο ηλεκτρονικές σελίδες (webpages) για να προστεθούν στις εκατοντάδες εκατομμύρια ήδη υπάρχουσες σελίδες του Παγκόσμιου Ιστού (World Wide Web). Ο τεράστιος αυτός όγκος πληροφορίας συνδέεται με περισσότερους από ένα δισεκατομμύριο συνδέσμους (hyperlinks). Λόγω της ραγδαίας και χαοτικής ανάπτυξης του Παγκόσμιου Ιστού, είναι φανερή η έλλειψη οργάνωσης και δομής του υπάρχοντος δικτύου πληροφορίας. Το ερώτημα που γενάτε είναι πως μπορεί κάποιος να ανακτήσει σελίδες υψηλής ποιότητας που να σχετίζονται άμεσα με τις συγκεκριμένες ανάγκες του. Για το σκοπό αυτό οι χρήστες έχουν καταφύγει στις μηχανές αναζήτησης (search engines). Ο κλασικός τρόπος λειτουργίας μιας μηχανής αναζήτησης είναι η συντήρηση ενός ευρετηρίου (index) με τις ήδη γνωστές σελίδες για κάθε δυνατό ερώτημα (query) του χρήστη, η ανάκτηση πληροφορίας μέσω αυτού του ευρετηρίου και η διαβάθμιση των σελίδων μέσω ευρετικών κανόνων (heuristics), τις περισσότερες φορές μη αποδοτικών. Προβλήματα όπως πως μια μηχανή αναζήτησης θα επιλέξει τις 20 «καλύτερες» σελίδες όταν το ερώτημα του χρήστη εμφανίζεται σε δεκάδες χιλιάδες σελίδες είναι εύλογα και δύσκολα να λυθούν αποτελεσματικά.

Την τελευταία δεκαετία έχουν αναπτυχθεί νέες τεχνικές και αλγόριθμοι που βασίζονται στη δομή του παγκοσμίου ιστού παρά στο περιεχόμενο της κάθε σελίδας. Ο παγκόσμιος ιστός μπορεί να θεωρηθεί σαν ένα τεράστιο κατευθυντικό γράφημα, με τις σελίδες να αποτελούν τους κόμβους του γραφήματος και τους συνδέσμους τις ακμές του. Οι σύνδεσμοι μεταξύ των σελίδων παρέχουν πληροφορία για το πόσο σχετίζονται δυο σελίδες μεταξύ τους και για το πόσο σημαντική είναι μια σελίδα ή όχι. Με βάση αυτές τις δύο υποθέσεις, μπορούν να αναπτυχθούν τεχνικές για την εύρεση ποιοτικών σελίδων και τον προσδιορισμό «κοινοτήτων» (σελίδες με κοινά ενδιαφέροντα) στο διαδίκτυο.

Η παρούσα πτυχιακή εργασία έχει σαν στόχο την μελέτη πρόσφατων αποδοτικών τεχνικών και αλγορίθμων για την εύρεση σελίδων υψηλής ποιότητας που σχετίζονται με το ερώτημα του χρήστη. Οι τεχνικές αυτές βασίζονται στην γραφική αναπαράσταση του Ιστού και στην ανάλυση των συνδέσμων αυτού. Η εργασία επικεντρώνεται στην περιγραφή και ανάλυση του αλγορίθμου Pagerank που αποτελεί τον βασικό τρόπο λειτουργίας της Google, της πιο δημοφιλούς μηχανής αναζήτησης.

Η εργασία ολοκληρώνεται με μια προσπάθεια να εισβάλουμε στα άδυτα της Google, αναφέροντας το μυστικό της επιτυχίας της αλλά και τα μελλοντικά σχέδια της.

ABSTRACT

In a daily basis more than one million webpages are created in addition to the existing hundred million pages of the World Wide Web. That huge amount of information is combined with more than one billion hyperlinks. Due to the rapidly and chaotic development of the World Wide Web, there is obviously a lack in the organization and structure of the existing information network. The question that rises is how someone can retrieve web pages of high quality that are related directly to his specific needs. For that reason, users have started using the search engines. The most common way of a search engine to work, is to maintain an index with the already known and existing pages for every single and possible query by the user's side, to retrieve the information through that index and to grade of the pages using heuristics – a procedure that the most of the times it is not effective enough. Problems like how a search engine will select the 20 "best" pages when user's query appears in thousands of pages are reasonably and hard to be solved effectively.

New techniques and algorithms have been developed within the last decades that are based in the structure of the World Wide Web rather than in the content of each page. The World Wide Web can be considered as a huge directed graph, where pages correspond to the nodes and hyperlinks correspond to the edge of the graph. The links between the pages provide information on how these pages are related to each other and on how important or not is a specific page. Based on these assumptions, techniques can be developed for finding qualitative pages and for the determination of the "communities" (pages with common interests) in the web.

The objective of this work is to study recent and efficient techniques and algorithms for the finding of highly qualitative web pages that are related to the user query. These techniques are based on the graphical representation of the Web and on the analysis of that links. This project is focused on the description and analysis of the Pagerank algorithm which is the basic operation for Google, the most popular search engine. Our work concludes with an attempt to invade in the sanctuary of Google, mentioning the secret of its success and its future plans.

ΚΕΦΑΛΑΙΟ 1ο Ο ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ

1.1 Ο ορισμός του Παγκόσμιου Ιστού

Ο Παγκόσμιος Ιστός, γνωστός και ως www ή w3 (World Wide Web-Παγκόσμιος Επεκτεινόμενος Ιστός), είναι ένα δίκτυο υπολογιστών συνδεδεμένων μεταξύ τους, όπου επιτρέπει στους χρήστες να έχουν πρόσβαση σε πληροφορίες που είναι αποθηκευμένες σε αυτό το δίκτυο [Τράκας]. Η προεργασία για τη δημιουργία του παγκόσμιου ιστού είχε ξεκινήσει στα μέσα του 1989 από τον Βρετανό Tim Berners-Lee στο CERN (Ευρωπαϊκό Εργαστήριο Φυσικής Σωματιδίων) στη Γενεύη της Ελβετίας, προσπαθώντας να βρει έναν τρόπο να αρχειοθετεί τις επιστημονικές μελέτες των συνεργατών του [GC2000]. Στις 16 Απριλίου 1991 έδωσε στη δημοσιότητα τεχνικές λεπτομέρειες για να μπορέσουν να τον χρησιμοποιήσουν και άλλοι. Στα μέσα του Δεκεμβρίου 1991 δημιουργήθηκε το πρώτο web site (ιστοσελίδα) στις ΗΠΑ στο πανεπιστήμιο Stanford από τον Πολ Κουντς και μέσα στον επόμενο χρόνο εμφανίστηκαν άλλα 25 web sites. Το CERN στις 30 Απριλίου 1993 ανακοίνωσε ότι ο παγκόσμιος ιστός αποκτάει ελεύθερη χρήση [ΨΣ2006]. Η επιτυχία του ήταν τόσο μεγάλη όπου και ενσωματώθηκε πολύ γρήγορα στις υπηρεσίες του διαδικτύου γνωρίζοντας τεράστια απήχηση χάρη στον εύκολο τρόπο περιήγησης και αναζήτησης πληροφοριών.

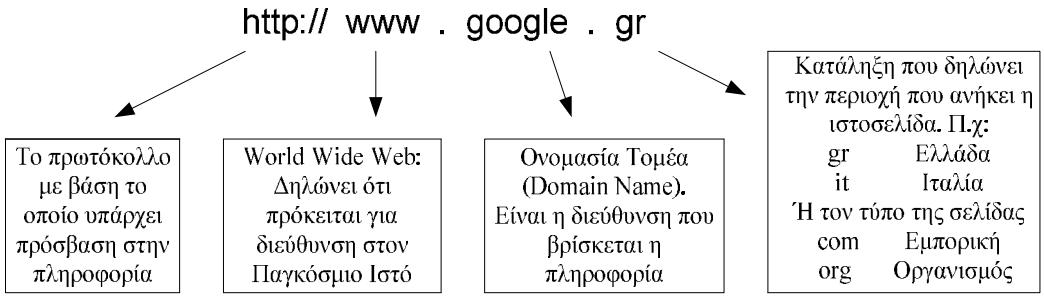
Ο Παγκόσμιος Ιστός ουσιαστικά αποτελεί έναν εικονικό χώρο όπου η επικοινωνία επιτυγχάνεται με ειδικά έγγραφα υπέρ-κειμένων (hypertexts). Τα υπέρ-κείμενα είναι κείμενα τα οποία περιέχουν συνδέσμους (links) όπου επιτρέπουν σε έναν χρήστη μέσω λέξεων-κλειδιών να μεταφέρετε από ένα κείμενο σε ένα άλλο κείμενο, εικόνα, ήχο ή ακόμα και βίντεο. Οι σύνδεσμοι βρίσκονται σε διάφορα σημεία μέσα στο κείμενο μιας σελίδας, είναι λέξεις υπογραμμισμένες με διαφορετικό χρώμα, συνήθως μπλε, όπου κάνοντας ο χρήστης «κλικ» με το ποντίκι μεταφέρεται σε μια άλλη σελίδα. Η διαδικασία αναζήτησης πληροφορίας από μια σελίδα σε μια άλλη ονομάζεται πλοήγηση (browsing).

Η λειτουργία του παγκόσμιου ιστού βασίζεται στην εφαρμογή πελάτη-εξυπηρετητή (client-web server), το οποίο απαιτεί τη συνεργασία αυτών των δύο. Κατά τη μεταξύ τους επικοινωνία χρησιμοποιείται το πρωτόκολλο HTTP (Hyper-

Text Transfer Protocol – Πρωτόκολλο Μεταφοράς Υπέρ-Κειμένου) το οποίο είναι το πιο συνηθισμένο στον ηλεκτρονικό χώρο του παγκόσμιου ιστού. Πριν τον Παγκόσμιο Ιστό και το HTTP, το πρωτόκολλο που χρησιμοποιούνταν για την μεταφορά αρχείων στο διαδίκτυο ήταν το FTP (File Transfer Protocol – Πρωτόκολλο Μεταφοράς Αρχείων) [Φιλίππου].

Οι σελίδες είναι αποθηκευμένες σε υπολογιστές (web servers – εξυπηρετητές ιστού), με μεγάλη υπολογιστική. Βρίσκονται μόνιμα συνδεδεμένοι στο διαδίκτυο για να μπορεί ο χρήστης οποιαδήποτε στιγμή να αντλεί πληροφορίες. Από την πλευρά του ο πελάτης (client) πρέπει να διαθέτει ένα πρόγραμμα ώστε να μεταφέρει τις σελίδες στην οθόνη του τοπικού υπολογιστή του. Το πρόγραμμα που εμφανίζει τις σελίδες καθώς και επιτρέπει τη μετάβαση μιας σελίδας σε μια άλλη μέσω συνδέσμων στον παγκόσμιο ιστό λέγεται φυλλομετρητής (browser). Οι πιο διαδεδομένοι φυλλομετρητές είναι ο Internet Explorer της Microsoft όπου εμφανίστηκε τον Αύγουστο το 1995, ο Firefox του Mozilla Foundation που εμφανίστηκε το 1998 και ο Chrome της Google ο οποίος παρουσιάστηκε τον Σεπτέμβριο του 2008. Οι φυλλομετρητές παρέχουν τις διάφορες λειτουργίες στους χρήστες όπως ανάγνωση e-mail, δημιουργία ιστοσελίδων, εκτέλεση αρχείων ήχου, προβολή βίντεο, κ.α.

Μέσω του Παγκόσμιου Ιστού είναι δυνατή η πρόσβαση σε κάθε πληροφορία που υπάρχει. Η χρήση του είναι πολύ απλή, αρκεί ο χρήστης να εισάγει στον φυλλομετρητή το όνομα της σελίδας και την ακριβή της διεύθυνση. Τα στοιχεία αυτά υπάρχουν στο URL (Uniform Resource Locator), όπου είναι μοναδικά για κάθε σελίδα και αποτελούνται από τα εξής βασικά μέρη (Εικόνα 1):



Εικόνα 1: Βασικά μέρη του URL

1.2 **Προγραμματισμός στον Παγκόσμιο Ιστό**

Ο Παγκόσμιος Ιστός χρησιμοποιεί τις Γλώσσες Σήμανσης (Markup Languages), οι οποίες χρησιμοποιούνται για να μπορούν οι φυλλομετρητές να διαβάζουν τις σελίδες και να τις εμφανίζουν στο χρήστη. Ονομάζονται και γλώσσες προγραμματισμού.

Η βασική γλώσσα προγραμματισμού που χρησιμοποιεί ο Παγκόσμιος Ιστός είναι η HTML (Hyper-Text Markup Language – Γλώσσα Επισήμανσης Υπερκειμένου), όπου δημιουργήθηκε το 1990 από τους Tim Berners – Lee και Anders Berglund. Η HTML είναι μια γλώσσα δημιουργίας ιστοσελίδων και περιγραφής εγγράφων. Τα έγγραφα αυτά ονομάζονται HTML documents τα οποία έχουν την κατάληξη .html ή .htm. Η γλώσσα αυτή είναι υποσύνολο της γλώσσας SGML (Standard Generalized Markup Language – Πρότυπη Γενικευμένη Γλώσσα Καταγραφής), η οποία επινοήθηκε από την IBM για τη λύση του προβλήματος της μη τυποποιημένης εμφάνισης κειμένων. Η HTML χρησιμοποιείται για να δημιουργεί κείμενα ώστε να εμφανίζονται καλύτερα στο χρήστη. Επίσης βοηθάει στην ενσωμάτωση εικόνων, βίντεο και ήχων στις σελίδες.

Το βασικό στοιχείο της γλώσσας HTML όπου γίνεται ο τρόπος παρουσίασης και διαμόρφωσης του εγγράφου είναι η εισαγωγή κατάλληλων ετικετών (tags). Οι ετικέτες είναι ένα είδος εντολής που δίνει κάποια ιδιαίτερα χαρακτηριστικά στο κείμενο καθώς καθορίζουν την τοποθεσία και τη μορφή που θα εμφανίζονται οι λέξεις μέσα σε αυτό [Μπουντουρίδης96]. Οι ετικέτες είναι λέξεις που περιβάλλονται από τα σύμβολα «< … >» και διακρίνονται σε απλές και διπλές.

Απλές ετικέτες είναι αυτές που δεν έχουν ετικέτα τερματισμού, ένα απλό παράδειγμα είναι η ετικέτα
 που δηλώνει την αλλαγή γραμμής μέσα στο κείμενο. Ένα παράδειγμα διπλής ετικέτας όπου έτσι ξεκινούν και όλες οι σελίδες HTML, είναι η ετικέτα έναρξης <html> και τελειώνουν με την ετικέτα τερματισμού </html>. Κάθε σελίδα HTML αποτελείται από δύο τμήματα, το <head> … </head> και το <body> … </body>. Όπου το πρώτο είναι η κεφαλή του κειμένου το οποίο καθορίζει τις παραμέτρους της σελίδας όπως τον τίτλο, τη μορφή που θα έχει καθώς και τη σχέση της με τις άλλες σελίδες. Αν για παράδειγμα θέλουμε να γράψουμε τον τίτλο της σελίδας θα χρησιμοποιήσουμε το <title> … </title>. Το δεύτερο τμήμα καθορίζει το σώμα της σελίδας, δηλαδή τις πληροφορίες που θα έχει η σελίδα μέσα είτε με τη μορφή κειμένου, εικόνων ή διασυνδέσεων προς άλλες σελίδες. Για να

γράψουμε την επικεφαλίδα του κειμένου θα χρησιμοποιήσουμε το <h1> … </h1>, ή αν θέλουμε να γράψουμε τη φράση «κάνει ζέστη» με έντονα γράμματα θα γράψουμε κάνει ζέστη ενώ αν θέλουμε να εμφανίζεται με πλάγια γράμματα θα γράψουμε κάνει ζέστη .

Ένα ακόμα βασικό στοιχείο της γλώσσας HTML είναι οι σύνδεσμοι (links). Οι σύνδεσμοι βρίσκονται ανάμεσα στις ετικέτες <α> </α> και έχουν την εξής μορφή <α href= «urlpage»> hyperlink – text</α>. Στο πεδίο «href» δίνουμε τη διεύθυνση της σελίδας προς την οποία δείχνει ο συγκεκριμένος σύνδεσμος [Μακρής08]. Ενδιάμεσα των δύο ετικετών δίνουμε το κείμενο που παρουσιάζει ο σύνδεσμος και η σελίδα προς την οποία δείχνει.

Μια άλλη γλώσσα σήμανσης που χρησιμοποιείται από τον παγκόσμιο ιστό είναι η XML (eXtensible Markup Language), όπου σχεδιάστηκε για να ικανοποιήσει ανάγκες οι οποίες δεν μπορούν να λυθούν από την γλώσσα HTML [Σάμψων2003]. Αναπτύχθηκε από το διεθνή οργανισμό W3C (World Wide Web Consortium) το 1996 όπου εδραιώθηκε από τον John Bosak της Sun Microsystems. Με την XML οι σχεδιαστές του παγκόσμιου ιστού έλυσαν πολλά προβλήματα καθώς δίνει καλύτερο στυλ και δομή στα έγγραφα από αυτό της HTML. Αποτελεί επέκταση της HTML καθώς χρησιμοποιεί τις περιγραφικές της εντολές και είναι υπεύθυνη για την XHTML, μια ανασχεδιασμένη HTML. Δίνει έμφαση στην απλότητα, τη γενικότητα και τη χρησιμότητα στο διαδίκτυο. Η XML δεν είναι απλά μία γλώσσα σήμανσης αλλά και μια μετά-γλώσσα η οποία χρησιμοποιείται για να καθορίσει νέες γλώσσες. Οι στόχοι της γλώσσας XML είναι [Μπαλής08]:

- Να είναι εύχρηστη στο διαδίκτυο.
- Να υποστηρίζει μεγάλη ποικιλία από εφαρμογές.
- Να είναι ευανάγνωστα τα XML έγγραφα.
- Να προετοιμάζεται γρήγορα ο σχεδιασμός XML, να είναι τυπικός και περιεκτικός.
- Να δημιουργούνται εύκολα τα XML έγγραφα.
- Να υπάρχει ευκολία στην ανάπτυξη των προγραμμάτων που επεξεργάζονται XML έγγραφα.

Τέλος για την περιγραφή εικονικών κόσμων ο παγκόσμιος ιστός χρησιμοποιεί τη γλώσσα δημιουργίας εικονικής πραγματικότητας VRML (Virtual Reality

Modeling Language). Συνελήφθη σαν ιδέα την άνοιξη του 1994 στο συνέδριο για τον παγκόσμιο ιστό στη Γενεύη της Ελβετίας [VRML]. Η VRML είναι μια γλώσσα η οποία αναλύει τη γεωμετρική περιγραφή των αντικειμένων που υπάρχουν διατεταγμένα μέσα σε ένα τρισδιάστατο χώρο. Δείχνει δηλαδή από τι αποτελείται ένα αντικείμενο, όπως το γεωμετρικό του σχήμα, τις διάφορες ιδιότητες της επιφάνειάς του (λαμπρότητα, χρώμα, ομαλότητα κ.α.) καθώς και τη θέση που έχει στον τρισδιάστατο αυτό χώρο. Για να δούμε όμως ένα αρχείο VRML θα πρέπει να υπάρχει ένας ειδικός φυλλομετρητής VRML (browser VRML) [Nadeau98].

1.3 **Υπερσύνδεσμοι** (hyperlinks)

Οι υπερσύνδεσμοι, ή απλούστερα οι σύνδεσμοι, μεταξύ των σελίδων μεταφέρουν μια πολύ σημαντική πληροφορία η οποία έχει να κάνει με τη σχέση των σελίδων που συνδέονται μέσω αυτών. Στη δομή που σχηματίζεται από τους συνδέσμους εμπεριέχεται ο παράγοντας της ανθρώπινης κρίσης, ο οποίος απουσιάζει από τις μηχανές αναζήτησης, για να καθοριστεί καλύτερα η έννοια της ποιότητας μιας σελίδας.

Πιο συγκεκριμένα, έστω ότι η σελίδα Α περιλαμβάνει στο κείμενο της έναν σύνδεσμο στη σελίδα Β. Η ύπαρξη του συνδέσμου αυτού φανερώνει ότι ο κατασκευαστής της σελίδας Α πιστεύει ότι η σελίδα Β περιέχει σημαντική και αξιόλογη πληροφορία και πιθανόν σχετική με αυτή που περιέχεται στη σελίδα Α. Με αυτόν τον τρόπο, μπορεί να χρησιμοποιηθεί ο αριθμός των υπερσυνδέσμων που δείχνουν στη σελίδα Β (in-degree) ως ένα μέτρο για την αξιολόγηση της ποιότητας της.

Ακολουθώντας την ίδια λογική, μπορεί να θεωρηθεί ότι εάν η σελίδα Α έχει υπερσυνδέσμους σε πολλές καλές και ποιοτικές σελίδες τότε η άποψη και η κρίση του κατασκευαστή της σελίδας Α αποκτά μεγαλύτερη σημασία και γίνεται αξιοπρόσεκτη. Επομένως, το γεγονός ότι η σελίδα Α έχει ένα υπερσύνδεσμο προς τη σελίδα Β υποδηλώνει ότι ίσως και η σελίδα Β είναι μια ποιοτική σελίδα.

Ωστόσο, η χρήση των υπερσυνδέσμων ως μια σημαντική πηγή πληροφορίας κρύβει πολλούς κινδύνους και παγίδες. Αυτοί προκύπτουν από τη διαφορετική σημασία που μπορεί να έχει κάποιος υπερσύνδεσμος. Δηλαδή πολλοί από τους

υπερσυνδέσμους που εμφανίζονται στο διαδίκτυο έχουν δημιουργηθεί για να ικανοποιήσουν άλλους σκοπούς. Για παράδειγμα έχουν δημιουργηθεί καθαρά και μόνο για λόγους πλοήγησης, ώστε να διευκολύνουν την μετακίνηση του χρήστη μέσα στη σελίδα. Ακόμα οι λόγοι μπορεί να είναι εμπορικοί ή και διαφημιστικοί, καθώς πολλές επιχειρήσεις παρουσιάζονται και διαφημίζονται μέσω του διαδικτύου, επακόλουθο της ανάπτυξής του.

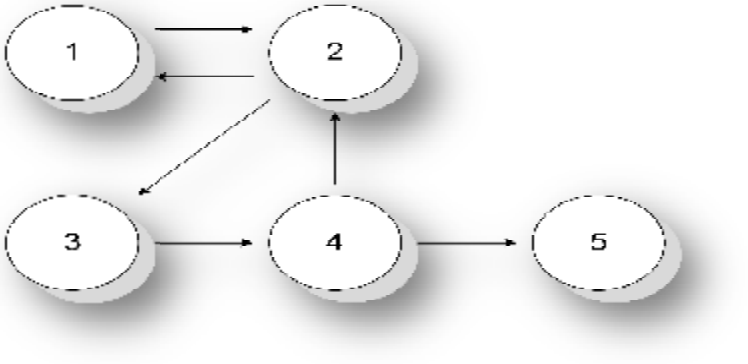
Λαμβάνοντας υπόψη και τους υπερσυνδέσμους μεταξύ των σελίδων, και όχι μόνο το περιεχόμενο αυτών ως πηγή πληροφορίας, ο παγκόσμιος ιστός μπορεί να αναπαρασταθεί μέσω ενός γραφήματος, όπως αναφέρουμε αναλυτικά στην επόμενη ενότητα.

1.4 **Ο Παγκόσμιος Ιστός ως γράφημα**

Ο παγκόσμιος ιστός αποτελεί μια συλλογή εγγράφων, όπου εκτός από τα δεδομένα που περιέχει στους κόμβους του χαρακτηρίζεται και από τους συνδέσμους που τους συνδέουν μεταξύ τους. Θα μπορούσε ο παγκόσμιος ιστός να θεωρηθεί ως ένα γράφημα[1] (Web Graph), και πιο συγκεκριμένα ένα κατευθυνόμενο γράφημα, στο οποίο οι κόμβοι είναι οι σελίδες και οι ακμές οι σύνδεσμοι. Ένα απλό παράδειγμα κατευθυνόμενου γραφήματος δίνουμε στο Σχήμα 1: Αποτελείται από 5 σελίδες (κόμβους) όπου η σελίδα 1 «δείχνει» στη σελίδα 2, η σελίδα 2 δείχνει στις σελίδες 1 και 3, η σελίδα 3 δείχνει στη σελίδα 4 και η σελίδα 4 στις σελίδες 2 και 5.

Το γράφημα του Παγκόσμιου Ιστού είναι ένα συναρπαστικό αντικείμενο μελέτης καθώς έχει εκατοντάδες εκατομμύρια κόμβους σήμερα και πάνω από ένα δισεκατομμύριο συνδέσμους.

[1] Γράφημα ή γράφος (graph) είναι μια αφηρημένη αναπαράσταση ενός συνόλου στοιχείων, όπου μερικά ζευγάρια στοιχείων συνδέονται μεταξύ τους με δεσμούς. Τα διασυνδεδεμένα στοιχεία ονομάζονται κορυφές ενώ οι δεσμοί που συνδέουν τα ζευγάρια των κορυφών ονομάζονται ακμές. Για περισσότερα στοιχεία από τη Θεωρία Γραφημάτων, ενδεικτικά αναφέρουμε το [Μαυρονικόλας11] και για αλγορίθμους σε γραφήματα το [TS2011].



Σχήμα 1: Παράδειγμα ενός κατευθυνόμενου γραφήματος

Τα τελευταία χρόνια πολλοί ερευνητές που έχουν προσπαθήσει ν' αποδώσουν τον ιστό με μορφή γραφήματος, κατέληξαν στο συμπέρασμα ότι αποτελείται από ένα σύνολο από ηλεκτρονικές σελίδες, μεγάλης πολυπλοκότητας οι οποίες εισάγονται και εξάγονται με μια διαδικασία εντελώς αποκεντρωμένη και χαοτική, καθώς ο καθένας μπορεί να δημιουργήσει μια σελίδα όπως θέλει χωρίς κάποια καθορισμένη δομή και πρότυπα περιεχομένου. Οι σελίδες του ιστού μπορούν να έχουν γραφτεί σε διάφορες γλώσσες, διαλέκτους ή μορφές από άτομα με διαφορετικό υπόβαθρο, μόρφωση, κουλτούρα, ενδιαφέροντα και κίνητρα. Όπως επίσης και να περιέχουν αλήθειες, ψέματα, προπαγάνδα, σοφία ή ανοησίες. Κάθε σελίδα μπορεί να διαφέρει σε επίπεδο μεγέθους, περιεχομένου κλπ, και αυτό το επιλέγει ο δημιουργός της. Σε καθημερινή βάση προστίθενται αρκετές εκατοντάδες καινούριες σελίδες με αποτέλεσμα το μέγεθος του παγκόσμιου ιστού να αυξάνει διαρκώς και με ιδιαίτερα γρήγορους ρυθμούς.

ΚΕΦΑΛΑΙΟ 2ο ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

2.1 Ανάκτηση πληροφορίας

Ο παγκόσμιος ιστός αποτελείται από ένα πολύ μεγάλο αριθμό σελίδων, που καθημερινά αυξάνεται. Ο καθένας έχει τη δυνατότητα να δημιουργήσει και να δημοσιεύσει μια σελίδα στον παγκόσμιο ιστό χωρίς να πάρει έγκριση από κάποιο κεντρικό οργανισμό. Ο άναρχος αυτός τρόπος εισαγωγής των σελίδων είναι και το βασικό μειονέκτημά του παγκόσμιου ιστού. Παράλληλα, με την ίδια ευκολία, πολλές σελίδες και δικτυακοί τόποι[2] καταργούνται καθημερινά από τους ιδιοκτήτες τους [BR2009]. Παρά το συνεχώς αυξανόμενο μέγεθός του ο παγκόσμιος ιστός είναι ένας ανοργάνωτος χώρος. Η έλλειψη προτύπων και η παντελής έλλειψη δομής και οργάνωσης των σελίδων αυτών έχει οδηγήσει σε ένα χαοτικό σύνολο στο οποίο η πρόσβαση στην πληροφορία είναι ιδιαίτερα δυσχερής. Μερικοί ακόμα λόγοι που μειώνεται η ορθή εύρεση σωστών πληροφοριών είναι οι εξής [Μακρής08]:

- Οι χρήστες συνηθίζουν να δίνουν μικρά ερωτήματα, της τάξης των τριών λέξεων το πολύ, χωρίς να είναι πρόθυμοι να δώσουν επιπλέον πληροφορίες για αυτό που αναζητούν. Δεν δίνουν σημασία στη σωστή διατύπωση της αναζήτησής τους με αποτέλεσμα τα ερωτήματά τους να είναι αρκετά ασαφή, και να παίρνουν λάθος απαντήσεις όπου δεν ταιριάζουν με την πραγματική τους πληροφοριακή ανάγκη.
- Στον παγκόσμιο ιστό όπως αναφέραμε καθημερινά προστίθενται νέες σελίδες, όπου πολλές από αυτές αλλάζουν μορφή ανά τακτά χρονικά διαστήματα. Αυτό επιβαρύνει το σύστημα διότι πρέπει να ενημερώνει διαρκώς τις ήδη αποθηκευμένες σελίδες και να προσθέτει συνέχεια καινούργιες.
- Υπάρχουν σελίδες στον παγκόσμιο ιστό όπου σκοπός τους δεν είναι η παροχή πληροφοριών, έτσι οι χρήστες που ανοίγουν τέτοιες σελίδες χάνουν αρκετό από το χρόνο τους. Επίσης μερικές σελίδες εστιάζουν σε κάποιο θέμα, άλλες δίνουν πληροφορίες για πολλά θέματα, και πολλές φορές άσχετα μεταξύ τους.

[2] Δικτυακός τόπος ή ιστότοπος είναι μια συλλογή από σελίδες, βίντεο, εικόνες κ.α., τα οποία φιλοξενούνται στην ίδια περιοχή (domain) του παγκόσμιου ιστού. Η υπηρεσία αυτή δίνει τη δυνατότητα στους χρήστες να δημιουργήσουν οποιουδήποτε είδους περιεχόμενο στις σελίδες τους. Το σύνολο των δικτυακών τόπων αποτελεί τον παγκόσμιο ιστό.

- Κάποιες σελίδες μπορεί να μην περιέχουν σωστή πληροφορία, να μην είναι αξιόπιστη, να αναφέρονται σε ανακρίβειες ή ακόμα και να γίνεται εσκεμμένη παραπλάνηση μέσω κάποιων από αυτές.
- Τέλος η επεξεργασία όλων των σελίδων που υπάρχουν στον παγκόσμιο ιστό απαιτεί μεγάλο κόστος χρόνου και χώρου και είναι ουσιαστικά ανέφικτη λόγω του μεγάλου όγκου της υπάρχουσας πληροφορίας.

Έτσι λοιπόν για να λυθεί το πρόβλημα της δύσκολης αναζήτησης μιας πληροφορίας στον παγκόσμιο ιστό είναι απαραίτητες οι υπηρεσίες αναζήτησης. Υπάρχουν τρία είδη υπηρεσιών αναζήτησης: οι Μηχανές Αναζήτησης (Search Engines), Μετά-Μηχανές Αναζήτησης (Meta-Search Engines) και οι Θεματικοί Κατάλογοι.

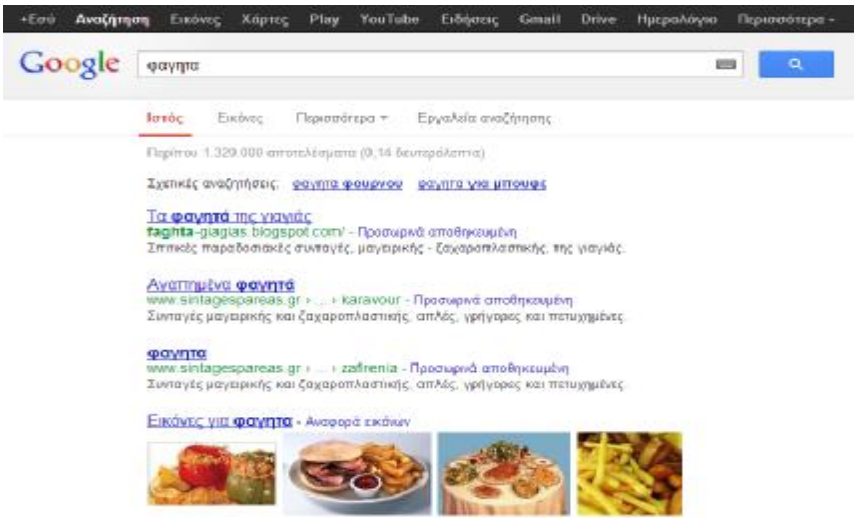
2.2 **Μηχανές Αναζήτησης** (Search Engines)

Οι μηχανές αναζήτησης είναι προγράμματα όπου επιτρέπουν στους χρήστες να κάνουν αναζήτηση των πληροφοριών που επιθυμούν. Διαθέτουν μια βάση δεδομένων όπου είναι καταγεγραμμένες όλες οι διευθύνσεις του παγκόσμιου ιστού και βάζουν σε μια τάξη όλες αυτές τις διευθύνσεις που παρέχουν τις πληροφορίες. Οι βάσεις δεδομένων των μηχανών αναζήτησης είναι ρυθμισμένες έτσι ώστε μέσα σε λίγα δευτερόλεπτα να δίνουν τα αποτελέσματα στους χρήστες [BR2009].

Η λειτουργία τους είναι πολύ εύκολη, αρκεί ο χρήστης να επιλέξει την αρχική σελίδα μιας μηχανή αναζήτησης και να πληκτρολογήσει τους όρους που περιγράφουν το θέμα που τον ενδιαφέρει, με όσο μεγαλύτερη σαφήνεια και περιεκτικότητα γίνεται. Η μηχανή αναζήτησης θα του επιστρέψει μια σελίδα αποτελεσμάτων, όπου τα αποτελέσματα είναι μια λίστα από σελίδες, ταξινομημένη ανάλογα με το ποσοστό σχετικότητας της κάθε σελίδας με βάση το ερώτημα του χρήστη (εικόνα 2). Έτσι ο χρήστης θα επιλέξει τη σελίδα που θα καλύψει καλύτερα την ανάγκη του.

Η χρήση των μηχανών αναζήτησης επιφέρει πολλά οφέλη, τόσο για τους χρήστες όσο και για τις επιχειρήσεις που διαθέτουν έναν δικτυακό τόπο. Αν σκεφτούμε ότι κάθε μέρα τεράστιος αριθμός χρηστών χρησιμοποιούν κάποια μηχανή αναζήτησης και ότι στα αποτελέσματά της εμφανίζεται και ο δικτυακός τόπος κάποιας επιχείρησης σημαίνει αυτόματα ότι υπάρχει αύξηση του αριθμού χρηστών-πελατών που θα επισκεφτούν την σελίδα της. Επομένως η επιχείρηση αυτή

επιτυγχάνει μέσω της μηχανής αναζήτησης να προσελκύσει πιο γρήγορα και εύκολα πελάτες όπου θα ενδιαφέρονται για τα προϊόντα της και χωρίς κανένα κόστος για την ίδια.



Εικόνα 2: Λίστα αποτελεσμάτων αναζήτησης

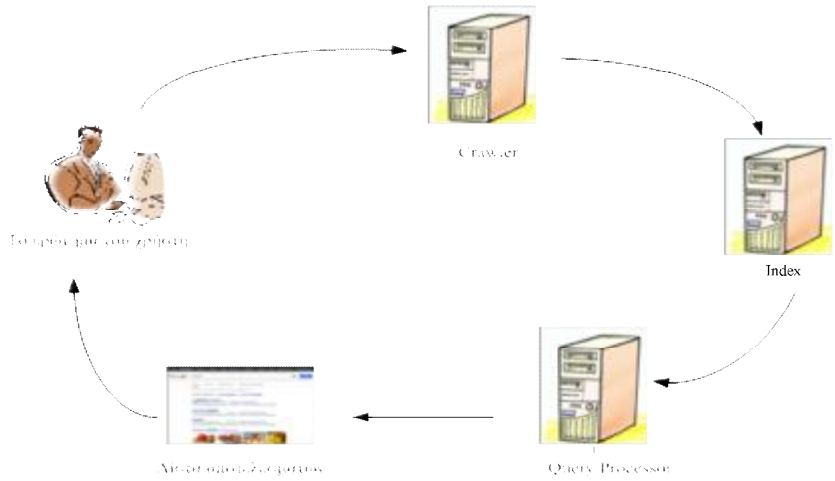
Μια μηχανή αναζήτησης αποτελείται από τρία βασικά μέρη [ΘΖ2005, ΜΥ2011]:

- Τον Crawler (ή Spider ή Robot), ο οποίος είναι ο ανιχνευτής της μηχανής αναζήτησης. Είναι ένα αυτόματο πρόγραμμα που επισκέπτεται σελίδες ακολουθώντας τους συνδέσμους στον παγκόσμιο ιστό, τις διαβάζει και τις μεταβιβάζει στη βάση δεδομένων της μηχανής αναζήτησης. Κατά διαστήματα ο Crawler επισκέπτεται τις σελίδες που ήδη είχε επισκεφτεί για να βρει τυχόν αλλαγές που έχουν γίνει έτσι ώστε η βάση δεδομένων της μηχανής αναζήτησης να μένει πάντα ενημερωμένη. Η λειτουργία του Crawler είναι περίπου η ίδια με αυτή ενός φυλλομετρητή όπου στέλνει αίτηση για μια σελίδα, την κατεβάζει και τη διαβιβάζει στη βάση δεδομένων. Βέβαια ο Crawler είναι πολύ πιο γρήγορος από τον φυλλομετρητή και αναζητά ταυτόχρονα εκατοντάδες διαφορετικές σελίδες.
- Το ευρετήριο (index), το οποίο είναι μια βάση δεδομένων που περιέχει αντίγραφα από τις σελίδες που επισκέφθηκε και διάβασε ο Crawler. Το ευρετήριο επεξεργάζεται αυτές τις σελίδες και αποφασίζει ποιες θα ταξινομήσει, καθώς μπορεί μερικές σελίδες να εμφανίζονται πολλές φορές. Επίσης αξιολογεί το κείμενο των σελίδων, τις συνδέσεις και τα άλλα στοιχεία

του περιεχομένου των σελίδων. Στη συνέχεια αρχειοθετεί τα αντίγραφα των σελίδων σε έναν κατάλογο και έτσι η μηχανή αναζήτησης έχει στη διάθεσή της έναν μεγάλο αριθμό σελίδων για να τις χρησιμοποιήσει στο επόμενο βήμα. Να τονίσουμε ότι οι σελίδες που σαρώνονται από τον Crawler αξιολογούνται με βάση κάποια κριτήρια για κάθε μηχανή αναζήτησης. Κάθε μηχανή αναζήτησης θέτει και κάποιες προϋποθέσεις, αν οι δικτυακοί τόποι τις πληρούν τότε εισάγονται στον κατάλογό της, αν όχι δεν συμπεριλαμβάνονται ή κατατάσσονται χαμηλά στα αποτελέσματα που θα επιστρέψει η μηχανή αναζήτησης σε μια αναζήτηση του χρήστη.

- Τον επεξεργαστή ερωτήματος (query processor), το οποίο είναι ένα πρόγραμμα που ερευνά το ευρετήριο για να βρει σελίδες όπου να ταιριάζουν με τις λέξεις-κλειδιά του ερωτήματος που έθεσε ο χρήστης. Πιο αναλυτικά, όταν ο χρήστης πληκτρολογήσει τις λέξεις-κλειδιά στη μηχανή αναζήτησης, ενεργοποιείται ο επεξεργαστής ερωτήματος για να εντοπίζει και κατατάσσει τις σελίδες χρησιμοποιώντας έναν αλγόριθμο αξιολόγησης βασισμένο σε κάποιους κανόνες, και να εξετάζει το βαθμό σχετικότητας της κάθε σελίδας. Έπειτα κατατάσσει στα αποτελέσματα τις σχετικές με το θέμα του χρήστη σελίδες και του τις παρουσιάζει με τη μορφή συνδέσμων.

Μπορούμε να συνοψίσουμε την αρχιτεκτονική της επεξεργασία του ερωτήματος ενός χρήστη στην εικόνα 3.



Εικόνα 3: Αρχιτεκτονική επεξεργασία του ερωτήματος του χρήστη

Κάποιες από τις πιο δημοφιλής μηχανές αναζήτησης είναι η Google (www.google.com), η Yahoo (www.yahoo.com), η Altavista (www.altavista.com) που αγοράστηκε από την Yahoo και από τις 8/7/2013 είναι εκτός λειτουργίας, η Ask (www.ask.com), και η Lycos (www.lycos.com)[3].

2.3 **Μειονεκτήματα των μηχανών αναζήτησης**

Με τη χρήση των μηχανών αναζήτησης οι χρήστες μπορούν να αναζητούν για πληροφορίες που τους ενδιαφέρουν στον παγκόσμιο ιστό. Υπάρχει όμως η πιθανότητα να τους επιστρέψουν ένα μεγάλο ποσοστό άσχετων σελίδων που δεν έχουν αξιολογηθεί ή είναι παλιές ή περιέχουν ανακριβείς ή ελλιπείς πληροφορίες. Αυτό είναι ένα από τα μεγαλύτερα μειονεκτήματα των μηχανών αναζήτησης καθώς ο χρήστης θα επισκεφθεί το πολύ τις 10 πρώτες σελίδες που θα του εμφανίσουν και οι περισσότερες από αυτές δεν θα ικανοποιούν την ανάγκη του. Έτσι ορισμένες μηχανές αναζήτησης αξιολογούν τις σελίδες με βάση το περιεχόμενό τους, το πόσο συχνά εμφανίζονται οι λέξεις-κλειδιά μέσα στο κείμενο και αν είναι σε εμφανή σημεία όπως στην επικεφαλίδα, στον τίτλο ή είναι γραμμένες με τέτοιο τρόπο ώστε να δίνεται περισσότερη βαρύτητα π.χ. να είναι υπογραμμισμένες.

Οι μηχανές αναζήτησης που αξιολογούν τις σελίδες με βάση το περιεχόμενό τους παρουσιάζουν ένα άλλο πρόβλημα. Είναι πολύ εύκολο να ξεγελαστούν από τους σχεδιαστές σελίδων όπου βάζουν σε κρυφά, από τους χρήστες, σημεία πολλές φορές κάποιες λέξεις-κλειδιά με αποτέλεσμα να βαθμολογούνται περισσότερο οι σελίδες τους [Μακρής08]. Επίσης με βάση αυτήν την μέθοδο που χρησιμοποιούν οι μηχανές αναζήτησης προκύπτει το πρόβλημα ότι ο χρήστης μπορεί να εισάγει μια λέξη η οποία να έχει δύο έννοιες, όπως για παράδειγμα η λέξει «puma» η οποία μπορεί να είναι είτε το ζώο είτε η μάρκα παπουτσιών. Έτσι αν ο χρήστης δεν δώσει άλλες πληροφορίες για αυτό που ψάχνει, η μηχανή αναζήτησης θα του επιστρέψει και σελίδες που δεν τον ενδιαφέρουν. Σημαντικό πρόβλημα είναι επίσης και οι συνώνυμες λέξεις, όπως για παράδειγμα αν ο χρήστης εισάγει τη λέξη «αμάξι» η μηχανή αναζήτησης δεν θα του εμφανίσει αποτελέσματα που να περιέχουν τη λέξη «αυτοκίνητο» με αποτέλεσμα να χαθεί αρκετή σημαντική πληροφορία.

[3] Για περισσότερες μηχανές αναζήτησης μπορείτε να επισκεφτείτε την ιστοσελίδα http://en.wikipedia.org/wiki/List_of_search_engines.

Όπως αναφέραμε παραπάνω ο παγκόσμιος ιστός δεν έχει κάποια δομή ή οργάνωση λόγω του μεγάλου όγκου πληροφοριών με αποτέλεσμα να προκύπτουν όλα αυτά τα προβλήματα. Όμως το βασικό στοιχείο που παρέχει κάποια μορφή οργάνωσης και δομής είναι οι σύνδεσμοι, οι οποίοι δίνουν μια πιο χαλαρή μορφή συνοχής στις αποθηκευμένες πληροφορίες του παγκόσμιο ιστού.

2.4 **Μετά-Μηχανές Αναζήτησης** (Meta-Search Engines)

Οι μετά-μηχανές αναζήτησης, σε αντίθεση με τις απλές μηχανές αναζήτησης, οι οποίες όπως είδαμε χρησιμοποιούν έναν crawler για να συγκεντρώσουν μια δική τους βάση δεδομένων, δεν διαθέτουν δικό τους ευρετήριο, αλλά αντλούν τα αποτελέσματά τους από τα ευρετήρια άλλων μηχανών αναζήτησης. Με άλλα λόγια, θα μπορούσαμε να πούμε ότι είναι μηχανές αναζήτησης πάνω σε άλλες μηχανές αναζήτησης.

Ο τρόπος λειτουργίας τους είναι ίδιος με τον τρόπο λειτουργίας των απλών μηχανών αναζήτησης. Ο χρήστης πληκτρολογεί στη φόρμα εισαγωγής ερωτήματος τις λέξεις-κλειδιά ή άλλες λέξεις που περιγράφουν το θέμα για το οποίο επιθυμεί την ανάκτηση πληροφορίας. Στην πορεία οι μετά-μηχανές στέλνουν το ερώτημα του χρήστη σε μια σειρά από προκαθορισμένες μηχανές αναζήτησης (ή και θεματικούς καταλόγους) και αφού αφαιρέσουν τις διπλοεγγραφές, παρουσιάζουν ένα μέρος από τα αποτελέσματα της κάθε μιας μέσα σε λίγα δευτερόλεπτα (συχνά ανακτά το 10% των αποτελεσμάτων) [MY2011].

Μια τέτοια μηχανή όμως απαιτεί, όπως είναι φυσικό, περισσότερο χρόνο για την εκτέλεση του ερωτήματος καθώς θα πρέπει να πραγματοποιήσει ελέγχους σε πολλές άλλες μηχανές αναζήτησης. Το πλεονέκτημα αυτών των μηχανών αναζήτησης έναντι των απλών μηχανών είναι σημαντικό καθώς ο χρήστης κερδίζει χρόνο εισάγοντας μόνο μια φορά το ερώτημά του και επιτυγχάνει καλύτερη κάλυψη, αφού τα ευρετήρια δεν ανήκουν σε μια, αλλά σε δύο ή και περισσότερες βάσεις δεδομένων. Επιπλέον, οι μετά-μηχανές αναζήτησης συχνά επιστρέφουν σχετικά αποτελέσματα σε ασαφή ερωτήματα του χρήστη που μια απλή μηχανή αναζήτησης μπορεί να μην τα καταφέρει [Lazar00]. Τέλος, ο χρήστης εξοικονομεί το χρόνο που θα απαιτούσε για να πραγματοποιήσει μια στοιχειώδης αξιολόγηση στις μηχανές αναζήτησης προκειμένου να επιλέξει αυτή που θα χρησιμοποιούσε.

Κάποιες από τις πιο δημοφιλής μετά-μηχανές αναζήτησης είναι οι www.metacrawler.com, www.dogpile.com, www.excite.com.[4]

2.5 **Θεματικοί Κατάλογοι**

Οι θεματικοί κατάλογοι είναι συνδέσεις σε διάφορες σελίδες ή δικτυακούς τόπους, οργανωμένες σε διάφορες θεματικές ενότητες και υποενότητες ανάλογα με το θέμα που οδηγεί στην σελίδα. Όπως για παράδειγμα συμβαίνει με τα βιβλία σε μια βιβλιοθήκη έτσι με παρόμοιο τρόπο και στους θεματικούς καταλόγους οι πληροφορίες είναι ιεραρχικά δομημένες και οργανωμένες. Ένας θεματικός κατάλογος μπορεί να περιέχει και σχόλια για τις συνδέσεις ώστε να μπορεί ο κάθε χρήστης να πληροφορείται για το τι περιλαμβάνει ένας θεματικός κατάλογος [ΘΖ2005]. Η οργάνωση και η κατηγοριοποίηση των θεματικών καταλόγων γίνεται από εξειδικευμένο προσωπικό. Κάθε καταχωρημένη σελίδα στον παγκόσμιο ιστό καταχωρείται στον ανάλογο θεματικό κατάλογο.

Υπάρχουν δύο είδη θεματικών καταλόγων, οι ακαδημαϊκοί ή επαγγελματικοί και οι εμπορικοί θεματικοί κατάλογοι. Στους ακαδημαϊκούς θεματικούς καταλόγους οι σελίδες αξιολογούνται με συγκεκριμένα κριτήρια πριν την κατάταξη τους σε κάποια θεματική κατηγορία, δημιουργούνται από ειδικούς και ο σκοπός τους είναι η υποστήριξη των ερευνητών και η ποιότητα. Οι εμπορικοί θεματικοί κατάλογοι δεν αξιολογούνται αλλά απλά κατατάσσονται στην ανάλογη θεματική κατηγορία. Ο σκοπός τους είναι το κέρδος, απευθύνονται στο ευρύ κοινό και έχουν έσοδα από διαφημίσεις.

Για τον σχεδιασμό των θεματικών καταλογών υπάρχουν δύο βασικοί μέθοδοι ανάλογα με το ποιος ταξινομεί τις κατηγορίες:

- Το κλειστό μοντέλο, στο οποίο η ταξινόμηση γίνεται από μια μικρή ομάδα ατόμων, π.χ. από κάποιους υπαλλήλους μιας εταιρείας. Σε αυτήν την ομάδα τα κριτήρια που χρησιμοποιούν για την ταξινόμηση είναι συγκεκριμένα και τα ακολουθούν πιστά με αποτέλεσμα τον σωστό σχεδιασμό σε όλη την έκταση του καταλόγου.

[4] Για περισσότερες μηχανές αναζήτησης μπορείτε να επισκεφτείτε την ιστοσελίδα http://en.wikipedia.org/wiki/List_of_search_engines.

- Το ανοικτό μοντέλο, στο οποίο η ταξινόμηση γίνεται από εθελοντές. Στο μοντέλο αυτό οποιοσδήποτε χρήστης μπορεί να προσθέσει περιγραφές, σχόλια, καθώς και συνδέσμους. Οι θεματικοί κατάλογοι αυτού του τύπου τείνουν να έχουν πολλά προβλήματα σε σχέση με την ποιότητα, αλλά λόγω του χαμηλού κόστους και ελευθερίας στην χρήση είναι πιο δημοφιλείς στους χρήστες του παγκόσμιου ιστού.

Η περιήγηση στους θεματικούς καταλόγους είναι μια απλή διαδικασία για τον κάθε χρήστη. Ενδεικτικά αναφέρουμε την ιστοσελίδα www.directoryworld.net όπου υπάρχουν οι βασικές θεματικές ενότητες και επιλέγοντας ο χρήστης μια ενότητα εμφανίζονται άλλες υποενότητες. Κάποιοι από τους πιο δημοφιλής θεματικούς καταλόγους είναι οι www.in.gr, www.pathfinder.gr, www.ert.gr, www.xo.gr, και www.ego.gr.

Οι σημερινές υπηρεσίες αναζήτησης δεν διακρίνονται πάντα σε μηχανές αναζήτησης ή θεματικούς καταλόγους και αυτό γιατί στην ίδια σελίδα μπορεί να συνυπάρχουν μια μηχανή αναζήτησης και ένας θεματικός κατάλογος. Το περιεχόμενο ενός θεματικού καταλόγου μπορεί να ερευνηθεί από μια μηχανή αναζήτησης, όπως και το ευρετήριο είτε χωριστά από αυτό είτε ταυτόχρονα. Επίσης οι περισσότεροι θεματικοί κατάλογοι διαθέτουν ένα δικό τους μηχανισμό αναζήτησης για να μπορεί να ερευνηθεί με λέξεις-κλειδιά το περιεχόμενό, οι κατηγορίες, οι συνδέσεις και οι περιγραφές, όχι όμως το πλήρες κείμενο των σελίδων τους.

2.6 **Αναζήτηση μέσω υπερσυνδέσμων**

Όπως αναφέραμε και παραπάνω, με τη δημιουργία των μηχανών αναζήτησης έγινε προσπάθεια να γίνει ευκολότερη η πλοήγηση μέσα στον παγκόσμιο ιστό και να περιορίσουν το χώρο της αναζήτησης μέσω της χρήσης συγκεκριμένων λέξεων κλειδιών. Στην αρχή που ο όγκος της πληροφορίας ήταν πολύ μικρός, οι μηχανές αναζήτησης χρησιμοποιούσαν λίστες οι οποίες περιείχαν τα πιο διαδεδομένα θέματα και κατασκευάζονταν από ανθρώπους. Διατηρούνταν δηλαδή σε ένα ευρετήριο, το οποίο περιείχε μια λίστα λέξεων με όλες τις σελίδες που περιείχαν κάθε λέξη και χρησιμοποιούνταν στη συνέχεια για να απαντηθούν ερωτήματα των χρηστών. Τα επόμενα χρόνια όμως που ο όγκος των πληροφοριών υπερδιπλασιάστηκε, η συντήρηση τέτοιων λιστών από ανθρώπους ήταν αδύνατη. Έτσι, δημιουργήθηκαν οι

αυτοματοποιημένες μηχανές αναζήτησης, οι οποίες βασίζονται στην ποιότητα των λέξεων-κλειδιών και δίνουν αποτελέσματα τα οποία περιέχουν χιλιάδες έως και εκατομμύρια σελίδες. Δυστυχώς όμως μόνο μερικές χιλιάδες από αυτές τις σελίδες ανταποκρίνονται στις ανάγκες των χρηστών, ενώ τα υπόλοιπα αποτελέσματα είναι χαμηλής ποιότητας. Επομένως παρουσιάστηκε η ανάγκη για κάποιου είδους βαθμολόγησης της σημαντικότητας και της σχετικότητας των ανακτημένων σελίδων.

Για να αντιμετωπίσουν αυτό το πρόβλημα οι μηχανές αναζήτησης χρησιμοποίησαν απλές ευρετικές μεθόδους (heuristics) ώστε να επιτύχουν βαθμολόγηση των σελίδων. Τέτοιες μέθοδοι λαμβάνουν υπ' όψιν τη συχνότητα παρουσίασης ενός όρου μέσα στο κείμενο, αν εμφανίζεται στην αρχή του κειμένου ή σε περιοχές που θεωρούνται σημαντικές (τίτλος, πλάγια γράμματα κλπ.), κ.α.

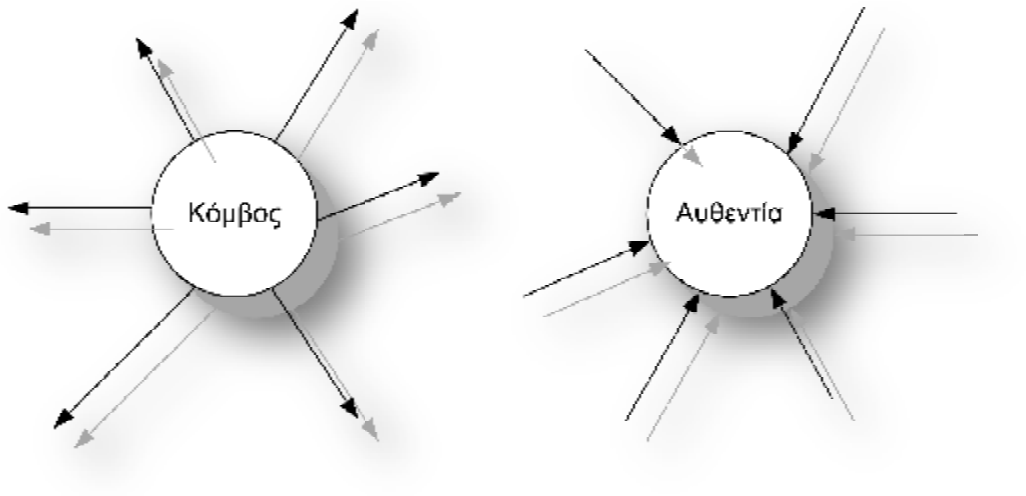
Παρόλα αυτά οι σχεδιαστές των σελίδων προσπάθησαν να εκμεταλλευτούν αυτές τις μεθόδους έτσι ώστε να επιτύχουν υψηλή βαθμολογία για τους δικτυακούς τους τόπους. Αυτό το πετύχαιναν εισάγοντας λέξεις ή φράσεις πολλές φορές μέσα στα κείμενά τους, ακόμα και σε σημεία που δεν ήταν εμφανή. Επομένως έγινε αντιληπτή η επιτακτική ανάγκη να βρεθούν ακόμα πιο έξυπνοι τρόποι για να βαθμολογηθούν οι σελίδες.

Η ιδέα που άλλαξε την αναζήτηση στον παγκόσμιο ιστό ήταν η χρήση της δομής του γραφήματος του παγκόσμιου ιστού, με την έννοια ότι η σημαντικότητα μιας σελίδας είναι ανάλογη του πλήθους των υπερσυνδέσμων που την δείχνουν, και ειδικά όταν οι υπερσύνδεσμοι αυτοί προέρχονται από «σημαντικές» σελίδες. Οι βασικότεροι αλγόριθμοι που στηρίζονται στην ιδέα αυτή είναι ο HITS του Kleinberg [Kleinberg98], ο PageRank των Brin και Page [BP98] και ο Salsa [LC98], τους οποίους και θα αναλύσουμε στο επόμενο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 3ο ΑΛΓΟΡΙΘΜΟΙ ΑΝΑΖΗΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

3.1 **Ο αλγόριθμος** HITS

Ο αλγόριθμος HITS του οποίου το όνομα είναι ακρωνύμιο των λέξεων Hypertext Induced Topic Search, αναπτύχθηκε από τον Jon Kleinberg το 1998 [Kleinberg98], την ίδια εποχή που οι Brin και Page επεξεργάζονταν τον αλγόριθμο Pagerank [BP98]. Έως το 2001 ο HITS δεν ενσωματώθηκε σε καμία μηχανή αναζήτησης, μέχρι που χρησιμοποιήθηκε στην τεχνολογία κατάταξης της νέας μηχανής αναζήτησης Teoma (www.teoma.com) [CS2002], όπου και τον υιοθέτησε ως το βασικό στοιχείο της τεχνολογίας της. Ο αλγόριθμος HITS χρησιμοποιεί τόσο τους εσωτερικούς όσο και τους εξωτερικούς συνδέσμους για να υπολογίσει το βαθμό κατάταξης των σελίδων. Η διαφορά του με τον αλγόριθμο Pagerank (τον οποίο θα αναλύσουμε στην επόμενη ενότητα) είναι ότι ο αλγόριθμος HITS είναι ερωτηματοεξαρτώμενος και υπολογίζει το βαθμό κατάταξης δύο σελίδων κάθε φορά. Ο HITS διακρίνει τις σελίδες σε κόμβους (hubs) και αυθεντίες (authorities), όπου κόμβοι είναι σελίδες με πολλούς εξωτερικούς συνδέσμους και αυθεντίες είναι σελίδες με πολλούς εσωτερικούς συνδέσμους (Εικόνα 4).



Εικόνα 4: Παράδειγμα ενός κόμβου και μιας αυθεντίας

Μια σελίδα μπορεί να είναι ταυτόχρονα και κόμβος και αυθεντία. Οι κόμβοι και οι αυθεντίες ονομάζονται καλοί/καλές όταν ισχύει ότι «οι καλές αυθεντίες δείχνονται από καλούς κόμβους και οι καλοί κόμβοι δείχνονται από καλές αυθεντίες».

3.1.1 **Περιγραφή της τεχνικής**

Κάθε σύνολο από διασυνδεδεμένες σελίδες μπορεί να αναπαρασταθεί σαν ένα κατευθυνόμενο γράφημα $G = (V, E)$, όπου για κάθε σελίδα του συνόλου υπάρχει ένας κόμβος στο γράφημα και για κάθε σύνδεσμο από τη σελίδα p στην q υπάρχει μια κατευθυνόμενη ακμή από τον κόμβο p στον q. Ο βαθμός εξόδου ή έξω-βαθμός (***out-degree***) ενός κόμβου p είναι ο αριθμός των σελίδων προς τις οποίες η σελίδα p έχει σύνδεσμο (οι σελίδες που δείχνει η p) και ο βαθμός εισόδου ή έσω-βαθμός (***in-degree***) ενός κόμβου p είναι ο αριθμός των σελίδων που έχουν σύνδεσμο προς αυτήν (οι σελίδες που δείχνουν την p).

Από το γράφημα G είναι δυνατό να απομονωθεί ένα υπογράφημα ακολουθώντας την παρακάτω διαδικασία. Εάν θεωρήσουμε ότι το W είναι ένα υποσύνολο των σελίδων V του γραφήματος τότε ως $G[W]$ ορίζεται το γράφημα που προκύπτει από αυτό το σύνολο των σελίδων. Έτσι το $G[W]$ περιέχει τόσους κόμβους όσες οι σελίδες του W και ακμές αυτές που προκύπτουν από τους αντίστοιχους συνδέσμους μεταξύ των σελίδων του W.[5]

Ας υποθέσουμε ότι δίνεται ως είσοδος στο σύστημα ένα ερώτημα με ευρύ θέμα καθορισμένο από τον όρο αναζήτησης σ. Καθώς η τεχνική δεν έχει νόημα να εφαρμοστεί σε όλες τις σελίδες του διαδικτύου, αλλά σε ένα μόνο κομμάτι του σχετικό με το θέμα του ερωτήματος, πρώτα πρέπει να επιλεγεί αυτό το υποσύνολο σελίδων του διαδικτύου.

Μια πρώτη ιδέα θα ήταν να επιλεγεί το σύνολο των σελίδων Q_σ που περιέχουν τον όρο του ερωτήματος. Αυτή η μέθοδος έχει όμως δύο σημαντικά μειονεκτήματα. Πρώτον αυτό το σύνολο είναι πολύ πιθανό να περιέχει ένα μεγάλο αριθμό σελίδων και να αυξήσει τόσο το υπολογιστικό κόστος που να είναι ανέφικτη η εκτέλεση του αλγορίθμου και δεύτερον πιθανότατα πολλές από τις αυθεντικές σελίδες να μην περιέχονται σε αυτό το σύνολο. Βάση των παραπάνω προκύπτει το ότι πρέπει να

[5] Ουσιαστικά το $G[W]$ αποτελεί ένα επαγόμενο (induced) υπογράφημα του G.

αποκτηθεί πρώτα ένα σύνολο από σελίδες, το S_σ, το οποίο θα ικανοποιεί τις ακόλουθες απαιτήσεις [Kleinberg98]:

- Να είναι σχετικά μικρό,
- Να είναι πλούσιο σε σχετικές με το θέμα σελίδες, και
- Να περιέχει τις περισσότερες ή έστω πολλές από τις αυθεντικές σελίδες.

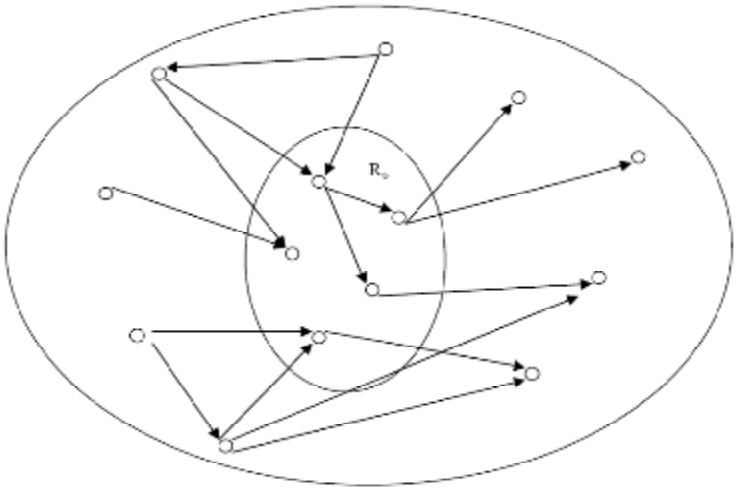
Καθώς το S_σ διατηρείται μικρό σε μέγεθος, το υπολογιστικό κόστος για την εφαρμογή του αλγορίθμου σε αυτό διατηρείται σε σχετικά μικρά μεγέθη. Επίσης εξασφαλίζοντας ότι το σύνολο αυτό είναι πλούσιο σε σχετικές με το θέμα σελίδες γίνεται πιο εύκολη η εύρεση καλών αυθεντιών. Επομένως το πρόβλημα που προκύπτει είναι η εύρεση ενός τέτοιου συνόλου.

,

3.1.2 **Εύρεση βασικού συνόλου σελίδων**

Για να δημιουργηθεί το βασικό σύνολο σελίδων που απαιτείται, αρχικά δημιουργείται ένα αρχικό σύνολο το R_σ, το οποίο να περιέχει τις πρώτες t (για παράδειγμα, έστω $t = 200$) σελίδες που δίνει σαν αποτέλεσμα μια term-based μηχανή αναζήτησης όπως για παράδειγμα η Alta Vista, δίνοντάς της σαν είσοδο τον όρο σ. Αυτό το σύνολο είναι εμφανές ότι ικανοποιεί την πρώτη απαίτηση, καθώς το μέγεθός του μπορεί εύκολα να καθοριστεί από την παράμετρο t. Επίσης ικανοποιείται και η δεύτερη απαίτηση καθώς το R_σ είναι ένα υποσύνολο του Q_σ που είναι η συλλογή όλων των σελίδων που περιέχουν τον όρο σ. Από την άλλη πλευρά όμως, το σύνολο αυτό απέχει πολύ από το να ικανοποιεί και την τρίτη απαίτηση, καθώς ακόμα και το σύνολο Q_σ δεν την ικανοποιεί.

Δεν είναι ιδιαίτερα δύσκολο να καταλήξουμε σε ένα σύνολο S_σ (εικόνα 5), χρησιμοποιώντας το R_σ, το οποίο να ικανοποιεί και την τρίτη απαίτηση. Θεωρώντας ότι μια καλή αυθεντία για το συγκεκριμένο θέμα δεν περιέχεται στο σύνολο R_σ είναι πολύ πιθανό να δείχνεται από τουλάχιστον μία σελίδα του R_σ. Έτσι ο αριθμός των καλών αυθεντιών μπορεί να αυξηθεί επεκτείνοντας το σύνολο R_σ προσθέτοντας τις σελίδες που δείχνουν σε σελίδες αυτού του συνόλου και αυτές που δείχνονται από σελίδες του R_σ [Kleinberg98, Kleinberg00].



Εικόνα 5: Επέκταση του αρχικού συνόλου σελίδων στο βασικό σύνολο

Το σύνολο S_σ τελικά προκύπτει μεγαλώνοντας το αρχικό σύνολο R_σ, περιλαμβάνοντας σε αυτό κάθε σελίδα προς την οποία υπάρχει σύνδεσμος από σελίδα του συνόλου R_σ και κάθε σελίδα από την οποία υπάρχει σύνδεσμος προς κάποια σελίδα του συνόλου R_σ, με την προϋπόθεση ότι μέχρι d[6] σελίδες μπορούν να προστεθούν στο σύνολο S_σ που δείχνουν σε μια μόνο σελίδα του R_σ. Η προϋπόθεση αυτή είναι ιδιαίτερα σημαντική, καθώς μπορεί να υπάρχει ένας πολύ μεγάλος αριθμός σελίδων που περιέχουν διασύνδεση προς μια σελίδα, και είναι εμφανές ότι δεν είναι δυνατό όλες αυτές οι σελίδες να συμπεριληφθούν στο σύνολο S_σ, του οποίου το μέγεθος απαιτείται να είναι σχετικά μικρό.

Από τις σελίδες που ανήκουν στο σύνολο S_σ προκύπτει ένα γράφημα το $G[S_\sigma]$, στο οποίο κόμβοι είναι οι σελίδες και ακμές οι σύνδεσμοι που συνδέουν τις σελίδες του συνόλου. Καθώς υπάρχουν στις σελίδες σύνδεσμοι οι οποίοι δεν μεταφέρουν κάποια σημαντική πληροφορία αλλά απλά διευκολύνουν την πλοήγηση του χρήστη, προτείνεται μια ευρετική μέθοδος, η οποία σκοπό έχει να αντισταθμίσει το αποτέλεσμα των συνδέσμων αυτών. Επομένως οι ακμές που υπάρχουν στο γράφημα $G[S_\sigma]$ χωρίζονται σε δύο κατηγορίες.

Μια ακμή χαρακτηρίζεται εγκάρσια (transverse) εάν συνδέει δύο σελίδες οι οποίες ανήκουν σε διαφορετικό domain, και φυσική (intrinsic) εάν συνδέει δύο

[6] Το d είναι φυσικός αριθμός.

σελίδες που βρίσκονται στο ίδιο domain. Το domain είναι το πρώτο επίπεδο της διεύθυνσης, η οποία σχετίζεται με κάποια σελίδα. Καθώς οι φυσικές ακμές είναι αυτές που διευκολύνουν την πλοήγηση μέσα σε ένα διαδικτυακό κόμβο, προκύπτει ότι μεταφέρουν πολύ λιγότερη πληροφορία για τη σπουδαιότητα και την ποιότητα της σελίδας προς την οποία δείχνουν από ότι οι εγκάρσιες ακμές. Για το λόγο αυτό οι φυσικές ακμές του γραφήματος αφαιρούνται με αποτέλεσμα να μένουν σε αυτόν μόνο οι εγκάρσιες ακμές. Το γράφημα που προκύπτει τελικά είναι το G_σ. Η μέθοδος αυτή της διαγραφής των φυσικών ακμών, είναι μεν ιδιαίτερα απλή, είναι όμως και αποτελεσματική.

3.1.3 **Υπολογισμών των Βαρών των** Hubs **και των** Authorities

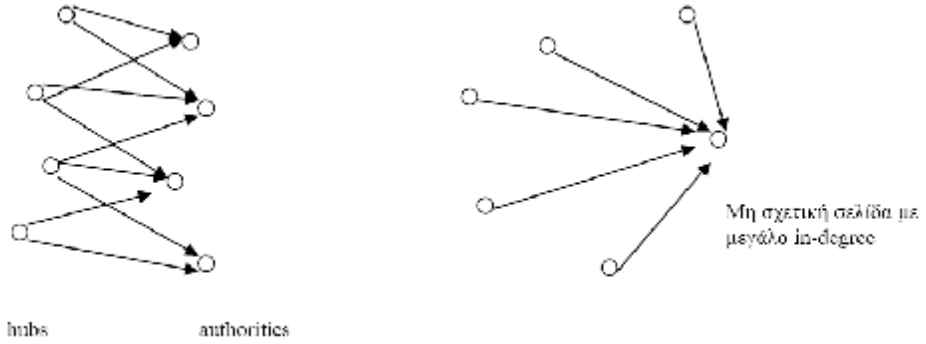
Το γράφημα G_σ που έχει δημιουργηθεί περιέχει πολλές σχετικές με το ερώτημα σελίδες και αρκετές σημαντικές σελίδες. Αυτό που χρειάζεται στη συνέχεια είναι να βρεθούν αυτές οι σημαντικές σελίδες, αναλύοντας τη δομή των ακμών του γραφήματος αυτού.

Μια απλή προσέγγιση είναι η ταξινόμηση των σελίδων βάση του έσω-βαθμού, δηλαδή του αριθμού των σελίδων που δείχνουν στη συγκεκριμένη σελίδα. Η ιδέα αυτή είχε απορριφθεί για το σύνολο όλων των σελίδων που περιέχουν το ερώτημα σ. Αλλά σε αυτή τη φάση το γράφημα που έχει δημιουργηθεί είναι χαρακτηριστικά μικρότερο και περιέχει πολύ περισσότερες σημαντικές σελίδες, προς τις οποίες υπάρχουν πολλές ακμές.

Παρόλο που αυτή η προσέγγιση δίνει καλύτερα αποτελέσματα για το γράφημα από ότι για το σύνολο των σελίδων, εφαρμόζοντάς τη στο γράφημα μπορεί να δημιουργήσει σημαντικά προβλήματα. Αυτό συμβαίνει γιατί δεν διαχωρίζει τις σημαντικές σελίδες, σε σχέση με το ερώτημα, που υπάρχουν στο γράφημα από τις γενικότερα δημοφιλείς σελίδες, καθώς και οι δύο αυτοί τύποι σελίδων έχουν μεγάλο έσω-βαθμό.

Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με την παρατήρηση ότι οι αυθεντικές σελίδες που είναι σχετικές με το ερώτημα σ του χρήστη δεν απαιτεί να έχουν μόνο μεγάλο έσω-βαθμό, αλλά και να έχουν αρκετά κοινά χαρακτηριστικά με τα σύνολα των σελίδων που δείχνουν προς αυτές. Επομένως εκτός από τις αυθεντικές

σελίδες θα πρέπει να προσδιοριστούν και ένα σύνολο άλλων σελίδων, οι λεγόμενες κομβικές σελίδες, οι οποίες έχουν διασυνδέσεις προς τις αυθεντικές σελίδες. Οι σελίδες αυτές συνενώνουν κατά κάποιο τρόπο τις αυθεντίες σε ένα κοινό θέμα, αγνοώντας σελίδες που απλά έχουν μεγάλο έσω-βαθμό. Ένα παράδειγμα αυτής της συνένωσης των αυθεντικών σελίδων από τις κομβικές σελίδες φαίνεται στην εικόνα 6.



Εικόνα 6: Ένα ισχυρά συνδεδεμένο σύνολο σελίδων hubs και authorities

Οι κομβικές και αυθεντικές σελίδες υποδηλώνουν ένα είδος σχέσης αμοιβαίας ενίσχυσης, καθώς όπως ήδη αναφέραμε ένας καλός κόμβος είναι μια σελίδα που δείχνει σε πολλές καλές αυθεντίες και μια καλή αυθεντία είναι μια σελίδα που δείχνεται από πολλούς καλούς κόμβους. Επομένως για να βρεθούν αυτά τα σύνολα σελίδων πρέπει να βρεθεί μια μέθοδος η οποία να μπορεί να ανιχνεύσει αυτή τη σχέση στο σύνολο G_σ.

3.1.4 **Ο επαναληπτικός αλγόριθμος**

Όπως ήδη έχουμε αναφέρει ο αλγόριθμος HITS εξαρτάται από το ερώτημα του χρήστη και διακρίνει τις ιστοσελίδες του σε δύο κατηγορίες (αυθεντίες και κόμβους), οι οποίες είναι αλληλένδετες μεταξύ τους και σε κάθε μία από αυτές τις σελίδες αντιστοιχεί ένα μέτρο (βάρος) αυθεντίας και ένα μέτρο (βάρος) κόμβου.

Ο επαναληπτικός αλγόριθμος που περιγράφουμε στη συνέχεια, και ο οποίος υπολογίζει και ενημερώνει τα βάρη των κόμβων και των αυθεντιών για κάθε σελίδα,

εκμεταλλεύεται αυτή την αμοιβαία σχέση των κομβικών και αυθεντικών σελίδων. Με κάθε σελίδα του γραφήματος συνδέονται δύο μη αρνητικά βάρη, το authority βάρος $x^{<p>}$ και το hub βάρος $y^{<p>}$. Κανονικοποιώντας τα βάρη κάθε τύπου ξεχωριστά έτσι ώστε το άθροισμα των τετραγώνων τους να είναι ίσα με τη μονάδα, παρατηρείται ότι οι σελίδες με τις μεγαλύτερες τιμές για αυτά τα βάρη είναι οι καλύτερες αυθεντίες και κόμβοι αντίστοιχα.

Αριθμητικά αυτή η σχέση αμοιβαίας ενίσχυσης αναπαριστάνεται ως εξής: εάν η σελίδα p δείχνει σε πολλές σελίδες με μεγάλες τιμές για το βάρος x (authority), τότε είναι αναμενόμενο να αποκτήσει μεγάλη τιμή για το βάρος y (hub). Ανάλογα εάν η σελίδα p δείχνεται από πολλές σελίδες με μεγάλες τιμές y είναι αναμενόμενο να αποκτήσει μεγάλη τιμή για το βάρος x. Αυτή η προσέγγιση ωθεί προς τον ορισμό δύο συναρτήσεων που εφαρμόζονται στα βάρη x *και* y, οι οποίες αναφέρονται σαν I και O. Έτσι η συνάρτηση I ενημερώνει τα βάρη x ως εξής:

$$x^{<p>.} \leftarrow \sum_{q:(q,p)eE} y^{<q>}$$

Ενώ η συνάρτηση O ενημερώνει τα βάρη y ως εξής:

$$y^{<p>} \leftarrow \sum_{q:(q,p)eE} x^{<q>}$$

Για να βρεθούν οι τιμές ισορρόπησης (σταθεροποίηση των τιμών) για τα βάρη, αρκεί να εφαρμοστούν οι συναρτήσεις I και O διαδοχικά αρκετές φορές μέχρι οι τιμές να σταθεροποιηθούν. Το σύνολο των βαρών x αναπαριστάνεται με ένα διάνυσμα όπου κάθε συντεταγμένη αντιστοιχεί σε μια σελίδα και αντίστοιχα το σύνολο των βαρών y με ένα άλλο διάνυσμα. Ο αλγόριθμος που καλείται είναι ο ακόλουθος:

Iterate (G, k)

1. *G: a collection of n linked pages*
2. *k: a natural number*
3. *Let z denote the vector (1, 1, 1, .. , 1) Є* R^n .
4. *Set x0 :=z.*
5. *Set y0 :=z.*
6. *For i =1, 2, ..., k*

7. *Apply the I operation to (x_i –1, y_i –1) ,obtaining new -weights x_i' .*
8. *Apply the O operation to (x_i', y_i –1) ,obtaining new y -weights y_i' .*
9. *Normalize x_i', obtaining x_i .*
10. *Normalize y_i', obtaining y_i .*
11. *End*
12. *Return (x_k ,y_k).*

Στην επαναληπτική αυτή διαδικασία διατηρούμε ένα σύνολο G από μια συλλογή από n διασυνδεδεμένες σελίδες και ορίζουμε ως z το διάνυσμα που ανήκει σε ένα σύνολο R. Στην πορεία θέτουμε ως τιμές $x0 = z$ και $y0 = z$ για κάθε i να είναι φυσικός αριθμός και εφαρμόζουμε τη συνάρτηση του I $(x_i - 1\ ,\ y_i - 1)$ και αποκτάμε ένα καινούριο βάρος το x_i'. Εφαρμόζοντας στη συνέχεια τη συνάρτηση του O $(x_i'\ ,\ y_i - 1)$ αποκτάμε ένα καινούριο βάρος y_i'. Ομαλοποιώντας το x_i' και y_i' στη συνάρτηση αποκτάμε δύο καινούρια βάρη x_i και y_i και τελειώνοντας μας επιστρέφει τη νέα συνάρτηση I (x_k, y_k).

Στη συνέχεια εφαρμόζεται μια συνάρτηση φιλτραρίσματος η οποία επιστρέφει τις c καλύτερες κομβικές σελίδες. Ο αλγόριθμος για αυτό το φιλτράρισμα είναι ο ακόλουθος:

Filter (G, k, c)

1. *G: a collection of n linked pages*
2. *k,c: natural numbers*
3. *(x_k, y_k) := Iterate(G, k)*
4. *Report the pages with the c largest coordinates in x_k as authorities*
5. *Report the pages with the c largest coordinates in y_k as hubs*

Έχοντας ως είσοδο τα βάρη της συνάρτησης I (x_k, y_k), ο αλγόριθμος μας επιστρέφει αρχικά τις c καλύτερες σελίδες με τις μεγαλύτερες συντεταγμένες στο x_k ως αυθεντία και τις c καλύτερες σελίδες με τις μεγαλύτερες συντεταγμένες στο y_k ως κόμβους.

Χρησιμοποιώντας προχωρημένες τεχνικές από την θεωρία της Γραμμικής Άλγεβρας αποδεικνύεται ότι όσο αυξάνεται ο αριθμός k, δηλαδή όσο πιο πολλές

φορές εκτελείται ο αλγόριθμος Iterate, τόσο οι τιμές των βαρών τείνουν να σταθεροποιηθούν. Προκύπτει επίσης και μια σημαντική παρατήρηση σε σχέση με τα «τελικά» βάρη. Έστω A, ο πίνακας γειτνίασης του γραφήματος των σελίδων δηλαδή ο πίνακας που προκύπτει θέτοντας την τιμή 1 στη θέση (i, j) αν υπάρχει ακμή στο γράφημα G_σ από τη σελίδα p_i στη σελίδα p_j και την τιμή 0 στις υπόλοιπες θέσεις του πίνακα. Εύκολα μπορεί να δειχτεί ότι οι συναρτήσεις I *και* O χρησιμοποιώντας τον πίνακα A μπορούν να γραφτούν ως εξής:

$$x \leftarrow A^T y \text{ και } y \leftarrow Ax \text{ αντίστοιχα.}$$

Έτσι παρατηρείται ότι το τελικό διάνυσμα x στο οποίο σταθεροποιείται ο αλγόριθμος Iterate είναι το πρωτεύον ιδιοδιάνυσμα του πίνακα $A^T A$ και αντίστοιχα το τελικό διάνυσμα y είναι το πρωτεύον ιδιοδιάνυσμα του πίνακα AA^T.

Μετά από μερικές εκτελέσεις του αλγόριθμου Iterate προκύπτει ότι ο αλγόριθμος συγκλίνει αρκετά γρήγορα στις τελικές τιμές των διανυσμάτων x και y, καθώς αρκούν 20 επαναλήψεις. Από την παρατήρηση που προέκυψε παραπάνω μπορεί κανείς να θεωρήσει ότι αρκεί να βρεθούν τα ιδιοδιανύσματα των ανωτέρω πινάκων για να βρεθούν και οι τελικές τιμές των βαρών. Η εύρεση των ιδιοδιανυσμάτων όμως δεν είναι εύκολη διαδικασία.

Η χρήση του αλγόριθμου Iterate προτιμάται για δύο λόγους. Πρώτον, ο αλγόριθμος αυτός υποδηλώνει την ώθηση σε αυτή την προσέγγιση λόγω της αμοιβαίας ενίσχυσης που προκύπτει από τις συναρτήσεις I και O. Δεύτερον, δεν χρειάζεται να εκτελεστεί ο αλγόριθμος αυτός μέχρι να συγκλίνει, καθώς αρκεί για να υπολογιστούν τα διανύσματα των βαρών να αρχικοποιηθούν και στη συνέχεια να εφαρμοστεί ένας καθορισμένος μικρός αριθμός διαδοχικών επαναλήψεων των συναρτήσεων I και O.

3.2 **Ο Αλγόριθμος** Pagerank

Το 1998, δυο φοιτητές οι Larry Page και Sergey Brin, δημιουργοί της μηχανής αναζήτησης Google (www.google.com), κατά τη διάρκεια ενός νέου ερευνητικού προγράμματος για ένα νέο είδος μηχανής αναζήτησης στο πανεπιστήμιο Stanford, δημιούργησαν τον αλγόριθμο Pagerank [BP98, PBMW98]. Το όνομα του οφείλεται

στον έναν από τους δύο δημιουργούς του, τον Larry Page[7]. Είναι σύνθεση δύο αγγλικών λέξεων «page» και «rank» οι οποίες μεταφρασμένες στην ελληνική γλώσσα σημαίνουν «σελίδα» και «κατάταξη» ή «βαθμός», αντίστοιχα. Ο αλγόριθμος Pagerank υπήρξε το βασικό συστατικό για την επιτυχία της Google, καθώς αποτέλεσε μια καινοτόμα ιδέα από την αρχή της λειτουργίας της [Franceschet11]. Είναι μια από τις καλύτερες μεθόδους που χρησιμοποιεί η Google για να αποφασίσει ποιες σελίδες είναι πιο σημαντικές.

Η βασική ιδέα του αλγόριθμου Pagerank είναι να προσδιορίσει, μέσω μιας μαθηματικής μεθόδου, τη σπουδαιότητα μιας σελίδας με βάση τις σημαντικές σελίδες που δείχνουν σε αυτήν. Πιο συγκεκριμένα ένας χρήστης ξεκινώντας την πλοήγησή του στον παγκόσμιο ιστό, αρχίζοντας από μια τυχαία σελίδα και μέσω της βοήθειας διαφόρων συνδέσμων καταλήγει σε μια συγκεκριμένη σελίδα. Οι σύνδεσμοι ουσιαστικά αποτελούν την ψήφο εμπιστοσύνης για μια σελίδα και γι' αυτό το λόγο όσοι περισσότεροι σύνδεσμοι υποδεικνύουν μια σελίδα τόσο μεγαλύτερη πιθανότητα έχει ο χρήστης να την ανακαλύψει. Η Google χρησιμοποιώντας τον αλγόριθμο Pagerank μπορεί να μετράει αυτούς τους ψήφους εμπιστοσύνης και να αξιολογεί κατά πόσο σημαντική είναι η σελίδα [Cormick10].

Ο αλγόριθμος Pagerank εκτός από την ποσότητα των συνδέσμων που δείχνουν σε μια σελίδα αξιολογεί επίσης και την επισκεψιμότητά τους, δηλαδή το πόσο συχνά την επισκέπτονται οι χρήστες. Ορισμένες σελίδες όμως μπορεί να έχουν μεγάλη επισκεψιμότητα αλλά όχι και συνδέσμους προς αυτήν, οι σελίδες αυτές θα έχουν χαμηλό βαθμό κατάταξης. Για παράδειγμα, μια σελίδα πορνογραφικού υλικού ενώ έχει μεγάλη επισκεψιμότητα, δεν έχει μεγάλο βαθμό κατάταξης καθώς δεν έχει πολλές σελίδες που δείχνουν σε αυτήν, διότι κανείς δεν θα ήθελε να βάλει έναν τέτοιον σύνδεσμο στη δική του προσωπική σελίδα. Έτσι ο αλγόριθμος Pagerank δίνει μεγαλύτερη βαρύτητα στην ποιότητα των εξωτερικών συνδέσμων (δηλαδή στον έξω-βαθμό) και των εσωτερικών συνδέσμων (δηλαδή στον έσω-βαθμό) της σελίδας. Για παράδειγμα, αν μια σελίδα που είναι σχετική με βιβλία έχει σύνδεσμο από μια σημαντική σελίδα όπως η Wikipedia (www.wikipedia.org) και μια άλλη διαφορετική σελίδα έχει συνδέσμους από πολλές άλλες απλές σελίδες, τότε η σελίδα η οποία

[7] Για ένα σύντομο βιογραφικό σημείωμα και πληροφορίες σχετικά με την επιστημονική και επαγγελματική πορεία των ιδρυτών της Google, μπορείτε να ανατρέξετε στους συνδέσμους http://en.wikipedia.org/wiki/Sergey_Brin και http://en.wikipedia.org/wiki/Larry_Page, αντίστοιχα.

δείχνεται από τη Wikipedia είναι σαφώς καλύτερη και λαμβάνει περισσότερη βαρύτητα σε σχέση με την άλλη σελίδα. Αυτή είναι και η λειτουργία του αλγόριθμου Pagerank, αναλύει την σπουδαιότητα των συνδέσμων. Επομένως μια σελίδα θα έχει υψηλό βαθμό κατάταξης αν ο βαθμός κατάταξης των υπόλοιπων σελίδων που συνδέουν προς αυτήν είναι υψηλός.

Ο βαθμός κατάταξης μιας σελίδας παίρνει τιμές από 0 έως 10. Όσο μεγαλύτερο βαθμό κατάταξης έχει μια σελίδα, τόσο πιο ψηλά θα φτάσει στα αποτελέσματα αναζήτησης του χρήστη [LM2007]. Σελίδες με βαθμό κατάταξης 0 είναι νέες ή σελίδες χωρίς συνδέσμους, ενώ σελίδες με βαθμό κατάταξης 10 είναι λίγες [8], όπως για παράδειγμα της Google (www.google.com), του Facebook (www.facebook.com) και του CNN (edition.cnn.com). Μια σελίδα που έχει βαθμό κατάταξης 3 και 50 συνδέσεις δεν σημαίνει πως αν έχει 100 συνδέσεις θα αποκτήσει βαθμό κατάταξης 6. Για να αποκτήσει μεγαλύτερο βαθμό κατάταξης θα πρέπει να έχει πολύ περισσότερες συνδέσεις. Ο βαθμός κατάταξης των σελίδων δεν μένει σταθερός, αλλά συχνά μεταβάλλεται, γι' αυτό και η Google ανανεώνει τις τιμές των σελίδων κάθε 3 με 4 μήνες.

Παλαιότερα οι μηχανές αναζήτησης αξιολογούσαν μια σελίδα βασιζόμενες στις λέξεις-κλειδιά και τη συχνότητα που αυτές εμφανίζονταν, και έτσι μπορούσαν εύκολα οι search engine spammers (άτομα που παραπλανούν τις μηχανές αναζήτησης) να τις ξεγελάσουν. Αυτοί χρησιμοποιούσαν τις λέξεις-κλειδιά στη σελίδα τους πολλές φορές με αποτέλεσμα η σελίδα να φτάνει πολύ ψηλά στα αποτελέσματα της αναζήτησης, χωρίς το θέμα τους να είναι σχετικό με αυτό που αναζητούσε ο χρήστης. Με τον αλγόριθμο Pagerank αυτό το πρόβλημα λύνεται καθώς όπως ήδη αναφέραμε ασχολείται με τους συνδέσμους και όχι τόσο με τις λέξεις-κλειδιά.

Ένα άλλο πρόβλημα που έχει δημιουργηθεί είναι το λεγόμενο web spamming (ανεπιθύμητες ιστοσελίδες). Ορισμένα άτομα μπορούν να παραποιήσουν τη λειτουργία των συνδέσμων για να επηρεάσουν την κατάταξη των δικών τους σελίδων. Δημιουργούν δηλαδή ένα μεγάλο αριθμό σελίδων οι οποίες έχουν σύνδεσμο στη δική τους σελίδα, με αποτέλεσμα να φτάνει ψηλά στην κατάταξη χωρίς να το

[8] Θα αναλύσουμε τον τρόπο εύρεσης του βαθμού κατάταξης μιας σελίδας σε επόμενη ενότητα.

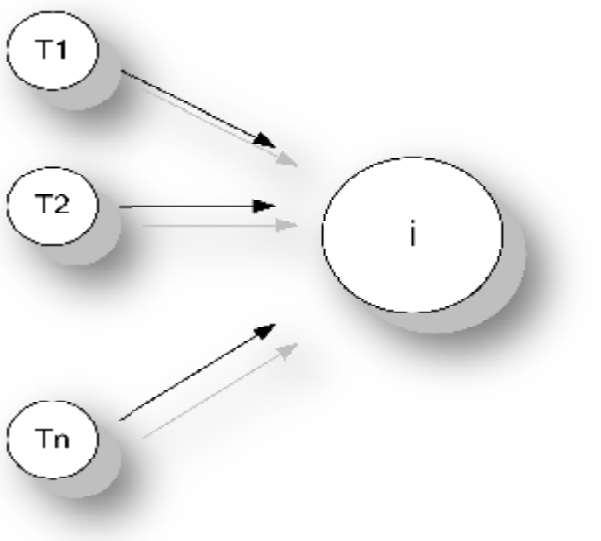
αξίζει. Έτσι αν η σελίδα τους ασχολείται για παράδειγμα με ηλεκτρονικούς υπολογιστές καταφέρνει να έχει και περισσότερες πωλήσεις.

Υπάρχουν αρκετοί τρόποι για την αντιμετώπιση του προβλήματος του web spamming. Ο πιο διαδεδομένος είναι ο αλγόριθμος Trustrank [ZGMP04], ο οποίος είναι συμπληρωματικός του αλγόριθμου Pagerank και χρησιμοποιείται στην μηχανή αναζήτησης Google. Ο Trustrank δίνει πληροφορίες για το αν μια σελίδα είναι έγκυρη και αν χρησιμοποιεί μεθόδους spam. Λειτουργεί σαν ένας ελεγκτής των σελίδων όπου ελέγχει όλα τα αρχεία τα οποία είναι αποθηκευμένα στη σελίδα. Τα τελευταία χρόνια ο αλγόριθμος Trustrank έχει γίνει αρκετά γνωστός, με αποτέλεσμα πολλοί ιστότοποι να δίνουν περισσότερη σημασία στο να ανεβάσουν τη βαθμολογία που τους δίνει ο αλγόριθμος Trustrank παρά ο αλγόριθμος Pagerank [Λαμπρόπουλος11].

3.2.1 **Υπολογισμός του βαθμού κατάταξης μιας σελίδας**

Οι δημιουργοί του αλγόριθμου Pagerank, Sergey Brin και Larry Page, ξεκίνησαν με μια απλή εξίσωση άθροισης για να υπολογίσουν το βαθμό κατάταξης μιας σελίδας i. Έστω ότι έχουμε ένα πλήθος σελίδων $T_1, T_2, \dots, T_n$ (Σχήμα 2), όπου όλες αυτές οι σελίδες δείχνουν την σελίδα i. Ο βαθμός κατάταξης της σελίδας i ισούται με το άθροισμα των βαθμών κατάταξης των σελίδων $T_1, T_2, \dots, T_n$, σύμφωνα με τον τύπο:

$$PR(i) = PR(T_1) + PR(T_2) + \cdots + PR(T_n) \qquad (3.1)$$



Σχήμα 2: Υπολογισμός του βαθμού κατάταξης της σελίδας i

Στο σχήμα 2 παρατηρούμε ότι οι σελίδες $T_1, T_2, \dots, T_n$ έχουν μόνο έναν εξωτερικό σύνδεσμο προς τη σελίδα i. Γενικότερα, αν $C(T_1), C(T_2), \dots, C(T_n)$ είναι το πλήθος των εξωτερικών συνδέσμων (έξω-βαθμός) των σελίδων $T_1, T_2, \dots, T_n$, αντίστοιχα, τότε για τον υπολογισμό του βαθμού κατάταξης της σελίδας i , διαιρούμε κάθε όρο του προηγούμενου αθροίσματος (1) με τον αντίστοιχο έξω-βαθμό και προκύπτει έτσι ο παρακάτω γενικός τύπος υπολογισμού του βαθμού κατάταξης:

$$\mathrm{PR(i)} = \frac{\mathrm{PR(T_1)}}{\mathrm{C(T_1)}} + \frac{\mathrm{PR(T_2)}}{\mathrm{C(T_2)}} + \dots + \frac{\mathrm{PR(T_n)}}{\mathrm{C(T_n)}} \qquad (3.2)$$

Παρατηρήστε ότι ο τύπος (1) προκύπτει από την (2) αν θέσουμε $C(T_1) = C(T_2) = \cdots = C(T_n) = 1$.

Η θεωρία του αλγόριθμου Pagerank υποστηρίζει ότι ένας τυχαίος χρήστης ο οποίος κάνοντας κλικ στους συνδέσμους θα σταματήσει τελικά σε μια σελίδα [BP98]. Η πιθανότητα ο χρήστης να συνεχίσει την πλοήγηση του ακολουθώντας τους συνδέσμους είναι ο συντελεστής απόσβεσης d που ορίζεται συνήθως ίσος με 0,85, ενώ $(1 - d)$ είναι η πιθανότητα μετάβασης του τυχαίου χρήστη σε μια νέα σελίδα, μέσω της εισαγωγής ενός καινούριου προορισμού URL στη γραμμή διευθύνσεων του φυλλομετρητή, πιθανώς άσχετη με την τρέχουσα σελίδα [Franceschet11]. Έτσι η εξίσωση υπολογισμού του βαθμού κατάταξης διαμορφώνεται ως εξής:

$$PR(i) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} \dots + \frac{PR(T_n)}{C(T_n)}\right) \quad (3.3)$$

όπου:

- $PR(i)$ είναι ο βαθμός κατάταξης της σελίδας i
- $T_1, T_2, \dots, T_n$ είναι οι σελίδες που δείχνουν στη σελίδα i
- $PR(T_1), \dots, PR(T_n)$ είναι οι βαθμοί κατάταξης των $T_1, \dots, T_n$, αντίστοιχα
- $C(T_1), \dots, C(T_n)$ είναι το πλήθος των εξερχόμενων συνδέσμων που περιέχουν οι $T_1, \dots, T_n$, αντίστοιχα
- $d \in (0 \dots 1)$ είναι ο συντελεστής απόσβεσης (η πιθανότητα ο χρήστης να βαρεθεί την πλοήγηση στη σελίδα που βρίσκεται και να μεταβεί σε μια οποιαδήποτε άλλη), συνήθως ίσος με 0,85.

Για κάθε σελίδα i ο βαθμός κατάταξής της συνυπολογίζεται από τους όρους $PR\ (T_i)$ / $C\ (T_i)$. Δηλαδή όσες περισσότερες σελίδες δείχνουν στη σελίδα i και όσο μεγαλύτερο βαθμό κατάταξης $PR(T_i)$ έχουν αυτές οι σελίδες, τόσο μεγαλύτερο βαθμό κατάταξης θα αποκτήσει η σελίδα i [BP98, Craven, Sobek02]. Όμως όσους περισσότερους εξερχόμενους συνδέσμους έχει η κάθε σελίδα $C(T_i)$ τόσο λιγότερο θα βοηθήσει τη σελίδα i. Αυτό συμβαίνει γιατί όταν μια σελίδα δείχνει σε μια άλλη της δίνει την ψήφο εμπιστοσύνης της, αν όμως δείχνει σε περισσότερες από μία σελίδες θα πρέπει ουσιαστικά να μοιράσει την ψήφο εμπιστοσύνης της και στις υπόλοιπες σελίδες.

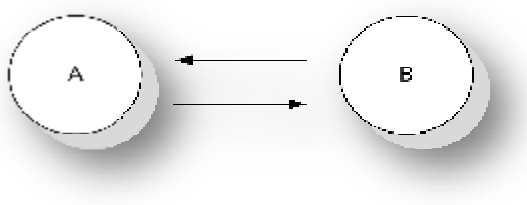
Το βαθμό κατάταξης των σελίδων που δείχνουν σε μια σελίδα i δεν μπορούμε να τον γνωρίζουμε πάντοτε μέχρι να τον υπολογίσουμε και αυτό είναι ένα από τα βασικά προβλήματα της εξίσωσης (3.3). Προκειμένου να παρακάμψουν αυτό το πρόβλημα οι Brin και Page χρησιμοποίησαν μια επαναληπτική διεργασία: Ας θεωρήσουμε ότι αρχικά όλες οι σελίδες έχουν τον ίδιο βαθμό κατάταξης (ίσο με $1/n$ όπου n είναι ο αριθμός των σελίδων στο ευρετήριο του παγκόσμιου ιστού). Εφαρμόζοντας την εξίσωση (3.3) με αυτή την παραδοχή υπολογίζουμε τον βαθμό κατάταξης $PR(i)$ για κάθε σελίδα i του ευρετηρίου. Κάθε φορά που θα κάνουμε την επανάληψη χρησιμοποιούμε τους βαθμούς κατάταξης $PR(Ti)$ που υπολογίστηκαν κατά την προηγούμενη επανάληψη. Για να ορίσουμε αυτήν την επαναληπτική διαδικασία θα εισαγάγουμε κάποιους επιπλέον συμβολισμούς: Έστω ότι $PR_{k+1}(i)$

είναι ο βαθμός κατάταξης της σελίδας i κατά την $(k + 1)$ -οστή επανάληψη [LM2007]. Τότε:

$$\mathrm{PR_{k+1}(i)} = (1 - \mathrm{d}) + \mathrm{d}\left(\frac{\mathrm{PR_k(T_1)}}{\mathrm{C(T_1)}} + \frac{\mathrm{PR_k(T_2)}}{\mathrm{C(T_2)}} \ldots + \frac{\mathrm{PR_k(T_n)}}{\mathrm{C(T_n)}}\right) \qquad (3.4)$$

Η επαναληπτική διαδικασία ξεκινά με $PR_0(i) = \frac{1}{n}$ για κάθε σελίδα i και επαναλαμβάνεται συνεχώς μέχρι οι βαθμοί κατάταξης να συγκλίνουν σε κάποιες σταθερές τελικές τιμές. Για να γίνει κατανοητή, δίνουμε στη συνέχεια μερικά παραδείγματα:

Παράδειγμα 1



Σχήμα 3: Παράδειγμα υπολογισμού βαθμού κατάταξης

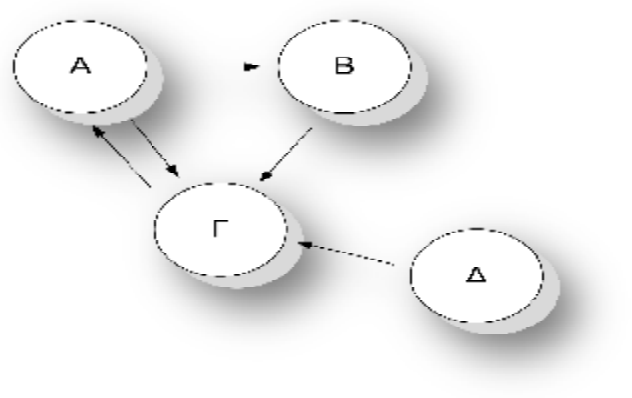
Στο σχήμα 3 έχουμε δύο σελίδες, την A και την B, οι οποίες έχουν σύνδεσμο η μια στην άλλη. Δεν γνωρίζουμε το βαθμό κατάταξής αυτών, οπότε υποθέτουμε ότι είναι ίσος με 1. Στη συνέχεια, εφαρμόζουμε τον τύπο (3.3):

$$PR(A) = (1 - d) + d\left(\frac{PR(B)}{C(B)}\right) = (1 - 0{,}85) + 0{,}85\,\frac{1}{1} = 0{,}15 + 0{,}85 = 1$$

$$PR(B) = (1 - d) + d\left(\frac{PR(A)}{C(A)}\right) = (1 - 0{,}85) + 0{,}85\frac{1}{1} = 0{,}15 + 0{,}85 = 1$$

Είδαμε ένα απλό παράδειγμα για τον υπολογισμό του βαθμού κατάταξης κάθε σελίδας. Στο επόμενο παράδειγμα θα δούμε και τη μέθοδο της επανάληψης η οποία μας δίνει πιο ακριβή αποτελέσματα σχετικά με τον υπολογισμό των σελίδων.

Παράδειγμα 2



Σχήμα 4: Παράδειγμα υπολογισμού βαθμού κατάταξης

Στο σχήμα 4 έχουμε τέσσερις σελίδες A, B, Γ και Δ. Η σελίδα A έχει σύνδεσμο στις σελίδες B και Γ, η σελίδα B έχει σύνδεσμο στη σελίδα Γ, η σελίδα Γ στην A και η σελίδα Δ στη Γ. Αρχικά υποθέτουμε ότι οι σελίδες έχουν αρχικό βαθμό κατάταξης $\frac{1}{n} = \frac{1}{4} = 0{,}25$. Στη συνέχεια εφαρμόζουμε τον τύπο (3.4), για κάθε σελίδα:

$$PR(A) = (1-d) + d\left(\frac{PR(\Gamma)}{C(\Gamma)}\right) = (1-0{,}85) + 0{,}85\frac{0{,}25}{1} = 0{,}15 + 0{,}2125$$
$$= 0{,}3625$$

$$PR(B) = (1-d) + d\left(\frac{PR(A)}{C(A)}\right) = (1-0{,}85) + 0{,}85\frac{0{,}25}{2} = 0{,}15 + 0{,}10625$$
$$= 0{,}25625$$

$$PR(\Gamma) = (1-d) + d\left(\frac{PR(A)}{C(A)} + \frac{PR(B)}{C(B)} + \frac{PR(\Delta)}{C(\Delta)}\right)$$
$$= (1-0{,}85) + 0{,}85\left(\frac{0{,}25}{2} + \frac{0{,}25}{1} + \frac{0{,}25}{1}\right) = 0{,}15 + 0{,}53125$$
$$= 0{,}68125$$

$$PR(\Delta) = (1-d) + d * 0 = (1-0{,}85) + 0{,}85 * 0 = 0{,}15 + 0 = 0{,}15$$

Οι παραπάνω τιμές αφορούν στην πρώτη επανάληψη, Στην επόμενη επανάληψη, χρησιμοποιούμε τις τιμές που προέκυψαν, αντικαθιστώντας τις προηγούμενες:

$$PR(A) = 0{,}15 + 0{,}85 * 0{,}68125 = 0{,}7290625$$

$$PR(B) = 0{,}15 + 0{,}85 * (0{,}3625/2) = 0{,}3040625$$

$$PR(\Gamma) = 0{,}15 + 0{,}85 * (0{,}3625/2 + 0{,}25625 + 0{,}15) = 0{,}649375$$

$$PR(\Delta) = 0{,}15 + 0{,}85 * 0 = 0{,}15$$

Συνεχίζοντας με τον ίδιο τρόπο, προκύπτει ο ακόλουθος πίνακας:

Επανάληψη	PR(A)	PR(B)	PR(Γ)	PR(Δ)
0	0,25	0,25	0,25	0,25
1	0,3625	0,25625	0,68125	0,15
2	0,7290625	0,3040625	0,649375	0,15
3	0,7019687	0,4598515	0,8458046	0,15
4	0,8689339	0,4483366	0,9667104	0,15
5	0,9717038	0,5192968	1,0278829	0,15
6	1,0237004	0,5629741	1,1318763	0,15
7	1,1120948	0,5850726	1,1911006	0,15
8	1,1624355	0,6226402	1,247452	0,15
9	1,2103342	0,644035	1,3007792	0,15
10	1,2556623	0,664392	1,3393217	0,15

Παρατηρούμε ότι ο βαθμός κατάταξης των σελίδων αυξάνεται συνεχώς (τείνοντας σε κάποια σταθερά) εκτός από τη σελίδα Δ η οποία δεν έχει εσωτερικούς συνδέσμους [Sobek02].

3.2.2 **Βαθμολόγηση των σελίδων χρησιμοποιώντας τον αλγόριθμο** PageRank

Με τις εξισώσεις (3.3) και (3.4) υπολογίζουμε το βαθμό κατάταξης μιας μόνο σελίδας κάθε φορά. Χρησιμοποιώντας όμως πίνακες (μήτρες) αποφεύγουμε αυτή τη διαδικασία και μπορούμε να υπολογίζουμε σε κάθε επανάληψη το *διάνυσμα βαθμών κατάταξης*, δηλαδή ένα απλό $1 \times n$ διάνυσμα το οποίο θα φέρει το βαθμό κατάταξης όλων των σελίδων ενός ευρετηρίου του παγκόσμιου ιστού [LM2007]. Για παράδειγμα για να υπολογίσουμε τους κόμβους (σελίδες) του κατευθυνόμενου γραφήματος του σχήματος 1 του πρώτου κεφαλαίου, θα χρησιμοποιήσουμε έναν τετραγωνικό πίνακα (square matrix) $n \times n$ ο οποίος λέγεται *πίνακας υπερσυνδέσμων* (hyperlink matrix) και συμβολίζεται με H και ένα $1 \times n$ διάνυσμα γραμμής π^T. Ο πίνακας H είναι ένας γραμμοκανονικοποιημένος (κανονικοποιημένος κατά γραμμές) πίνακας υπερσυνδέσμων, όπου το στοιχείο H_{ij} ισούται με $1/|P_i|$ αν υπάρχει σύνδεσμος από τον κόμβο i προς τον κόμβο j, και με 0 σε αντίθετη περίπτωση [Franceschet11].

Ο τετραγωνικός πίνακας H για το γράφημα του σχήματος 1 θα είναι ο εξής:

$$H = \begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Στον πίνακα H τα μη μηδενικά στοιχεία της γραμμής i είναι οι εξωτερικοί σύνδεσμοι της σελίδας i, ενώ τα μη μηδενικά στοιχεία της στήλης j είναι οι εσωτερικοί σύνδεσμοι της σελίδας i. Στη συνέχεια εισάγουμε το διάνυσμα γραμμής $\pi^{(\kappa)T}$, που αποτελεί το διάνυσμα των βαθμών κατάταξης κατά την κ-οστή επανάληψη. Μέσω αυτού του συμβολισμού η εξίσωση (3.4) μπορεί να γραφεί στην πιο συνεπτυγμένη μορφή:

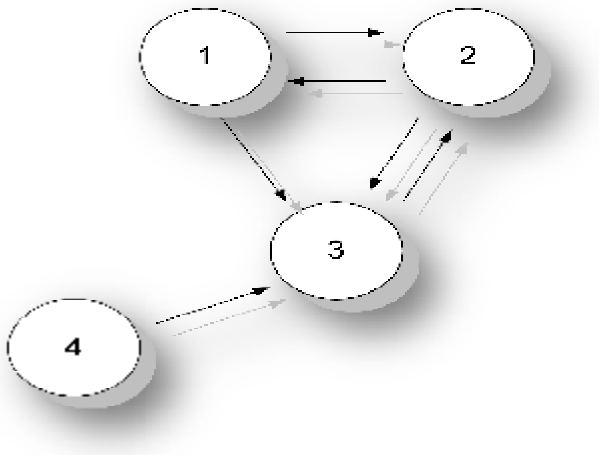
$$\pi^{(\kappa+1)T} = \pi^{(\kappa)T}H \qquad (3.5)$$

Από τη μελέτη της εξίσωσης (3.5) παρατηρούμε ότι σε κάθε επανάληψη εκτελείται πολλαπλασιασμός ενός διανύσματος με ένα πίνακα. Ο πίνακας *H* μοιάζει πολύ με στοχαστικό πίνακα πιθανοτήτων μετάβασης μιας αλυσίδας Markov (δηλαδή ένας μη αρνητικός πίνακας όπου το άθροισμα των στοιχείων κάθε γραμμής είναι ίσο με 1). Παρατηρούμε όμως ότι ο κόμβος 5 δεν έχει συνδέσεις σε άλλες σελίδες, είναι ένας κόμβος αδιεξόδου (dangling node) και γι' αυτό όλες οι εγγραφές στην γραμμή 5 θα είναι 0, άρα η γραμμή 5 δεν είναι στοχαστική. Όλες οι υπόλοιπες γραμμές που δεν αντιστοιχούν σε κόμβους αδιεξόδου είναι στοχαστικές. Έτσι ο πίνακας *H* ονομάζεται υποστοχαστικός. Επιπλέον ο *H* είναι πολύ αραιός πίνακας (μεγάλο ποσοστό των στοιχείων του είναι 0), διότι οι περισσότερες σελίδες υποδεικνύουν ελάχιστες άλλες σελίδες. Οι αραιοί πίνακες είναι θετικό να υπάρχουν καθώς απαιτούν ελάχιστο αποθηκευτικό χώρο, γιατί υπάρχουν κατάλληλοι τρόποι αποθήκευσής τους, όπου καταχωρίζονται μόνο τα μη μηδενικά στοιχεία και η θέση τους. Ο πολλαπλασιασμός ενός διανύσματος με έναν αραιό πίνακα έχει μικρότερες απαιτήσεις από ότι ο αντίστοιχος υπολογισμός με έναν πυκνό πίνακα [LM2007]. Οι παρατηρήσεις αυτές είναι σημαντικές για την ανάπτυξη και εφαρμογή του αλγόριθμου Pagerank.

3.2.2.1 **Τα προβλήματα της επαναληπτικής διεργασίας**

Με την εξίσωση (3.5) προκύπτει μια σειρά ερωτημάτων, όπως αν η επαναληπτική διεργασία θα συγκλίνει ή θα συνεχίζεται επ' άπειρον, αν θα συγκλίνει σε ένα μοναδικό διάνυσμα ή σε πολλά, και αν εξαρτάται η σύγκλιση από το εναρκτήριο διάνυσμα $\pi^{(0)T}$. Και αν τελικά συγκλίνει σε πόσο χρόνο θα γίνει αυτό, σε πόσες δηλαδή επαναλήψεις. Τα ερωτήματα αυτά απασχόλησαν τους δημιουργούς του Pagerank, τα οποία και αντιμετώπισαν, αρχικά ξεκινώντας την επαναληπτική διαδικασία θέτοντας $\pi^{(0)T} = \frac{1}{n}e^T$ όπου e^T είναι το διάνυσμα γραμμής με όλα τα στοιχεία να είναι ίσα με το 1, αλλά και με πιο εξελιγμένες τεχνικές που περιγράφουμε παρακάτω.

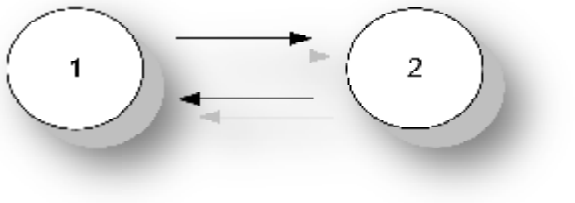
Ένα σημαντικό πρόβλημα είναι οι καταβόθρες τάξης, δηλαδή σελίδες οι οποίες συγκεντρώνουν όλο και μεγαλύτερο βαθμό κατάταξης σε κάθε επανάληψη, χωρίς να «μοιράζουν» τη βαθμολογία τους σε άλλες σελίδες.



Σχήμα 5: Γράφημα με σελίδες καταβόθρες

Για παράδειγμα, στο σχήμα 5, οι κόμβοι 1, 2 και 3 αποτελούν καταβόθρες τάξης, καθώς «συνωμοτούν» για να συγκεντρώσουν μεγάλο βαθμό κατάταξης χωρίς να τον «μοιράζονται» με τον κόμβο 4. Το παράδειγμα αυτό φανερώνει ένα ακόμα πρόβλημα που δημιουργούν οι καταβόθρες, αφού κάποιοι κόμβοι μένουν χωρίς βαθμολογία και όταν οι περισσότεροι κόμβοι ισοβαθμούν με βαθμολογία 0 η κατάταξη τους με βάση τη βαθμολογία είναι δύσκολη.

Το δεύτερο πρόβλημα είναι οι κύκλοι που δημιουργούνται όταν μια σελίδα δείχνει μόνο μια άλλη σελίδα και αντιστρόφως.



Σχήμα 6: Παράδειγμα γραφήματος με κύκλο

Αν υποθέσουμε ότι η επαναληπτική διεργασία της εξίσωσης (3.3) ξεκινάει με $\pi^{(0)T} = (1 \quad 0)$, τότε δεν πρόκειται να υπάρχει σύγκλιση σε όσο χρόνο και αν συνεχίσει να εκτελείται, καθώς τα διαδοχικά διανύσματα $\pi^{(\kappa)T}$ θα είναι συνεχώς μεταξύ του $(1 \quad 0)$, όταν το κ είναι άρτιο και $(0 \quad 1)$ όταν το κ είναι περιττό.

3.2.2.2 **Οι αρχικές προσαρμογές στο βασικό μοντέλο**

Μέσω της θεωρίας Markov[9] μπορούμε να τροποποιήσουμε την εξίσωση (3.5) έτσι ώστε να εξασφαλίσουμε επιθυμητά αποτελέσματα, καλές ιδιότητες σύγκλισης και ενθαρρυντικές απαντήσεις στους προβληματισμούς που αναφέραμε. Αν για οποιοδήποτε εναρκτήριο διάνυσμα εφαρμόσουμε τη δυναμομέθοδο (η οποία θα αναλυθεί σε επόμενη ενότητα) σε έναν πίνακα Markov τότε αυτή θα συγκλίνει σε ένα μοναδικό θετικό διάνυσμα που ονομάζεται στάσιμο διάνυσμα, υπό την προϋπόθεση ότι ο πίνακας είναι στοχαστικός, μη αναγώγιμος [10] και απεριοδικός [11] . Η απεριοδικότητα σε συνδυασμό με τη μη αναγωγιμότητα συνεπάγεται πρωταρχικότητα[12]. Έτσι τα προβλήματα σύγκλισης που δημιουργούν οι καταβόθρες τάξης και οι κύκλοι στον αλγόριθμο Pagerank μπορούν να αντιμετωπιστούν, αν ο αρχικός πίνακας *H* τροποποιηθεί ελαφρά ώστε να πάρει τη μορφή ενός πίνακα Markov.

Οι προσαρμογές που έκαναν στο βασικό μοντέλο οι Brin και Page, περιγράφονται στα πρώτα τους άρθρα που δημοσιεύτηκαν το 1998, αλλά πουθενά δεν αναφέρεται η φράση «αλυσίδα Markov». Για να περιγράψουν αυτές τις προσαρμογές χρησιμοποίησαν τον όρο του *αδιάφορου περιηγητή*, ο οποίος είναι ένας χρήστης που μεταπηδά τυχαία από τη μια σελίδα στην άλλη ακολουθώντας τους συνδέσμους του ιστού. Όταν φτάσει σε μια σελίδα με πολλούς εξωτερικούς συνδέσμους επιλέγει έναν σύνδεσμο στην τύχη και μεταβαίνει στην σελίδα που δείχνει αυτός. Αυτή η διεργασία συνεχίζεται επ' αόριστον. Ο χρόνος που αφιερώνει ο χρήστης σε μια συγκεκριμένη σελίδα αποτελεί ένα μέτρο της σχετικής σπουδαιότητας της σελίδας. Αν αφιερώνει πολύ χρόνο σε μια συγκεκριμένη σελίδα σημαίνει ότι ακολουθώντας τους συνδέσμους του ιστού τυχαία θα επανέλθει πάλι σε αυτή τη σελίδα. Οι σελίδες στις οποίες επανέρχεται συχνά ο χρήστης θα πρέπει να είναι σημαντικές αφού θα υποδεικνύονται από άλλες σημαντικές σελίδες. Ο αδιάφορος περιηγητής όμως

[9] Η θεωρία Markov ορίζει ότι η στοχαστική διεργασία είναι «αμνήμων», δηλαδή ότι η κατάσταση της αλυσίδας την επόμενη χρονική στιγμή εξαρτάται μόνο από την τρέχουσα κατάστασή της και όχι από το προηγούμενο ιστορικό της αλυσίδας. Για περισσότερες πληροφορίες της θεωρίας Markov, ενδεικτικά αναφέρουμε το [BR2010].

[10] Μη αναγώγιμος καλείται ένας πίνακας αν και μόνο αν το αντίστοιχο κατευθυντικό γράφημά είναι ισχυρά συνεκτικό.

[11] Η απεριοδικότητα οφείλεται στους βρόχους του γραφήματος.

[12] Πρωταρχικός ονομάζεται ένας πίνακας αν είναι μη αναγώγιμος και έχει ένα τουλάχιστον θετικό διαγώνιο στοιχείο.

αντιμετωπίζει ορισμένα προβλήματα. Όπως όταν επισκέπτεται έναν αδιέξοδο κόμβο, παγιδεύεται. Ο παγκόσμιος ιστός αποτελείται από πολλούς τέτοιους κόμβους, όπως εικόνες, αρχεία pdf, πίνακες δεδομένων κ.α., οι οποίοι παγιδεύουν τους χρήστες.

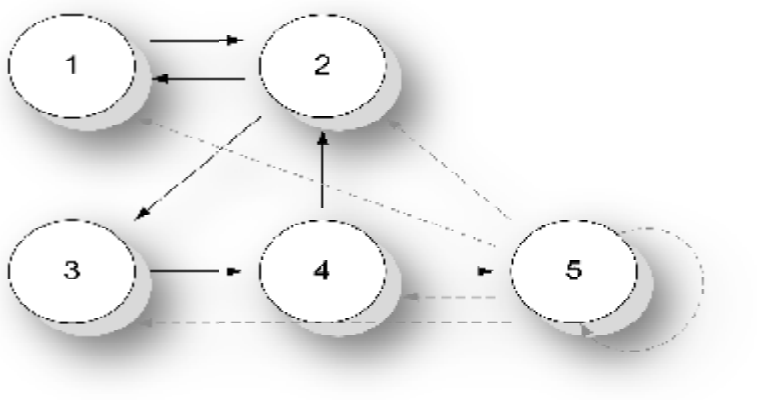
Για να διορθώσουν αυτό το πρόβλημα οι Brin και Page εισήγαγαν την πρώτη τους προσαρμογή, που ονόμασαν *προσαρμογή στοχαστικότητας*, σύμφωνα με την οποία οι γραμμές του πίνακα H που ισούνται με 0^T αντικαθίστανται με το $\frac{1}{n}e^T$, έτσι ώστε ο H να γίνει στοχαστικός. Με αυτόν τον τρόπο όταν ο χρήστης εισέλθει σε έναν αδιέξοδο κόμβο, μπορεί να μεταβεί με τυχαίο τρόπο σε οποιαδήποτε άλλη σελίδα. Έτσι ο νέος πίνακας του σχήματος 1 που προκύπτει είναι ο εξής:

$$S = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

Αν γράψουμε την προσαρμογή στοχαστικότητας σε μαθηματική μορφή, θα προκύψει ο πίνακας S όπου είναι ο τροποποιημένος πίνακας H:

$$\mathrm{S} = \mathrm{H} + \mathrm{a}(\frac{1}{\mathrm{n}}\mathrm{e}^{\mathrm{T}})$$

όπου το a_i ισούται με 1 αν η σελίδα i είναι αδιέξοδος κόμβος και με 0 σε αντίθετη περίπτωση. Το a είναι ένα δυαδικό διάνυσμα που ονομάζεται διάνυσμα αδιέξοδων κόμβων. Σύμφωνα με αυτό, για το σχήμα 1, με την τροποποίηση του πίνακα H θα έχουμε το παρακάτω σχήμα:



Σχήμα 7: Το νέο γράφημα μετά τη διόρθωση του αδιέξοδου κόμβου 5

Παρατηρούμε ότι στο νέο αυτό γράφημα γίνεται ανακύκλωση του κόμβου 5, δηλαδή ξεκινάει και καταλήγει πάλι στον κόμβο 5. Η σύνδεση αυτή ουσιαστικά δίνει τη δυνατότητα στον χρήστη να εκτελέσει τη λειτουργία “refresh” (ανανέωση) για όσο θα βρίσκεται στη σελίδα αυτή.

Η προσέγγιση αυτή εξασφαλίζει ότι ο S είναι στοχαστικός και αποτελεί τον πίνακα πιθανοτήτων μετάβασης μιας αλυσίδας Markov, όμως δεν αρκεί από μόνη της για να εξασφαλίσει τις επιθυμητές ιδιότητες σύγκλισης, δηλ. ότι υπάρχει ένα μοναδικό θετικό διάνυσμα π^T και ότι η εξίσωση (3.5) συγκλίνει γρήγορα σε αυτό. Έτσι οι Brin και Page χρειάστηκε να κάνουν και μια δεύτερη προσαρμογή, όπου ονόμασαν *προσαρμογή πρωταρχικότητας*. Με αυτή την προσαρμογή ο πίνακας που θα προκύψει δεν θα είναι μόνο στοχαστικός αλλά και πρωταρχικός.

Οι Brin και Page περιέγραψαν αυτές τις ιδιότητες και πάλι μέσω του αδιάφορου περιηγητή, ο οποίος ακολουθώντας τους συνδέσμους του παγκόσμιου ιστού μπορεί να βαρεθεί τη μέθοδο των συνδέσμων και να εισάγει κατευθείαν μια διεύθυνση στη γραμμή διευθύνσεων του φυλλομετρητή [LM2007]. Έτσι κάθε φορά που συμβαίνει αυτό ο χρήστης τηλεμεταφέρεται στη νέα σελίδα όπου και συνεχίζει να περιηγείται μέσω συνδέσμων μέχρι την επόμενη τηλεμεταφορά του. Για να εκφράσουν τη συμπεριφορά αυτή του αδιάφορου περιηγητή με μαθηματικό τρόπο κατασκεύασαν έναν νέο πίνακα:

$$G = dS + (1 - d)E$$

Ο πίνακας G ονομάζεται πίνακας Google. Ο συντελεστής απόσβεσης d είναι ένας αριθμός ανάμεσα στο 0 και το 1 και εκφράζει το ποσοστό του χρόνου όπου ο χρήστης ακολουθεί τους συνδέσμους, έναντι της τηλεμεταφοράς. Για παράδειγμα αν το $d = 0{,}7$ τότε κατά το 70% του χρόνου ο χρήστης θα ακολουθήσει τους συνδέσμους του παγκόσμιου ιστού και 30% τηλεμεταφέρεται με τυχαίο τρόπο σε νέες σελίδες. Η τηλεμεταφορά γίνεται τυχαία καθώς ο πίνακας τηλεμεταφοράς $E = \frac{1}{n}ee^T$ είναι ομοιόμορφος, που σημαίνει ότι κάθε φορά που ο χρήστης τηλεμεταφέρεται έχει την ίδια πιθανότητα να βρεθεί σε οποιαδήποτε σελίδα.

Έτσι ο πίνακας G με την προσαρμογή αυτή γίνεται:

- Στοχαστικός, δεν έχει αδιέξοδους κόμβους
- Μη αναγώγιμος, καθώς κάθε κόμβος συνδέεται με όλες τις άλλες σελίδες.
- Πρωταρχικός, διότι υπάρχει ακέραιος k ώστε $G^k > 0$. Υπάρχει δηλαδή ένα μοναδικό θετικό διάνυσμα π^T, και αν η δυναμομέθοδος εφαρμοστεί στον G θα συγκλίνει σίγουρα σε αυτό το διάνυσμα
- Απεριοδικός, εφόσον όπως είδαμε ο πίνακας είναι πρωταρχικός και μη αναγώγιμος
- Απολύτως πυκνός, αν και δυσάρεστο από υπολογιστικής πλευράς
- Τεχνητός, αφού έχει προκύψει από τη διπλή τροποποίηση του αρχικού πίνακα Η με σκοπό να επιτευχθούν οι επιθυμητές ιδιότητες σύγκλισης.

Έτσι, η προσαρμοσμένη μέθοδος Pagerank της Google εκφράζεται από τη σχέση:

$$\pi^{(\kappa+1)\mathrm{T}} = \pi^{(\kappa)\mathrm{T}}\mathrm{G} \qquad (3.6)$$

Που ουσιαστικά είναι η δυναμομέθοδος (περιγράφεται στην επόμενη ενότητα) εφαρμοσμένη στον πίνακα Google.

Για να ολοκληρώσουμε το παράδειγμά του σχήματος 1, θα φτιάξουμε τον στοχαστικό, πρωταρχικό πίνακα Google σύμφωνα με τον τύπο:

$$G = dS + (1-d)E = d\left(H + \frac{1}{n}\mathrm{a}e^T\right) + (1-d)\frac{1}{n}ee^T$$

$$= dH + (d\mathrm{a} + (1-\mathrm{d})\mathrm{e})\frac{1}{n}e^T$$

Θέτοντας $d = 0{,}85$, ο πίνακας G είναι:

$$G = 0{,}85H + (0{,}85\begin{pmatrix}0\\0\\0\\0\\1\end{pmatrix} + 0{,}15\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix})\frac{1}{5}(1 \quad 1 \quad 1 \quad 1 \quad 1)$$

$$G = \begin{pmatrix}\frac{3}{100} & \frac{88}{100} & \frac{3}{100} & \frac{3}{100} & \frac{3}{100}\\ \frac{27}{60} & \frac{3}{100} & \frac{27}{60} & \frac{3}{100} & \frac{3}{100}\\ \frac{3}{100} & \frac{3}{100} & \frac{3}{100} & \frac{27}{60} & \frac{3}{100}\\ \frac{3}{100} & \frac{27}{60} & \frac{3}{100} & \frac{3}{100} & \frac{27}{60}\\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5}\end{pmatrix}.$$

3.2.2.3 **Υπολογισμός του διανύσματος** Pagerank

Για μικρούς πίνακες (όπως στο παράδειγμά μας) ο υπολογισμός του διανύσματος Pagerank είναι μια εύκολη και μη απαιτητική διαδικασία από άποψη υπολογιστικού κόστους. Για έναν πίνακα όμως του πραγματικού παγκόσμιου ιστού, όπου έχει περισσότερες από 25 δισεκατομμύρια γραμμές και στήλες, χρειάζεται υπερβολικά μεγάλη υπολογιστική ισχύ και χρόνο για να υπολογιστεί ο βαθμός κατάταξής τους, κάτι το οποίο δεν είναι εφικτό [LM2007]. Έτσι οι Brin και Page πρότειναν τη δυναμομέθοδο για να προσεγγίσουν το διάνυσμα Pagerank με σχετικά μεγάλη ακρίβεια.

Η δυναμομέθοδος είναι μια από τις παλαιότερες και απλές επαναληπτικές μεθόδους που χρησιμοποιείται για την εύρεση επικρατών ιδιοτιμών και ιδιοδιανυσμάτων ενός πίνακα. Η μέθοδος αυτή είναι όμως πολύ αργή, αλλά εξαιρετικά απλή. Η υλοποίηση της και ο απαιτούμενος προγραμματισμός της δεν παρουσιάζουν καμία δυσκολία. Αν η δυναμομέθοδος εφαρμοστεί στον πίνακα G (εξίσωση 3.6) μπορεί να εκφραστεί μέσω του πολύ αραιού πίνακα H:

$$\pi^{(\kappa+1)\mathrm{T}} = \pi^{(\kappa)\mathrm{T}}\mathrm{G}$$

$$= \mathrm{d}\pi^{(\kappa)\mathrm{T}}\mathrm{S} + \frac{1-\mathrm{d}}{\mathrm{n}}\pi^{(\kappa)\mathrm{T}}\mathrm{e}\mathrm{e}^{\mathrm{T}} = \mathrm{d}\pi^{(\kappa)\mathrm{T}}\mathrm{H} + (\mathrm{d}\pi^{(\kappa)\mathrm{T}}\mathrm{a} + 1 - \mathrm{d})\frac{\mathrm{e}^{\mathrm{T}}}{\mathrm{n}} \qquad (3.7)$$

Η εξίσωση αυτή υπολογίζεται συνεχώς μέχρι να εκπληρωθεί κάποιο κριτήριο σύγκλισης (δηλ. μέχρι τα διανύσματα $\pi^{(\kappa)T}$ και $\pi^{(\kappa+1)T}$ να διαφέρουν το πολύ κατά

0,01 ανά στοιχείο μεταξύ τους). Οι πολλαπλασιασμοί μεταξύ διανύσματος και πίνακα $\pi^{(\kappa)T}H$ αφορούν τον αραιό πίνακα H. Οι πίνακες S και G δεν χρειάζεται να υπολογιστούν ούτε να αποθηκευτούν ποτέ. Το μόνο που απαιτείται είναι τα διανύσματα a και e από τα οποία σχηματίζονται οι πίνακες αυτοί.

Ο λόγος όπου η δυναμομέθοδος παραμένει μέχρι και σήμερα η επικρατέστερη μέθοδος επανάληψης είναι γιατί είναι μια μέθοδος «άνευ πίνακα» (ο όρος αυτός αναφέρεται στην αποθήκευση και τη διαχείριση του πίνακα συντελεστών), όπου ο πίνακας συντελεστών δεν υφίσταται καμία επεξεργασία αλλά απλώς προσπελαύνεται από ένα πρόγραμμα για τον υπολογισμό διανύσματος με πίνακα. Η δυναμομέθοδος δεν έχει μεγάλες απαιτήσεις όσον αφορά τον αποθηκευτικό χώρο, εκτός από τον πίνακα H και το διάνυσμα αδιέξοδων κόμβων a, χρειάζεται η αποθήκευση ενός μόνο διανύσματος, του τρέχοντος της επανάληψης $\pi^{(\kappa)T}$. Το διάνυσμα αυτό είναι απολύτως πυκνό που σημαίνει ότι θα πρέπει να αποθηκεύονται n πραγματικοί αριθμοί. Στην περίπτωση της Google ο αριθμός n ισούται με 8,1 δισεκατομμύρια. Άλλες επαναληπτικές μέθοδοι, αν και ταχύτερες απαιτούν την αποθήκευση πολλών διανυσμάτων όπως οι GMRES [SS86] και BICGSTAB [Vorst92]. Για παράδειγμα, η μέθοδος GMRES απαιτεί την αποθήκευση 10 διανυσμάτων μήκους n σε κάθε επανάληψη, μέγεθος το οποίο ισοδυναμεί με τον χώρο που χρειάζεται για όλο τον πίνακα H. Η δυναμομέθοδος για τον υπολογισμό του διανύσματος Pagerank σύμφωνα με τους Brin και Page απαιτεί 50-100 δυναμό-επαναλήψεις [LM2007].

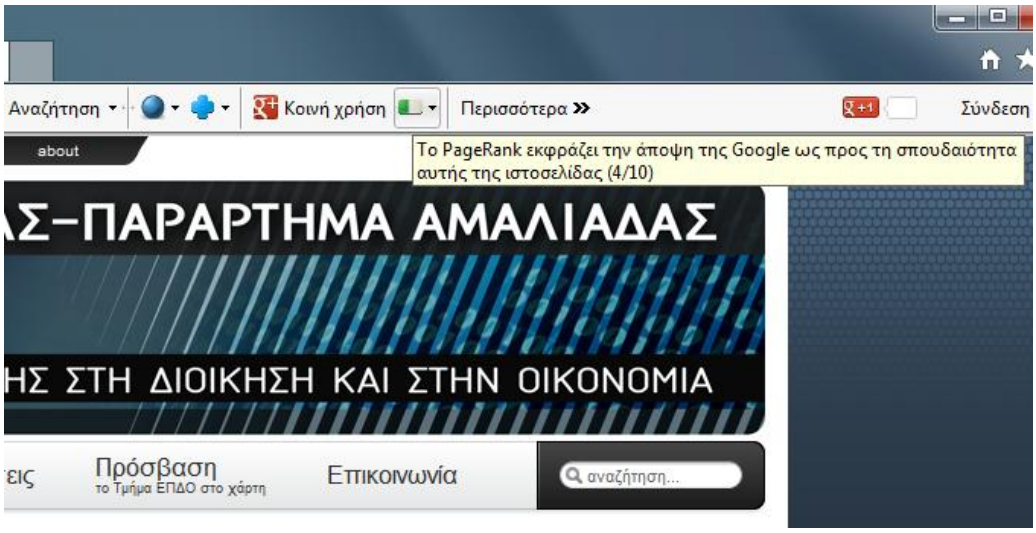
Με τις προσαρμογές της στοχαστικότητας και πρωταρχικότητας έχουμε θετικά αποτελέσματα στα προβλήματα των προηγούμενων ενοτήτων. Με την εφαρμογή της δυναμομεθόδου στον πίνακα G, ανεξάρτητα από το εναρκτήριο διάνυσμα, η διαδικασία συγκλίνει σε ένα μοναδικό θετικό διάνυσμα που ονομάζεται διάνυσμα Pagerank. Επειδή το διάνυσμα που προκύπτει είναι θετικό, δεν θα έχουμε ανεπιθύμητες ισοβαθμίες στο 0.

3.2.3 **Πως θα δούμε το βαθμό κατάταξης μιας σελίδας**

Για να δει κάποιος την τιμή του βαθμού κατάταξης μιας σελίδας υπάρχουν πολλοί τρόποι. Ένας απλός και συνηθισμένος τρόπος είναι με την εγκατάσταση της

μπάρας εργαλείων της Google (Google Toolbar)[13]. Σε αυτήν, πέρα από τις διάφορες ρυθμίσεις και δυνατότητες που προσφέρει, εμφανίζεται μια μπάρα μέτρησης. Αφού ολοκληρωθεί η φόρτωση μιας σελίδας τότε η μπάρα γεμίζει με πράσινο χρώμα Μετακινώντας το ποντίκι πάνω στην μπάρα αυτή εμφανίζεται ο βαθμός κατάταξης της σελίδας. Ένας άλλος τρόπος είναι μέσο της ιστοσελίδας PageRank Checker (www.prchecker.info/check_page_rank.php). Μια απλή αναζήτηση στην Google με λέξη κλειδί «how to view pagerank» θα εμφανίσει πολλές σελίδες που υπολογίζουν το βαθμό κατάταξης.

Στην εικόνα 7 βλέπουμε τον βαθμό κατάταξης που δίνει η Google για την ιστοσελίδα του Τμήματος Εφαρμογών Πληροφορικής στη Διοίκηση και στην Οικονομία του ΤΕΙ Δυτικής Ελλάδας.



Εικόνα 7: Ο βαθμός κατάταξης της Google για την ιστοσελίδα www.amaliada.teipat.gr

3.3 **Ο αλγόριθμος** Salsa

Ο αλγόριθμος Salsa [LC98, LM2000] είναι ακρώνυμο της φράσης Stochastic Approach to Link Structure Analysis (στοχαστική προσέγγιση στη συνδεσμοδομική ανάλυση), χρησιμοποιεί ιδέες τόσο από τον αλγόριθμο HITS όπως και από τον αλγόριθμο Pagerank για να δημιουργήσουν μια νέα μέθοδο κατάταξης ιστοσελίδων.

[13] Η μπάρα εργαλείων της Google (www.google.com/toolbar/ie/index.html) είναι διαθέσιμη μόνο για τους φυλλομετρητές Internet Explorer της Microsoft και Chrome της Google.

Ο αλγόριθμος αυτός υπολογίζει τους κόμβους και τις αυθεντίες, όπως ο αλγόριθμος HITS, οι οποίες όμως προσδιορίζονται μέσω αλυσίδων Markov, όπως ο αλγόριθμος Pagerank.

Για να γίνει πιο κατανοητή η παραπάνω αναφορά, ας θεωρήσουμε μια συλλογή σελίδων η οποία περιέχει τις ακόλουθες δύο κοινότητες. Η κοινότητα *y* αποτελείται από ένα μικρό αριθμό σελίδων κόμβων και αυθεντιών, αλλά κάθε κόμβος της δείχνει στις περισσότερες αυθεντίες της. Η κοινότητα *z* αποτελείται από ένα μεγάλο αριθμό σελίδων στην οποία όμως κάθε κομβική σελίδα δείχνει σε λιγότερες αυθεντικές σελίδες από ότι οι αντίστοιχες του συνόλου *y*. Το βασικό θέμα της συλλογής αυτών των σελίδων εμφανίζεται στην κοινότητα *z* και πιθανότατα ενδιαφέρει περισσότερο τους χρήστες. Καθώς όμως υπάρχουν πολλές αυθεντικές σελίδες στην κοινότητα *z* και οι κομβικές σελίδες της κοινότητας αυτής δεν δείχνουν σε πολλές από αυτές, και ενώ η κοινότητα *y* είναι ισχυρά διασυνδεδεμένη τελικά οι σελίδες της κοινότητας *y* θα αξιολογηθούν καλύτερα από ότι αυτές της *z*. Έτσι τελικά δεν θα εμφανιστεί το κεντρικό θέμα της αρχικές συλλογής των σελίδων.

Συνδυάζοντας την θεωρία των τυχαίων περιπάτων και την έννοια των δύο διαφορετικών τύπων σελίδων του διαδικτύου, κόμβοι και αυθεντίες, αναλύονται τελικά δύο αλυσίδες Markov μια για κάθε τύπο σελίδων. Σε αντίθεση με τους συμβατικούς τυχαίους περιπάτους σε γραφήματα, οι αλλαγές της κατάστασης σε αυτές τις αλυσίδες δημιουργούνται διασχίζοντας δύο διασυνδέσεις κάθε φορά, είτε πρώτα μια προς τα εμπρός διασύνδεση και μετά μια προς τα πίσω ή το αντίστροφο. Αναλύοντας τις δύο αλυσίδες που προκύπτουν δίνεται η δυνατότητα να δοθούν σε κάθε σελίδα του γραφήματος δύο βάρη, ένα κόμβος και μια αυθεντία. Για πιο αναλυτική περιγραφή του αλγορίθμου, ο αναγνώστης προτείνεται να ανατρέξει στις βιβλιογραφικές αναφορές [LC98, LM2000].

ΚΕΦΑΛΑΙΟ 4ο **Η ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ** GOOGLE

4.1 **Η ιστορία της** Google

Οι δημιουργοί της μηχανής αναζήτησης Google όπως ήδη αναφέραμε ήταν οι Larry Page και Sergey Brin, απόφοιτοι του τμήματος υπολογιστών στο πανεπιστήμιο Stanford, όπου και συναντήθηκαν για πρώτη φορά το 1995 [K2008]. Ο Larry ήταν 22 ετών και ο Sergey ένα χρόνο μικρότερος. Το 1996 ξεκίνησαν τη συνεργασία τους για τη δημιουργία της μηχανής αναζήτησης BackRub. Το όνομα της προερχόταν από την ιδιότητα της μηχανής αυτής να αναλύει «προς τα πίσω» τους συνδέσμους που έδειχναν προς κάποια σελίδα. Το 1997 ο Larry και ο Sergey αποφάσισαν να αλλάξουν το όνομα της μηχανής αναζήτησης BackRub. Μετά από πολλή σκέψη κατέληξαν στο όνομα Google, ένα λογοπαίγνιο με τη λέξη "googol", η οποία είναι ένας μαθηματικός όρος για τον αριθμό 10^{100} δηλαδή τον αριθμό που ξεκινάει με το ψηφίο 1 ακολουθούμενο από 100 μηδενικά. Η χρήση του όρου υποδηλώνει το στόχο τους, ο οποίος είναι η οργάνωση ενός φαινομενικά άπειρου όγκου πληροφοριών στον ιστό. Ο διαδικτυακός τόπος της Google ήταν αρχικά στο πανεπιστήμιο Stanford και χρησιμοποιούσε το URL google.stanford.edu. Το Σεπτέμβριο του 1997 κατοχυρώθηκε η διεύθυνση www.google.com όπου παραμένει έως σήμερα.

Το 1998 αναζητώντας χρηματοδότηση, έδειξαν τη μηχανή τους στον Andy Bechtolsheim έναν από τους ιδρυτές της Sun Microsystems, όπου εξέδωσε μια επιταγή 100.000 δολαρίων με δικαιούχο την ανύπαρκτη ακόμη εταιρία Google Inc. Το πρόβλημα της ανύπαρκτης ακόμα εταιρίας που εμπόδιζε την εξαργύρωση της επιταγής, ξεπεράστηκε όταν οι νεαροί απόφοιτοι κατόρθωσαν να συγκεντρώσουν συνολικά ένα εκατομμύριο δολάρια από φίλους, γνωστούς και συγγενείς, Έτσι η Google Inc εγκαταστάθηκε στο γκαράζ της Susan Wojcicki στη διεύθυνση 232 Santa Margarita, Menlo Park. Από τη δοκιμαστική λειτουργία της μηχανής αναζήτησης Google, γίνονταν 10.000 αναζητήσεις κατά μέσο όρο που ολοένα και αυξάνονταν. Το Σεπτέμβριο του 1999 τελείωσε η δοκιμαστική λειτουργία της Google και έκανε το επίσημο ξεκίνημά της. Από τότε συνέχισε να αναπτύσσεται τεχνολογικά και να προσελκύει όλο και περισσότερους χρήστες καθώς έβρισκαν την πληροφορία που αναζητούσαν.

Η Google θεωρείται σήμερα η μεγαλύτερη μηχανή αναζήτησης. Σύμφωνα με την υπηρεσία διαδικτυακών στατιστικών και αναλύσεων HitsLink (www.hitslink.com), το Δεκέμβριο του 2011 στη μηχανή αναζήτησης Google λάμβανε μέρος στο 80% των αναζητήσεων παγκοσμίως. Ο κύριος λόγος για τη δημοτικότητά της είναι η χρήση του αλγόριθμου Pagerank, τον οποίο και αναλύσαμε διεξοδικά στο 2^{ο} κεφάλαιο, όπου αναλύει τους συνδέσμους και κατατάσσει τις σελίδες με βάση τους όρους αναζήτησης.

4.2 **Λόγοι επιτυχίας της** Google

Η μηχανή αναζήτησης Google δραστηριοποιείται στις νέες τεχνολογίες, δίνοντας βαρύτητα σε δύο κατευθύνσεις. Στη δυνατότητα καλύτερης εύρεσης μέσα στη χαοτική παράθεση πληροφοριών και σελίδων στον παγκόσμιο ιστό και στη δυνατότητα προβολής εταιριών και ιδιωτών με το καλύτερο δυνατό αποτέλεσμα στο χαμηλότερο δυνατό κόστος. Οι λόγοι επιτυχίας της είναι οι ακόλουθοι:

- Η Google κατάφερε σε σχέση με όλες τις άλλες μηχανές αναζήτησης να δίνει καλύτερη δυναμική στην εύρεση, να μειώσει την άσκοπη υπέρ-πληροφόρηση και τέλος να βελτιώσει τα ποσοστά επιτυχίας σε σχέση με την αξιοπιστία των πληροφοριών τις οποίες βρίσκει ο χρήστης.
- Οι χρήστες μπορούν να βρίσκουν εύκολα και χωρίς κόπο αυτό που ψάχνουν, έχοντας άμεση πρόσβαση σε εκατομμύρια σελίδες.
- Η βάση της στρατηγικής της αναφέρεται στη διαφοροποίηση με εστίαση στην ποιότητα. Λόγω της καινοτομίας του κλάδου η διαφοροποίηση είναι αναγκαία, πράγμα που πέτυχε η εταιρία καθιστώντας τον εαυτό της ως αρχηγό και αναγκάζοντας τις άλλες εταιρίες να την ακολουθούν. Από την άλλη ως ποιότητα για το κλάδο μεταφράζεται η ευκολία προσβασιμότητας των χρηστών και οι έξτρα παροχές, οι οποίες θα τον διευκολύνουν στην καθημερινή του ζωή [Παπαδάκης07].
- Προώθησε στην αγορά πρώτη τη μέθοδο αξιολόγησης σελίδων Pagerank, η οποία βοηθά από τη μια τους χρήστες να παίρνουν πάντα τα καλύτερα αποτελέσματα και από την άλλη τους διαφημιζόμενους να προβάλλονται σε σχετικές με το αντικείμενο τους ομάδες χρηστών.
- Η Google προώθησε πρώτη την αποτελεσματική διαφήμιση, δίνοντας τη

δυνατότητα στο χρήστη να βλέπει μόνο διαφημίσεις που τον ενδιαφέρουν, αλλά και στους διαφημιζόμενους να αυξήσουν τη κερδοφορία τους.

- Η Google έχοντας την παρούσα στιγμή 112 γραφεία σε όλο τον κόσμο, έχει τη δυνατότητα και την ευκαιρία να επεκτείνεται με γρήγορους ρυθμούς και να έχει πελάτες σ' όλα τα μήκη και τα πλάτη του πλανήτη. Αυτό την καθιστά δελεαστική για τους διαφημιζόμενους και τη βοηθά να αυξάνει συνεχώς τη κερδοφορία της.
- Η εταιρία εστιάζει στο ομαδικό μάνατζμεντ, έχοντας με αυτό τον τρόπο σε συνεχή εγρήγορση το προσωπικό της και παίρνοντας από αυτό το καλύτερο δυνατό αποτέλεσμα.
- Η εταιρία έχει ισχυρή κουλτούρα, ενώ ακόμα και η εσωτερική διαρρύθμιση των γραφείων (εικόνα 8) δίνει τη δυνατότητα στην εύκολη διάχυση των πληροφοριών στο εσωτερικό της, αλλά και στην καλύτερη συνεργασία μεταξύ των στελεχών της. Η συνεργασία γεννά ιδέες και οι ιδέες κάνουν συνεχώς όλο και πιο αποτελεσματική την εταιρία στην αγορά.

Εικόνα 8: Τα γραφεία της Google

Σε επίπεδο διοίκησης ανθρωπίνων πόρων μπορούμε να διαπιστώσουμε τα εξής ισχυρά σημεία της εταιρίας [QS98]:

- Τα γραφεία στο εσωτερικό της εταιρίας, είναι πολύ κοντά μεταξύ τους (εικόνα 9), με αποτέλεσμα να υπάρχει καλύτερη επικοινωνία και μεταξύ των εργαζόμενων στην ίδια ομάδα αλλά και με άλλες ομάδες.

- Το σκεπτικό 70/20/10 δηλαδή 70% παραγωγή, 20% σκέψη και 10% δημιουργία, έχει βοηθήσει την εταιρία να έχει ένα σκεπτόμενο προσωπικό το οποίο μπορεί και καταφέρνει να επιφέρει στην εταιρία το καλύτερο δυνατό αποτέλεσμα βοηθώντας την να διαφοροποιείται.
- Η εταιρία επενδύει στη δημιουργικότητα.
- Υπάρχει ένα καλά σχεδιασμένο σύστημα προσλήψεων, το οποίο βοηθά την εταιρία να επιλέγει πάντα τα καλύτερα και πιο αποτελεσματικά στελέχη της αγοράς.
- Η εταιρία δε μένει στάσιμη μελετά συνεχώς την αγορά και αναπτύσσει νέες τεχνολογίες.

Στην εικόνα 9 βλέπουμε ότι, εκτός από τα τυπικά γραφεία της, η Google έχει διαμορφώσει εσωτερικά χώρους αναψυχής, όπως ήσυχα δωμάτια, στους οποίους οι εργαζόμενοι μπορούν ακόμα και να κοιμηθούν.

Εικόνα 9: Τα γραφεία της Google εσωτερικά

4.3 **Η αρχιτεκτονική του συστήματος της** Google

Η μηχανή αναζήτησης Google σε γενικές γραμμές ακολουθεί την αρχιτεκτονική που παρουσιάσαμε στο 1^{ο} κεφάλαιο. Η Google εκτελείται σε ένα περιβάλλον από χιλιάδες υπολογιστές μικρού κόστους όπου και μπορεί να εκτελεί γρήγορα και συγχρόνως πολλούς υπολογισμούς. Έτσι με αυτόν τον τρόπο μπορεί να επεξεργαστεί παράλληλα ένα μεγάλο όγκο δεδομένων, αφού οι υπολογισμοί εκτελούνται την ίδια χρονική στιγμή [LM2007]. Το λογισμικό της μηχανής

αναζήτησης Google αποτελείται από τα εξής κύρια μέρη τα οποία θα δούμε και πιο αναλυτικά παρακάτω [BP2007]:

- Το Googlebot, το οποίο είναι ένας Crawler που ανιχνεύει και αποθηκεύει τις σελίδες του παγκόσμιου ιστού.
- Το λογισμικό ευρετηριοποίησης (Google Indexer), το οποίο ταξινομεί κάθε λέξη σε κάθε σελίδα και αποθηκεύει τα αποτελέσματα σε ένα ευρετήριο λέξεων σε μια τεράστια βάση δεδομένων.
- Τον επεξεργαστή ερωτήματος (Query Processor), ο οποίος συγκρίνει το ερώτημα του χρήστη με τα δεδομένα που είναι αποθηκευμένα στο ευρετήριο και δίνει ως αποτέλεσμα τα έγγραφα τα οποία θεωρεί ότι είναι πιο σχετικά.

4.3.1 Googlebot (Google's Crawler)

Το Googlebot χρησιμοποιείται για την εύρεση και αποθήκευση των σελίδων που υπάρχουν στον παγκόσμιο ιστό, όπου μετά παραδίδονται στον Google Indexer. Το Googlebot επισκέπτεται τις σελίδες με πολύ πιο αργό ρυθμό από ότι θα μπορούσε και αυτό για να μην καταναλώνει πολλούς πόρους για την ανανέωση μιας σελίδας. Όταν το Googlebot επισκέπτεται μια σελίδα βρίσκει τους συνδέσμους που δείχνουν σε αυτήν και τις προσθέτει σε μια ουρά αναμονής για να τις εξετάσει αργότερα. Η λειτουργία του είναι πολύ απλή και είναι προγραμματισμένο έτσι ώστε να αντιμετωπίζει διάφορες δυσκολίες που παρουσιάζονται. Όπως για παράδειγμα κάποιες σελίδες εμφανίζονται πολλές φορές μέσα στην ουρά προτεραιότητας για μελλοντική εξέταση του Googlebot ή τα δεδομένα τους να έχουν ήδη ευρετηριοποιηθεί στο παρελθόν. Επίσης ένα άλλο πρόβλημα είναι το κάθε πότε το Googlebot πρέπει να επισκέπτεται τις σελίδες που είναι ήδη ευρετηριοποιημένες για να ελέγξει τυχόν αλλαγές. Ο χρόνος επανεξέτασης θα πρέπει να μην είναι πολύ συχνός για να μην ελεγχθούν ξανά σελίδες που δεν έχουν αλλαγές.

Οι δημιουργοί των σελίδων για να διαπιστώσουν ότι το Googlebot επισκέφθηκε τη σελίδα τους, μπορούν να ελέγξουν τα «αρχεία πρακτικών» του διακομιστή τους. Η Google παρέχει επίσης τη δυνατότητα στους δημιουργούς να δηλώσουν την σελίδα τους [14]. Μέσω μιας φόρμας υποβολής, η σελίδα τους

[14] Οι διαχειριστές ιστοτόπων μπορούν να ανατρέξουν στο https://www.google.com/webmasters/

προστίθεται στη λίστα προς επίσκεψη της Google, αν και δεν εγγυάται το αν και πότε θα ελέγξει τη σελίδα το Googlebot. Όμως με αυτόν τον τρόπο πολλοί spammers μπορούν να εισάγουν αυτοματοποιημένα εκατομμύρια διευθύνσεις, με στόχο την διαφημιστική ή άλλου είδους προπαγάνδα. Έτσι η Google προσπαθεί να ελέγχει όλες τις διευθύνσεις που εισάγονται στη φόρμα. Έχει προσθέσει ένα πεδίο όπου εμφανίζονται ορισμένα περίεργα γράμματα και ζητά να πληκτρολογεί ο κάθε δημιουργός, κάτι σαν τεστ των ματιών, ώστε να μπορούν να ξεγελάσουν τους spammers και να σταματήσει η κακόβουλη χρήση.

4.3.2 **Λογισμικό ευρετηριοποίησης** (Google's Indexer)

Αφού το Googlebot εξετάσει τις σελίδες, μεταβιβάζει το πλήρες κείμενο των σελίδων αυτών στο λογισμικό ευρετηριοποίησης (Google Indexer). Οι σελίδες αυτές αποθηκεύονται στη βάση δεδομένων του ευρετηρίου της Google. Το ευρετήριο αυτό είναι ταξινομημένο αλφαβητικά με βάση την κάθε λέξη για την αποδοτική αναζήτηση των σελίδων. Για κάθε όρο αποθηκεύεται μια λίστα με όλα τα έγγραφα στα οποία εμφανίζεται η λέξη, καθώς και τη θέση της μέσα στο κείμενο. Αυτή η δομή δεδομένων επιτρέπει την ταχεία πρόσβαση σε έγγραφα που περιέχουν τους όρους αναζήτησης του χρήστη.

Η Google για να βελτιώσει την απόδοση της αναζήτησης δεν ευρετηριοποιεί ορισμένες λέξεις οι οποίες είναι κοινές και παρουσιάζονται σε όλες τις σελίδες. Αυτές ονομάζονται «stop words» και είναι για παράδειγμα οι λέξεις «as» «the» «is» «in» «of» «how» «why», όπως επίσης και μονοψήφιους αριθμούς και γράμματα. Οι λέξεις αυτές είναι συνηθισμένες, παρουσιάζονται πάντα μέσα σε ένα κείμενο και δεν προσφέρουν ουσιαστική πληροφορία σχετικά με την αναζήτηση. Ο Google Indexer αγνοεί επίσης και ορισμένα σημεία στίξης, τους κενούς χαρακτήρες και μετατρέπει όλα τα γράμματα σε μικρά και όχι κεφαλαία για να βελτιώσει τις επιδόσεις της Google.

4.3.3 **Επεξεργαστής ερωτήματος** (Google's Query Processor)

Ο επεξεργαστής ερωτήματος προσπαθεί να επεξεργαστεί το ερώτημα του χρήστη το οποίο ενδεχομένως να περιέχει λέξεις και ορισμένους τελεστές που έχουν

ειδική σημασία για τη μηχανή αναζήτησης. Με τη βοήθεια του Google Indexer εντοπίζει τις σελίδες στις οποίες εμφανίζονται οι λέξεις αυτές και αποκτά πρόσβαση στα δεδομένα τους. Το πιο σημαντικό κομμάτι της διαδικασίας αυτής είναι ο τρόπος εμφάνισης των αποτελεσμάτων το οποίο έχει κάνει και πολύ γνωστή τη μηχανή αναζήτησης Google. Ο επεξεργαστής ερωτήματος κατατάσσει τα αποτελέσματα σε φθίνουσα σειρά με βάση το βαθμό κατάταξής τους. Έτσι μια σελίδα με υψηλό βαθμό κατάταξης είναι πιο σημαντική από μια με χαμηλό και συνεπώς θα εμφανιστεί πιο πάνω στα αποτελέσματα αναζήτησης. Ο επεξεργαστής ερωτήματος λαμβάνει υπ' όψιν του πολλούς παραμέτρους για τον υπολογισμό βαθμολογίας μιας σελίδας, όπως η δημοτικότητα της σελίδας δηλ. σε πόσες άλλες σελίδες εμφανίζεται ως σύνδεσμος, η θέση και το μέγεθος των όρων αναζήτησης μέσα στη σελίδα, το πόσο κοντά είναι μεταξύ τους οι όροι κτλ.

Για να βελτιώσει την απόδοσή του ο επεξεργαστής ερωτήματος εφαρμόζει τεχνικές μηχανικής μάθησης (machine learning). Αν για παράδειγμα στο ερώτημα του χρήστη κάποιος όρος περιέχει ορθογραφικά λάθη τότε η μηχανή αναζήτησης χρησιμοποιεί μια τέτοια τεχνική και προτείνει εναλλακτικές ερωτήσεις διορθώνοντας τις λάθος ορθογραφικά λέξεις. Η Google δεν περιορίζεται μόνο στην αναζήτηση απλών λέξεων μέσα στις σελίδες, αλλά και στην αναζήτηση εικόνων, επιστημονικών εγγράφων, ολόκληρων φράσεων κτλ.

4.4 **Τα έσοδα της** Google

Η Google όπως ήδη αναφέραμε είναι η πιο δημοφιλής μηχανή αναζήτησης και χρησιμοποιεί αυτή την ιδιότητα ως πηγή εσόδων της. Από το πλήθος των υπηρεσιών που προσφέρει οι κυριότερες είναι η υπηρεσία AdWords και η υπηρεσία AdSense, τις οποίες και περιγράφουμε στη συνέχεια.

4.4.1 **Η υπηρεσία** AdWords

Η Google σχεδίασε το AdWords για διαφημιστές που θέλουν να απευθυνθούν σε ένα συγκεκριμένο κοινό με τον πιο αποτελεσματικό τρόπο. Οι διαφημιστές επιλέγουν κάποιες λέξεις-κλειδιά και πληρώνουν στη Google μόνο όταν οι χρήστες κάνουν κλικ στις διαφημίσεις τους. Η δημιουργία του διαφημιστικού κειμένου και η

διαχείριση των ηλεκτρονικών διαφημιστικών λογαριασμών είναι εύκολη υπόθεση και με χαμηλό κόστος.

Η ομάδα πωλήσεων της Google είναι υπεύθυνη για τη βελτιστοποίηση των διαφημιστικών καμπανιών για τους μεγαλύτερους διαφημιστές της. Οι ειδικοί του προγράμματος AdWords συνεργάζονται με τους διαφημιστές ώστε να επιλέξουν τις κατάλληλες λέξεις-κλειδιά, με τέτοιο τρόπο ώστε να βελτιωθεί η πορεία της καμπάνιας, απομονώνοντας τις προβληματικές λέξεις-κλειδιά. Δεν υπάρχει κανένα όριο στον αριθμό λέξεων-κλειδιών που θα επιλέξει ο διαφημιστής και κάθε μια από αυτές μπορεί να αντιστοιχεί σε διαφορετική δημιουργική εκτέλεση. Οι διαφημιστές μπορούν να αυξήσουν ακόμα περισσότερο την απόδοση της καμπάνιας τους, τοποθετώντας τις διαφημίσεις τους σε συγκεκριμένες γεωγραφικές περιοχές ή χρησιμοποιώντας διάφορες γλώσσες.

Η ταξινόμηση των αποτελεσμάτων γίνεται με βάση το ποσό που διαθέτει ο διαφημιστής, αλλά και με βάση τη βαθμολογία ποιότητας που έχει η κάθε σελίδα που διαφημίζεται. Η βαθμολογία ποιότητας είναι ένα μέγεθος της Google όπου λαμβάνεται υπ' όψιν το ιστορικό κάθε σελίδας όσον αφορά τις προηγούμενες επιλογές των χρηστών, καθώς επίσης και με βάση το περιεχόμενό της, την ευκολία πλοήγησης κτλ.

4.4.2 **Η υπηρεσία** AdSense

Η Google πιστεύει ότι οι διαφημίσεις μπορεί να είναι εξίσου χρήσιμες με τα αποτελέσματα αναζήτησης ή τις άλλες μορφές περιεχομένου. Η υπηρεσία AdSense συνδυάζει την τεχνολογία αναζήτησης της Google με τη βάση των διαφημιστών που επιλέγουν την προώθηση μέσω της χρήσης λέξεων-κλειδιών για την παροχή ακριβούς στοχοθέτησης των αποτελεσμάτων αναζήτησης ή του περιεχομένου στις σελίδες ενός ιστότοπου, ανεξάρτητα από το βαθμό εξειδίκευσης του θέματος. Με βάση αυτό επωφελούνται οι διαφημιστές, οι εκδότες και όσοι αναζητούν πληροφορίες.

Η εγγραφή στην υπηρεσία AdSense είναι πολύ εύκολη και διαρκεί μόνο λίγα λεπτά. Το όφελος από αυτή την υπηρεσία είναι ότι ο ιδιοκτήτης του διαδικτυακού τόπου πληρώνεται κάθε φορά που κάποιος χρήστης επιλέξει κάποιον σύνδεσμο από αυτούς που παρουσιάζονται στις διαφημίσεις.

Οι διαφημίσεις οι οποίες εμφανίζονται είναι σχετικές με το περιεχόμενο του διαδικτυακού τόπου και ο σχεδιαστής του μπορεί να επιλέξει το σημείο που θα τις τοποθετήσει. Η υπηρεσία AdSense έγινε δημοφιλής κυρίως επειδή οι διαφημίσεις οι οποίες παρουσιάζονται δεν έχουν τη μορφή διαφημιστικών εικόνων και έτσι είναι περισσότερο καλαίσθητες. Πλέον πολλοί δικτυακοί τόποι χρησιμοποιούν αυτή την υπηρεσία, η οποία είναι αρκετά βολική για μικρές επιχειρήσεις που δεν μπορούν να ανταπεξέλθουν οικονομικά ώστε να δημιουργήσουν ένα ξεχωριστό τμήμα marketing, το οποίο θα δραστηριοποιείται στην εύρεση πελατών που ενδιαφέρονται να διαφημιστούν με αυτόν τον τρόπο.

4.5 **Άλλες υπηρεσίες της** Google

Εκτός από τις υπηρεσίες που αναφέραμε παραπάνω, οι οποίες αποτελούν την κύρια πηγή εσόδων της, η Google προσφέρει και μια πληθώρα άλλων υπηρεσιών, όπως:

- Την υπηρεσία ηλεκτρονικού ταχυδρομείου, Gmail
- Την υπηρεσία χαρτών σε ψηφιακή μορφή, Google Maps
- Την υπηρεσία Google Storage
- Την υπηρεσία αναζήτησης βιβλίων
- Τις διαφημιστικές υπηρεσίες Double-click, Audio Ads, Click to call.

4.6 **Στα άδυτα της** Google

Στις μέρες μας, όπου η πληροφορική επανάσταση είναι σε πλήρη εξέλιξη, τα εργοστάσια εξακολουθούν να καταναλώνουν ενέργεια αλλά με διαφορετικό ρόλο. Καταναλώνουν ασύλληπτα υψηλό ποσό ενέργειας και είναι συνδεδεμένα με τον υπολογιστή μας.

Η Google έχει 13 data centers και καταναλώνει (μετά από ανακοίνωσή της το φθινόπωρο του 2011) 260 εκατομμύρια Watt ετησίως, ηλεκτρισμός που αρκεί να τροφοδοτήσει μια πόλη 200.000 κατοίκων, καθώς οι υπολογιστές της ημερησίως καταχωρίζουν πάνω από 20 δισ. σελίδες την ημέρα, επεξεργάζονται 100 δις αναζητήσεις το μήνα, διαχειρίζονται τα e-mails 425 εκατ. χρηστών και δέχονται 72

ώρες βίντεο κάθε λεπτό της ημέρας στο YouTube. Και σύντομα αν όλα πάνε καλά με το σύστημα Glass, όλα τα ανωτέρω θα προβάλλονται με γυαλιά, όπως τα συνηθισμένα με το κρύσταλλο, και στο κρύσταλλο θα βλέπουμε διάφανο το διαδίκτυο πάνω στον πραγματικό κόσμο.

Στα εργοστάσια πληροφορίας που διαθέτει η Google υπάρχουν εγκαταστάσεις όπου διαθέτουν σωλήνες ψύξης στο κάτω πάτωμα (εικόνα 10), όπου βρίσκονται οι servers, προκειμένου η θερμοκρασία να παραμένει χαμηλά. Ελάχιστα άτομα έχουν τη δυνατότητα πρόσβασης σε αυτό το χώρο (ούτε 10 από τα 100 που εργάζονται σε κάθε data center) και φοράνε υποχρεωτικά ωτοασπίδες ή ακουστικά με ενδοεπικοινωνία, καθώς ο θόρυβος που παράγουν είναι ανυπόφορος, αλλά και προκειμένου να κρατηθούν κρυφά τα βιομηχανικά μυστικά για τη λειτουργία των data centers.

Εικόνα 106: Σωλήνες ψύξης της Google

Παρόλο τη μεγάλη κατανάλωση ρεύματος που καταναλώνει η Google, το 2012 η Greenpeace την επαίνεσε για τις επενδύσεις της σε ανεμογεννήτριες και ηλιακές κυψέλες, καθώς και για το ότι υιοθετεί σχεδόν όλες τις οδηγίες για μείωση του ενεργειακού αποτυπώματός της [K2012].

ΚΕΦΑΛΑΙΟ 5ο **ΤΟ ΜΕΛΛΟΝ ΤΗΣ** GOOGLE

5.1 **Ο αλγόριθμος** PigeonRank

Κάθε χρήστης της Google είναι εξοικειωμένος με την ταχύτητα και τη δυναμική της μηχανής αναζήτησης της. Το ερώτημα το οποίο τίθεται είναι πώς η Google καταφέρνει να βρίσκει τα σωστά αποτελέσματα τόσο γρήγορα. Η βάση της ερευνητικής διαδικασίας για τη Google βασίζεται στην τεχνολογία εύρεσης Pigeon Rank[15] (εικόνα 11), η οποία αποτελεί ένα σύστημα κατηγοριοποίησης σελίδων με βάση το περιεχόμενό τους αλλά και με βάση τη σχετικότητά τους με το ζητούμενο της έρευνας. Η τεχνολογία αυτή ανακαλύφθηκε μέσα από την έρευνα των Larry Page and Sergey Brin.

Οι προαναφερόμενοι ερευνητές επιδίωξαν να βρουν ένα χαμηλού κόστους μηχανισμό ο οποίος θα έκανε γρήγορη εύρεση ανάμεσα σε εκατομμύρια ιστοσελίδες, προκειμένου να δώσει το επιθυμητό αποτέλεσμα στο χρήστη. Ο μηχανισμός αυτός λειτουργεί φυσικά πιο γρήγορα από οποιαδήποτε άλλη διαδικασία ακολουθούσε ο άνθρωπος ακόμα και αν αυτή βασιζόταν σε κάποιο εξειδικευμένο μαθηματικό αλγόριθμο. Η Google σήμερα χρησιμοποιεί χιλιάδες τεχνικούς οι οποίοι βελτιώνουν τη λειτουργία του Pigeon ενισχύοντας συνεχώς την ερευνητική του δυναμική.

Η επιτυχία του προγράμματος αναφέρεται από τη μια στον τοπικό server της εταιρίας τον Columbia livia και στη συνέχεια στην ικανότητα του να αναγνωρίζει τις κατάλληλες ιστοσελίδες με βάση τη διεξαχθείσα έρευνα του χρήστη. Το pigeon μελετά όλες τις παραδοσιακές ιστοσελίδες και συνεχώς ενημερώνεται για νέες, τις αξιολογεί με βάση τη δυναμική, το περιεχόμενο, τις λέξεις κλειδιά αλλά και την επισκεψιμότητα τους. Η κατάταξη διευκολύνει τους ιδιοκτήτες των σελίδων ώστε να τοποθετούνται σε υψηλότερα σημεία και να τις βρίσκουν πιο εύκολα και πιο γρήγορα οι χρήστες. Κάποιες φορές το πρόγραμμα επηρεάζεται από παράγοντες που δεν μπορεί να υπολογίσει ότι θα συναντήσει όπως φωτογραφίες, γραφήματα κ.λ.π.

[15] http://www.google.com/technology/pigeonrank.html

Εικόνα 11: Το σύστημα Pigeon Rank

5.2 **Από τον Παγκόσμιο Ιστό στο** Semantic Web

Ο Σημασιολογικός Ιστός (Semantic Web) είναι ένα όραμα και μια πρόταση για την μετεξέλιξη του Διαδικτύου και ειδικότερα του Παγκόσμιου Ιστού (World Wide Web). Ο όρος Semantic Web, καθώς και η αρχιτεκτονική για την υλοποίησή του, προτάθηκαν από τον Tim Berners-Lee, τον εφευρέτη του σημερινού Παγκόσμιου Ιστού. Ο Σημασιολογικός Ιστός υιοθετήθηκε από το World Wide Web Consortium (W3C)[16], έναν οργανισμό που στοχεύει στην προώθηση, ανάπτυξη και εξέλιξη του Web και των πρωτοκόλλων που το υποστηρίζουν. Πρόκειται για μια επέκταση και βελτίωση του σημερινού Web στην κατεύθυνση, κυρίως της δόμησης της πληροφορίας, έτσι ώστε αυτή να είναι προσπελάσιμη από εφαρμογές υπολογιστών, με τελικό στόχο την αυτοματοποίηση πολλών λειτουργιών στο διαδίκτυο. Η σημερινή αναπαράσταση της πληροφορίας που προορίζεται για χρήση από ανθρώπους θα αντικατασταθεί από μια αναπαράσταση κατανοητή από υπολογιστές [BHL2001].

Μερικά από τα πεδία στα οποία αναμένεται να έχει την μεγαλύτερη επίδραση είναι στην υγεία, στην παιδεία και στην επιχείρηση. Υπάρχουν ήδη πολλές προσπάθειες από εταιρείες, ερευνητές και μη κερδοσκοπικές οργανώσεις για να πράξουν πρότυπα οντολογιών, κυρίως για τα παραπάνω πεδία, για να μπορούμε να έχουμε κοινές γλώσσες και περισσότερα δεδομένα τα οποία να μπορούμε να συνδυάσουμε για να έχουμε καλύτερα αποτελέσματα. Στην υγεία προσπαθούμε να δημιουργήσουμε ενοποιημένες γλώσσες ιατρικής ορολογίας και υπηρεσίες οι οποίες θα βοηθάνε το ιατρικό προσωπικό και θα κατευθύνουν τους καταναλωτές σε αξιόπιστες πληροφορίες υγείας σχετικά με την κατάστασή τους. Στην εκπαίδευση ο Σημασιολογικός Ιστός θα συμβάλει σημαντικά στην μάθηση κυρίως στον τρόπο με τον οποίο αναζητάμε και μας επιστρέφονται οι πληροφορίες, στην οργάνωση των αποτελεσμάτων και στην δημιουργία ενός προγράμματος μάθησης ειδικό για το

[16] www.w3.org

καθένα. Στην επιχείρηση θα δούμε καλύτερη οργάνωση των εταιρειών, καλύτερες εμπειρίες για τους χρήστες στις διαδικτυακές αγορές και καλύτερο συντονισμό μεταξύ διαφορετικών εταιρειών. Στην καθημερινότητά μας θα δούμε τις επιδράσεις του Web3 (το οποίο θα αναλύσουμε παρακάτω) στα κοινωνικά δίκτυα και εικονικές κοινότητες. Θα έχουμε εφαρμογές οι οποίες θα δίνουν περισσότερες, πιο έμπιστες, πληροφορίες και θα διευκολύνουν σημαντικά τις διαδικτυακές μας δραστηριότητες [AH2009].

Ο σημασιολογικός ιστός προσπαθεί να επιλύσει το πρόβλημα της αναπαράσταση της γνώσης από τους υπολογιστές [FHLW03]. Βασικό συστατικό του είναι ο μηχανισμός επεξεργασίας της γνώσης που διαχειρίζεται λογικά τις πληροφορίες με σκοπό την εξαγωγή συμπερασμάτων, τη δημιουργία νέας γνώσης, την υποστήριξη στη λήψη αποφάσεων και τέλος την αυτόματη εκτέλεση ενεργειών.

Οι βασικές αρχές του σημασιολογικού ιστού είναι:

- Η διατήρηση του κατανεμημένου περιεχομένου του διαδικτύου.
- Η αναπαράσταση και ανάκτηση της πληροφορίας, καθώς οι εφαρμογές των υπολογιστών προσπελαύνουν δομημένες πηγές πληροφορίας και κανόνες, οι οποίοι χρησιμοποιούνται για να αιτιολογούν τις σχέσεις μεταξύ των πληροφοριών.
- Η αναπαράσταση των εννοιών μιας θεματικής περιοχής (λ.χ. του τουρισμού) επιτυγχάνεται με τη χρήση των οντολογιών.
- Η ύπαρξη πρακτόρων λογισμικού (software agents), δηλαδή προγραμμάτων που θα αναλαμβάνουν για λογαριασμό του χρήστη να κινούνται στο διαδίκτυο και να συλλέγουν την πληροφορία από διάφορες πηγές που διαθέτουν σημασιολογικό περιεχόμενο.

5.2.1. **Βασικές τεχνολογίες και εργαλεία για τον Σημασιολογικό Ιστό**

Οι βασικές τεχνολογίες που χρησιμοποιούνται στον σημασιολογικό ιστό, συνοπτικά, είναι οι εξής:

- Η XML (eXtensible Markup Language – επεκτάσιμη γλώσσα σήμανσης) όπως αναφέραμε και στο πρώτο κεφάλαιο είναι η επικρατέστερη γλώσσα για την περιγραφή και ανταλλαγή δεδομένων και κειμένων στο διαδίκτυο.

Παρέχει τη δυνατότητα δημιουργίας κειμένων με απεριόριστα πολύπλοκη δομή και συντακτικό. Έτσι, μπορούν να δομηθούν οι πληροφορίες που περιέχονται στα κείμενα για να επεξεργάζονται πιο εύκολα από τους υπολογιστές [Dick03].

- Το πρότυπο XML συμπληρώνεται με το XML Schema, μια γλώσσα με την οποία γράφουμε «λεξικά» και «γραμματικές» για XML κείμενα. Το XML Schema ορίζει τα επιτρεπόμενα στοιχεία, τις ιδιότητές ους, καθώς και τον τρόπο με τον οποίο συνδυάζονται μεταξύ τους μέσα στο XML κείμενο. Με απλά λόγια, το XML Schema αποτελεί το «συντακτικό» του XML κείμενο [Vlist2002].
- Η γλώσσα RDF (Resource Description Framework – Περιβάλλον Περιγραφής Πόρων) είναι το πρότυπο που υιοθετήθηκε από το W3C για την περιγραφή πληροφοριακών πόρων και γενικότερα για την αναπαράσταση της γνώσης στο περιβάλλον του διαδικτύου. Μέσω του RDF είναι δυνατή η μετατροπή της πληροφορίας σε σημασιολογική. Πόρος (resource) είναι οτιδήποτε θέλουμε να δηλώσουμε ή να περιγράψουμε. Για παράδειγμα, πόρος μπορεί να είναι μια ιστοσελίδα, ένας δικτυακός τόπος, ένα αντικείμενο, μια έννοια κλπ. Κάθε πόρος προσδιορίζεται με το Καθολικό Αναγνωριστικό Πόρου (Universal Resource Identifier – URI).
- Το RDF Schema είναι η οντοκεντρική επέκταση του RDF. Είναι μια γλώσσα με την οποία το μοντέλο δεδομένων του RDF εμπλουτίζεται με τα χαρακτηριστικά αντικειμενοστρεφούς αναπαράστασης, όπου ο πόρος αντιστοιχεί σε αντικείμενο. Συγκεκριμένα, το RDF Schema ορίζει ένα λεξικό για να εκφράζονται οι κατηγορίες (κλάσεις) των πόρων, οι πόροι, οι ιδιότητές τους και οι μεταξύ τους σχέσεις [BKH2002]
- Η OWL (Web Ontology Language – Γλώσσα Οντολογιών Ιστού) είναι μια γλώσσα που χρησιμοποιείται για την περιγραφή των οντολογιών που υπάρχουν στο διαδίκτυο [GH2003]. Η Οντολογία είναι μια αυστηρή περιγραφή των πόρων και των μεταξύ τους σχέσεων. Συγκεκριμένα, η οντολογία είναι η αποδεκτή σημασιολογικά κωδικοποίηση της πληροφορίας ενός θεματικού χώρου. Οι οντολογίες επιτρέπουν στους χρήστες να έχουν κοινή ονοματολογία και αντίληψη για τα αντικείμενα που δηλώνουν ή χρησιμοποιούν. Βοηθούν τον χρήστη να πλοηγηθεί στον

θεματικό χώρο της πληροφορίας που βασίζεται σε σημασιολογικές και όχι σε λεξικολογικές έννοιες. Στις οντολογίες, η δυσκολία εντοπίζεται στο ότι οι κοινότητες χρηστών με κοινά ενδιαφέροντα θα πρέπει να συμφωνήσουν στην οντολογική περιγραφή του θεματικού χώρου ενδιαφέροντός τους. Για την περιγραφή των οντολογιών έχει αναπτυχθεί η γλώσσα DAML ενώ μια αξιόλογη υποδομή οντολογιών για τον σημασιολογικό ιστό είναι η OIL [Fensel01].

- Οι πράκτορες λογισμικού (software agents) είναι προγράμματα που εκτελούν κάποια λειτουργία και παράγουν αποτελέσματα με το πέρας της εκτέλεσης αυτής. Συνήθως, οι πράκτορες λογισμικού περιδιαβαίνουν το διαδίκτυο και επεξεργάζονται τις πληροφορίες που βρίσκουν στις ιστοσελίδες που επισκέπτονται. Συχνά, οι πράκτορες λογισμικού χρησιμοποιούνται για λειτουργίες όπως η εύρεση, ταξινόμηση και επιλογή δεδομένων. Στο ηλεκτρονικό εμπόριο, μερικά παραδείγματα λειτουργιών τους είναι η σύγκριση τιμών του ίδιου προϊόντος σε πολλά ηλεκτρονικά καταστήματα, η ειδοποίηση για την εμφάνιση νέου περιεχομένου σε δικτυακούς τόπους ειδήσεων και ενημέρωσης κ.α.

Τα εργαλεία που χρησιμοποιούνται για τον σημασιολογικό ιστό είναι τα εξής:

- Επίσημες γλώσσες για την έκφραση και την αναπαράσταση των οντολογιών.
- Επεξεργαστές για την ημιαυτόματη δόμηση και δημιουργία νέων οντολογιών.
- Οντολογικά περιβάλλοντα δημιουργίας νέων οντολογιών από τις ήδη υπάρχουσες, δηλαδή περιβάλλοντα επαναχρησιμοποίησης και συγχώνευσης των οντολογιών.
- Υπηρεσίες αιτιολόγησης-εκλογίκευσης (reasoning).
- Εργαλεία συμβολισμού (annotation tools) για τη σύνδεση μη δομημένων και ημιδομημένων πηγών πληροφορίας με τη χρήση μεταδεδομένων (metadata).
- Εργαλεία έξυπνης πρόσβασης σε πηγές πληροφορίας.
- Εργαλεία μετάφρασης και ολοκλήρωσης υπηρεσιών ανάμεσα σε διαφορετικές οντολογίες που ανταλλάσσουν δεδομένα πολλαπλών προτύπων και ορισμών.

5.3 **Το** Web3

Ο ορισμός του web3 έχει δοθεί από τον Tim Berners-Lee [BLF99], όπου μιλάει για τον σημασιολογικό ιστό και αυτή ουσιαστικά είναι η μεγαλύτερη διαφορά σε σχέση με το παρελθόν. Περιγράφει έναν κόσμο στον οποίο το διαδίκτυο θα έχει τη δυνατότητα ανάλυσης όλων των δεδομένων που υπάρχουν σε αυτό, με αποτέλεσμα οι περισσότερες υποθέσεις μας να διεκπεραιώνονται από τις διαμεσολαβούσες μηχανές. Αυτό το όραμα κινητοποίησε αρκετούς επιστήμονες που με το έργο τους άρχισαν να δίνουν σάρκα και οστά στο web3.

Το web3 είναι με λίγα λόγια η νέα γενιά του παγκόσμιου ιστού, όπου πλέον ο χρήστης θα είναι το επίκεντρο και οι δυνατότητές του δε θα σταματούν πουθενά. Αν για παράδειγμα ένας χρήστης έχει αποφασίσει να δει μια ταινία στο σινεμά και αργότερα να φάει σε ένα κινέζικο εστιατόριο, με το σημερινό web θα περιπλανηθεί σε αρκετές σελίδες μέχρι να βρει τις ταινίες που παίζονται, τους κινηματογράφους που τον βολεύουν και τα εστιατόρια που είναι κοντά στους κινηματογράφους. Με το web3 αυτή η αναζήτηση θα είναι πιο εύκολη και πιο γρήγορη. Για παράδειγμα, θα υπάρχουν μηχανές αναζήτησης όπου θα πληκτρολογεί την φράση «Θέλω να δω μια αστεία ταινία και μετά να φάω σε ένα καλό εστιατόριο. Ποιες είναι οι επιλογές μου;». Ο φυλλομετρητής θα λάβει το αίτημα, αφού το αναλύσει θα επιστρέψει όλες τις πιθανές απαντήσεις.

Αυτό μπορεί να εξελιχτεί και σε ένα βήμα παραπέρα. Να λειτουργεί ως ένας προσωπικός βοηθός, δηλαδή να επιλέγει τις προτιμήσεις του χρήστη με βάση τις αρέσκειες του, την τοποθεσία του, την τιμή, την διαθεσιμότητα των τραπεζιών κλπ. Τέλος οι συσκευές θα έχουν αναβαθμιστεί σε τέτοιο επίπεδο που θα λαμβάνουν φωνητικές εντολές, αισθητήρες κίνησης κλπ.

Το Web3 δίνει δηλαδή με τα εργαλεία προγραμματισμού που χρησιμοποιεί τη δυνατότητα να συνδέσουμε το συντακτικό με το νόημα που θέλουμε να προσδώσουμε στις έννοιες, πράγματα που χρησιμοποιούμε. Αυτό επιτυγχάνεται με τις γλώσσες προγραμματισμού OWL [GH2003], RDF [BKH2002] και Ajax [ZMF2011].

5.3 **Η γλώσσα** HTML5

Η γλώσσα HTML5 [Keith10], η οποία αναμένεται να κυκλοφορήσει το 2014, είναι εξοπλισμένη από ένα μεγάλο όγκο νέων χαρακτηριστικών που θα φέρει τη συμβατότητα των προγραμμάτων περιήγησης σε ένα νέο επίπεδο πολυμέσων και βίντεο υποστήριξης, επιτάχυνσης, συστηματικής αποθήκευσης, πινέλα σχεδίασης και διάφορων άλλων που θα ανακαλυφθούν με την υλοποίησή της.

Μέχρι τώρα όταν ένα πρόγραμμα περιήγησης ανέλυε μια λανθασμένη εντολή ο κάθε φυλλομετρητής αντιμετώπιζε το θέμα με χρήση της δικιάς του ενσωματωμένης τεχνολογίας και αυτό οδηγούσε σε δυσάρεστα αποτελέσματα καθώς δεν υπήρχε συμβατότητα με τους φυλλομετρητές μεταξύ τους. Με την HTML5 όμως η ερμηνεία του κώδικα της ιστοσελίδας θα είναι καθορισμένη, επομένως και οι φυλλομετρητές θα πρέπει να είναι συμβατοί με αυτό.

Σημαντική βελτίωση θα προσφέρει η γλώσσα HTML5 και στα πολυμέσα (multimedia), αφού μέχρι τώρα για να μπορέσει να τα διαβάσει ένας φυλλομετρητής πρέπει να εγκατασταθούν σε αυτόν ως πρόσθετα (plugins) προκειμένου να δείξουν καλά και παραμείνουν ως πρόσθετα. Θα πρέπει δηλαδή ο χρήστης να εγκαταστήσει πρώτα τα πρόσθετα για να μπορέσει να παρακολουθήσει μια ταινία. Με τη νέα γλώσσα θα υπάρχουν καθορισμένες ετικέτες για το περιεχόμενο των πολυμέσων και θα μπορεί ευκολότερα ο φυλλομετρητής να υποστηρίξει ένα πρόγραμμα περιήγησης για ταινίες και ήχο. Επιπλέον ο φυλλομετρητής θα μπορεί να επιλέξει την καλύτερη μορφή για να παίξει ένα βίντεο.

Το geolocation είναι ακόμα μια καινοτομία της HTML5, όπου θα δίνει τη δυνατότητα στο χρήστη να επισημαίνει την γεωγραφική θέση του στον χάρτη. Επίσης υποστηρίζει σχέδιο και 3D επιτάχυνση στον φυλλομετρητής. Ο χρήστης δηλαδή θα μπορεί να παίζει ένα παιχνίδι online με ένα κλικ, αφού ήδη οι ταχύτητες σύνδεσης στο διαδίκτυο είναι αρκετά μεγάλες στις περισσότερες περιοχές και αυξάνονται καθημερινά όλο και περισσότερο.

5.4 **Το λειτουργικό σύστημα** Android

Το λειτουργικό σύστημα Android (βλέπουμε το λογότυπο του στην εικόνα 12) είναι μια πλατφόρμα για κινητές συσκευές η οποία περιλαμβάνει το λειτουργικό

σύστημα, το ενδιάμεσο λογισμικό και τις βασικές εφαρμογές. Δημιουργήθηκε από την Android Inc., η οποία μεταγενέστερα αγοράστηκε από την Google, για αυτό και παρέχει πολλές από τις λειτουργίες της Google όπως Αναζήτηση, Χάρτες, Gmail, Google Earth, ημερολόγιο της Google κ.α.

Εικόνα 12: **Το λογότυπο του λειτουργικού** Android

Πρόκειται για μια πλατφόρμα κινητών συσκευών που στηρίχθηκε στον ελεύθερο πυρήνα του Linux και το μεγαλύτερο μέρος του κώδικα του Android δημοσιεύτηκε από την Google. Ως πλατφόρμα ανοικτού κώδικα, μπορεί εύκολα να επεκταθεί και να τροποποιηθεί με σκοπό να συμβαδίζει με τις τελευταίες τεχνολογίες και εξελίξεις. Επομένως, θα εξελίσσεται συνεχώς και έχει μια συνεχή πρόοδο.

Είναι η πρώτη ελεύθερη πλατφόρμα για κινητές συσκευές, κάτι το οποίο επιτρέπει στο χρήστη τη δυνατότητα επιλογής των εφαρμογών που θα εγκαταστήσει. Επίσης, είναι δωρεάν, που σημαίνει ότι ο χρήστης μπορεί να εγκαταστήσει οποιαδήποτε έκδοση του λειτουργικού θέλει σε οποιαδήποτε συσκευή ή να αναβαθμίσει την υπάρχουσα χωρίς οικονομική επιβάρυνση.

Επιπλέον η ύπαρξη της αγοράς Android εφαρμογών (Android Market) επιτρέπει στους χρήστες να κατεβάζουν εφαρμογές (μόνο εάν έχουν δημιουργήσει και εγκαταστήσει στη συσκευή τους το gmail) και να βρίσκουν ενημερώσεις εύκολα και στους προγραμματιστές να μοιράζονται τις εφαρμογές που δημιουργούν δωρεάν ή με πολύ μικρό κόστος.

Αυτό που ουσιαστικά έχει καταφέρει να πετύχει η Google με αυτό το λειτουργικό σύστημα είναι ο χρήστης να μπορεί να χειρίζεται την κινητή συσκευή του όπως ακριβώς και τον υπολογιστή του. Να έχει δηλαδή τη δυνατότητα να μεταδίδει δεδομένα από το κινητό στο διαδίκτυο (φωτογραφίες, λίστα με επαφές, τρέχουσα θέση) και να λάβει όλα όσα μπορεί να χρειαστεί διαδικτυακά και να εμφανίζονται στην οθόνη της συσκευής του.

ΕΠΙΛΟΓΟΣ

Στις μέρες μας η χρήση των ηλεκτρονικών υπολογιστών και ειδικότερα η χρήση του διαδικτύου αυξάνεται αλματωδώς και όλο και περισσότερες νέες σελίδες εντάσσονται στις ήδη υπάρχουσες του Παγκόσμιου Ιστού.

Στην εργασία αυτή είδαμε ότι εξαιτίας αυτής της ραγδαίας αύξησης του όγκου πληροφορίας στο διαδίκτυο, που συνδέεται με περισσότερες από ένα δισεκατομμύριο συνδέσμους, κατέστη απαραίτητο να επινοηθούν νέοι τρόποι ταξινόμησης και ευρετηρίασης των ιστοσελίδων και των εγγράφων που υπάρχουν στον Παγκόσμιο Ιστό, ο οποίος μπορεί να χαρακτηριστεί και ως ένα κατευθυντικό γράφημα. Αυτό μπορεί να πραγματοποιηθεί μέσω των τεχνικών και των αλγορίθμων που χρησιμοποιεί η κάθε μηχανή αναζήτησης και αναπτύχθηκαν εκτενέστερα στο τρίτο κεφάλαιο.

Η Google, όπου σύμφωνα με έρευνες είναι η δημοφιλέστερη μηχανή αναζήτησης που υπάρχει αυτή την στιγμή στο διαδίκτυο, αποτελεί φαινόμενο και αντικείμενο μελέτης όχι μόνο για τους φανατικούς χρήστες του διαδικτύου αλλά και για πολλούς οικονομικούς αναλυτές. Είναι μια μηχανή αναζήτησης που ξεκινώντας από ένα ερευνητικό πρόγραμμα στο πανεπιστήμιο Stanford και χάρη στον αλγόριθμο Pagerank ακολούθησε μια ραγδαία ανοδική πορεία και έγινε η κυρίαρχη μηχανή αναζήτησης που κατέκτησε τον κόσμο. Προκειμένου να διατηρήσει την κυρίαρχή της θέση και τη πορεία της ανακαλύπτει συνεχώς νέες τεχνικές και αλγορίθμους, όπως το PigeonRank, τους οποίους αναφέραμε στο πέμπτο κεφάλαιο.

Το φαινόμενο Google, σύμφωνα με κάποιους αναλυτές οφείλει την επιτυχία της στο γεγονός ότι δεν αλλάζει τον προσανατολισμό της και παραμένει μια απλή και εύκολη στην χρήση μηχανή αναζήτησης. Η αποστολή της Google, σύμφωνα με την ίδια την εταιρεία παραμένει αναλλοίωτη, να οργανώνει την παγκόσμια πληροφορία και να την κάνει προσβάσιμη και χρήσιμη για κάθε άνθρωπο.

Βιβλιογραφία

[BHL2001] T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, Scientific American, 285 (5), 34 – 43, 2001.

[BKH2002] Jeen Broekstra Arjohn Kampman,and Frank van Harmelen. *Sesame: A generic Architecture for Storing and Querying RDF and RDF Schema*, in Proc. of the 1st International Semantic Web Conference, LNCS 2342, pp. 54 – 68, 2002.

[BLF99] Tim Berners-Lee and Mark Fischetti, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor*, Britain: Orion Business, 1999.

[BP2007] Nancy Blachman and Jerry Peek, *How Google Works,* 2007. (www.googleguide.com/google_works.html)

[BP98] Sergey Brin and Lawrence Page, *The anatomy of a large-scale hypertextual Web Search Engine*, in Proc. of the Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

[BR2009] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieva*l, ACM Press, 2009.

[BR2010] Albert T. Bharucha-Reid, *Elements of the theory of Markov processes and their applications*, Dover Publications, 2010.

[Cormick10] Jason Mc Cormick, *Seo Made Simple for 2011*, Royal House Publishing Inc., 2010.

[Craven] Phil Craven, *Google's PageRank Explained and how to make the most of it*. (www.webworkshop.net/pagerank.html)

[CS2002] Chris Sherman, *Teoma vs Google Round 2*, 2002. (http://searchenginewatch.com/article/2067648/Teoma-vs.-Google-Round-Two)

[Dick03] Kevin Dick, *XML: A Manager's Guide,* Addison-Wesley, 2003.

[Fensel01] D Fensel et al,. *OIL: An ontology Infrastructure for the Semantic Web*, IEEE Intelligent Systems 16, pp. 38 – 44, 2001.

[FHLW03] Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster (eds), *Spinning the Semantic Web*, MIT Press, 2003.

[Franceschet11] Massimo Franceschet, *Pagerank: Standing on the Shoulders of Giants*, Communications of the ACM, Vol. 54 (6), pp. 92 – 101, June 2011.

[GC2000] James Gillies and Robert Cailliau, *How the web was born – The Story of the World Wide Web*, Oxford University Press, 2000.

[GH2003] Deborah McGuinness and Frank van Harmelen (eds.) *OWL Web Ontology Language Overview*, 2003 (http://www.w3.org/TR/2003/WD-owl-features-20030331/)

[Keith10] Jeremy Keith, *HTML5 for Web Designers*, Jeffrey Zeldman, 2010.

[Kleinberg98] Jon Kleinberg, *Authoritative sources in a hyperlinked environment*, in Proc. of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[Kleinberg00] Jon Kleinberg, *The Small-World Phenomenon: an Algorithmic Perspective*, in Proc. of the 32nd ACM Symposium on Theory of Computing, pp. 163-170, 2000.

[LC98] François Laburthe and Yves Caseau, *SALSA: A Language for Search Algorithms*, in Proc. of Constraint Programing 1998 (CP'98), LNCS 1520, 1998, pp. 310 – 324.

[LM2007] Amy Langville and Carl D. Meyer, *Google's Pagerank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006.

[Lazar00] J. Lazar, *User-Centered Web Development*, Jones & Bartlett Learning, 2000.

[LM2000] R. Lempel and S. Moran, *The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect*, Computer Networks 33, pp. 387 – 401, 2000.

[MY2011] Weiyi Meng and Clement T. Yu, *Advanced Meta-search Engine Technology (Synthesis Lectures on Data Management)*, Morgan & Claypool Publishers, 2011.

[Nadeau98] David R. Nadeau, *Introduction to VRML 97*, 1998. (http://www.unibuc.ro/prof/niculae_c_m/bioinfo/vrml/vrml97.pdf)

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, *The PageRank Citation Ranking: Bringing order to the web*, Technical Report. Stanford InfoLab, January 1998.

[QS98] Joshua Quittner and Michelle Slatalla, Speeding the Net: The Inside Story of Netscape and How it Challenged Microsoft. Atlantic Monthly Press, New York, 1998.

[Sobek02] Markus Sobek, *The PageRank Algorithm*, 2002. (pr.efactory.de/e-pagerank-algorithm.shtml)

[SS86] Y. Saad and M.H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7:856-869, 1986.

[TS2011] K. Thulasiraman, and M.N.S. Swamy, *Graphs: Theory and Algorithms*", John Wiley & Sons, 1992.

[Vlist2002] Eric van der Vlist, XML Schema: *The W3C's Object-Oriented Descriptions for XML*, O' Reilly 2002.

[Vorst92] H.A. Van der Vorst, *Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems*, SIAM J. Sci. and Stat. Comput. 13 (2): 631–644, 1992.

[VRML] *Εγχειρίδιο VRML 1.0*. (www.it.uom.gr/project/vrml/vrml01/index.html)

[ZGMP04] Gyöngyi Zoltán, Hector Garcia-Molina, and Jan Pedersen, *Combating Web Spam with TrustRank*, in Proc. of the 30th International Conference on Very Large Data Bases, Toronto Canada, pp. 576 – 587, 2004.

[ZMF2011] Nicholas C. Zakas, Jeremy McPeak, Joe Fawcett, ***Professional Ajax***,Wiley, 2011.

[ΑΗ2009] Γρηγόρης Αντωνίου και Frank van Hatmelen, «Εισαγωγή στο Σημασιολογικό Ιστό», εκδόσεις Κλειδάριθμος, 2009.

[ΘΖ2005] Στέλλα Θεοδώρου και Κωνστάνς Ζαφείρη, *Μηχανές αναζήτησης & **Directories***, Inertia Design, 2005.

[Κ2012] Εφημερίδα Καθημερινή, ένθετο Κ, τεύχος 491, σελ. 27-28, Οκτώβριος 2012.

[Κούδας07] Παναγιώτης Κούδας, *Τι είναι το **Google PageRank***, 2007. (www.starfish.gr/seo/what-is-google-page-rank), 2007.

[Λαμπρόπουλος11] Μιχάλης Λαμπρόπουλος, *Η μηχανή αναζήτησης **Google: περιγραφή λειτουργίας, επιθέσεις και τρόποι αντιμετώπισης***, διπλωματική εργασία, Πανεπιστήμιο Πειραιώς, Τμήμα Ψηφιακών Συστημάτων, 2011. (http://digilib.lib.unipi.gr/dspace/bitstream/unipi/4051/1/Lampropoulos%2c%20Michalis.pdf)

[Μακρής08] Χρήστος Μακρής, *Αποθήκευση και Ανάκτηση Πληροφορίας*, Διδακτικές Σημειώσεις Τμήμα Εφαρμογών Πληροφορικής στη Διοίκηση και Οικονομία, ΤΕΙ Πάτρας, Παράρτημα Αμαλιάδας2008.

[Μπαλής08] Παντελής Μπαλής, *Τεχνολογίες Πληροφορικής-Επικοινωνιών – Πληροφορική **VI**: Δυναμικές εφαρμογές Παγκόσμιου Ιστού*, 2008. (http://kee.ideke.edu.gr/epms/files/N22_PLHROFORIKI-6.pdf)

[Μπουντουρίδης96] Μωυσής Α. Μπουντουρίδης, *Μια γενική παρουσίαση του **Internet** και του παγκόσμιου ιστού*, Δεκέμβριος 1996. (www.math.upatras.gr/~mboudour/articles/gpipi.html)

[Μαυρονικόλας11] Μάριος Μαυρονικόλας, *Θεωρία Γράφων: Διακριτά Μαθηματικά και Μαθηματική Λογική*, Τόμος Β, 26/09/2011.

[Παπαδάκης07] Βασίλειος Παπαδάκης, *Στρατηγική των Επιχειρήσεων: Ελληνική και Διεθνής Εμπειρία*, Εκδόσεις Μπένου,2007.

[Σάμψων03] Δημήτριος Σάμψων, *Η Γλώσσα Σήμανσης XML*, Πανεπιστημιακές Σημειώσεις. Πανεπιστήμιο Πειραιώς, Τμήμα: Διδακτικής της Τεχνολογίας και Ψηφιακών Συστημάτων, Δεκέμβριος 2003. (http://www.fme.aegean.gr/sites/default/files/dsampson_xml_lectures-notes-dec2003.pdf)

[Τράκας] Ν. Δ. Τράκας, *Από που ξεκίνησε το WWW… και τι είναι το CERN*. Εθνικό Μετσόβιο Πολυτεχνείο, Τομέας Φυσικής. (www.physics.ntua.gr/POPPHYS/articles/wwwhistory.html)

[Φιλίππου] Φιλίππου Στέφανος, *Διαδίκτυο και Παγκόσμιος Ιστός (Web)*. (www.netdevelop.gr/web_design_a.pdf)

[ΨΣ2006] Ψηφιακή Σύγκλιση, περιοδικό Infosoc, *15 Χρόνια Παγκόσμιος Ιστός*, τ. 47, 2006. (www.infosoc.gr/infosoc/el-GR/grafeiotypou/infosoc_magazine/previous_editions/infosoc_magazine_2006/infosoc47/infosoc47-04.htm)