Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης

Σχολή Τεχνολογικών Εφαρμογών
Τμήμα Εφαρμοσμένης Πληροφορικής & Πολυμέσων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΤΩΝ:

ΚΑΛΟΥΤΑ ΚΑΛΛΙΟΠΗΣ 2588
&
ΚΑΛΑΦΑΤΗ ΚΩΣΤΑ 2626

ΕΜΠΛΟΥΤΙΣΜΟΣ ΔΙΕΠΑΦΩΝ ΑΝΕΥΡΕΣΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΟΙΝΟΤΙΚΕΣ ΥΠΗΡΕΣΙΕΣ ΔΙΚΤΥΩΣΗΣ

Επιβλέπων Καθηγητής : ΑΚΟΥΜΙΑΝΑΚΗΣ ΔΗΜΟΣΘΕΝΗΣ
Επιτροπή Αξιολόγησης: Ακουμιανάκης Δημοσθένης
Βιδάκης Νικόλαος
Μανιφάβας Χαράλαμπος

Ημερομηνία Παρουσίασης: 27.08.13

ΗΡΑΚΛΕΙΟ, ΑΥΓΟΥΣΤΟΣ 2013

ΕΥΧΑΡΙΣΤΊΕΣ

Ευχαριστούμε θερμά:

- Το iSTLab και όλο το επιτελείο του (Καραδημητρίου Ν., Βελλή Γ., Βαρχαλαμά Π., Βλασόπουλο Α., Κότσαλη Δ., Βλαχάκη Γ., Συντυχάκη Μ., Μιχαηλίδη Κ., Κριτσωτάκη Β., Καφούση Γ.) και ιδιαίτερα τους: κ. Ακουμιανάκη Δημοσθένη, κ. Βιδάκη Νίκο, κ. Μηλολιδάκη Γιάννη και κ. Κτιστάκη Γιώργο που από την αρχή πίστεψαν σε μας και μας στήριξαν με κάθε τρόπο σε κάθε πρόβλημα που αντιμετωπίσαμε,
- Τους γονείς μας για την κατανόηση και την αμέριστη συμπαράσταση τους όλα αυτά τα χρόνια(!),τα αδέλφια μας,
- Και τέλος τους φίλους μας που ήταν για μας πηγή έμπνευσης!!!

ABSTRACT

The aim of this thesis is to create a tool that allows users to issue queries that cross boundaries of different and often disconnected community networks to compile data and present results in a consistent and unified manner.

Initially we studied the current technological state of the art, both in the conventional Internet search (e.g. Google) as well as in popular community networks (e.g., Facebook and YouTube). The focus was on algorithms, methodologies and search architectures and the challenges posed by the compelling need to cross boundaries of different sorts, including technical and semantics/thematic. On the grounds of this analysis, we then designed a tool that allows the user to issue queries that span boundaries of virtual settlements and return results which are aggregation of data from multiple community networks.

KEYWORDS: Search engines, Community networks, semantic search.

ΣΥΝΟΨΗ

Σκοπός αυτής της πτυχιακής εργασίας είναι η δημιουργία ενός εργαλείου το οποίο θα επιτρέπει στους χρήστες να πραγματοποιούν cross-boundary ερωτήματα σε διαφορετικά και συχνά ασύνδετα μεταξύ τους, κοινοτικά δίκτυα και η παρουσίαση των αποτελεσμάτων αυτών.

Αρχικά μελετήθηκε η τρέχουσα τεχνολογική στάθμιση, τόσο στον τομέα της συμβατικής αναζήτησης στο Διαδίκτυο (π.χ. Google), όσο και στον τομέα της αναζήτησης μέσα στα κοινοτικά δίκτυα(π.χ. Facebook και YouTube). Η εργασία επικεντρώθηκε σε μεθοδολογίες αλγορίθμων και αρχιτεκτονικών αναζήτησης, καθώς επίσης και στις προκλήσεις που τίθενται από την επιτακτική ανάγκη να ξεπεραστούν διαφόρων ειδών σύνορα, συμπεριλαμβανομένων τεχνικών και σημασιολογικών/θεματικών ορίων. Με βάση τη συγκεκριμένη ανάλυση σχεδιάσαμε ένα εργαλείο το οποίο επιτρέπει στο χρήστη να συντάσσει ερωτήματα που εκτείνονται πέραν των ορίων των εικονικών οικισμών και να επιστρέφει αποτελέσματα που αποτελούν μια συνάθροιση δεδομένων από πολλαπλά κοινοτικά δίκτυα.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Μηχανές Αναζήτησης, Κοινοτικά Δίκτυα, Σημασιολογική Αναζήτηση

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

HITS	Hyperlink-Incluced Topic Search
IBM	International Business Machines Corporation
PHITS	Probabilistic Hyperlink-Incluced Topic Search
WPR	Weighted PageRank
WLRank	Weighted Links Rank
HTML	HyperText Markup Language
FTP	File Transfer Protocol
Veronica	Very Easy Rodent-Oriented Net-wide Index to Computerized Archives
Jughead	Jonzy's Universal Gopher Hierarchy Excavation & Display
CSEs	Custom Search Engine
HTTP	HyperText Transfer Protocol
URL	Uniform Resource Locator
MIME	Multipurpose Internet Mail Extensions
IP	Internet Protocol
PR	PageRank
Wanderer	Wide World Web Wanderer
SEO	Search Engine Optimization
DNS	Domain Name System
CSS	Cascading Style Sheets
SQL	Structure Query Language
GRQL	Graphical RDF Query Language
DAO	Data Access Object
YASNS	Yet Another Social Networking Service
SWED	Semantic Web Environmental Directory
RDF	Resource Description Framework
DL	Description Logic
Json	JavaScript Object Notation

ΓΛΩΣΣΑΡΙ

Meta-search engine	Μηχανές μετα-αναζήτησης
Semantic Search	Σημασιολογική αναζήτηση
Page Rank, Distance Rank, Time Rank, WPR, WL Rank, Google, Google Panda, Google Penguin, Al, HITS	Αλγόριθμος ανάλυσης συνδέσμων
Focuced graph	Εστιασμένος γράφος
Hub	Διανομέας
Authority score	Τιμές αρχής
Hub Score	Τιμές διανομέα
Filter bubble	Φυσαλίδα
CSEs	Συνεργατικές Μηχανές Αναζήτησης
Enque/Dequeue	Προσθήκη/Αφαίρεση στην ουρά
Batch mode	Δέσμες
W3Catalog	Κατάλογος Διαδικτύου
Crawlers,bots,wanderers,robots,spiders,fishes,worms	Μηχανισμοί που πραγματοποιούν αυτόματη περιήγηση στις σελίδες του Διαδικτύου
Browser	Φυλλομετρητής
Wanderer,Aliweb,Google,WebCrawler,JumpStation,Bing, DuckDuckGo,SearchTogether,Cerciamo,CLEVER	Μηχανές αναζήτησης
SEO	Βελτιστοποίηση αποτελεσμάτων σε μηχανές αναζήτησης
Document Classifier	Έγγραφο ταξινόμησης
Information Retrieval Technology	Τεχνολογία ανάκτησης πληροφοριών
Scheduler	Χρονοπρογραμματιστής
Downloader	Μέθοδος ανάκτησης σελίδων
Bandwidth	Εύρος ζώνης
Hyperlink	Υπερσύνδεσμος
Damping Factor	Συντελεστής απόσβεσης
LinkedIn, VisiblePath, Xing, Dogster, Care2, Couchsurfing, My Church, My Space, Hi5, Facebook	Κοινωνικά δίκτυα
Affinity	Συγγένεια (σχέση)
Weight	Βάρος
Time Decay	Χρονική φθορά
News Feed	Τελευταίες ειδήσεις
Search Engine Crawler	Μηχανή αναζήτησης αράχνη

Data Repository	Αποθετήριο δεδομένων
Wine Agent demonstration portal	Σημασιολογική δικτυακή πύλη
Flickr.com, Youtube.com, Last.FM	
WordNet	Λεξιλογική βάση δεδομένων στα αγγλικά
AND, OR, NOT	Λογικοί τελεστές
Perl	Γλώσσα προγραμματισμού
Onto Views	Εργαλείο δημιουργίας σημασιολογικών δικτυακών πυλών
RDF	Πλαίσιο περιγραφής πόρων
Directory Portal	Κατάλογος δικτυακής πύλης
Path Ascending	Φθίνουσα πορεία
DAO	Αντικείμενο Πρόσβασης Δεδομένων
Frontier List	Λίστα με συνδέσμους προς έρευνα

1. ΕΙΣΑΓΩΓΗ

1.1. Περίληψη

Σκοπός της πτυχιακής αυτής εργασίας ήταν η μελέτη, σχεδίαση και ανάπτυξη ενός μηχανισμού αναζήτησης δεδομένων σε κοινοτικά δίκτυα και για το λόγο αυτό μελετήθηκαν οι δυνατότητες που παρέχονται από την τάση παροχής δημόσιων APIs από τα κοινοτικά αυτά δίκτυα. Επίσης μελετήθηκαν υπάρχουσες αρχιτεκτονικές συμβατικών μηχανών αναζήτησης δεδομένων στο Διαδίκτυο όσο και μηχανών αναζήτησης σε κοινοτικά δίκτυα. Μέσα από τη μελέτη αυτή καθορίστηκαν ανάγκες σχεδίασης ενός συστήματος ειδικού σκοπού που να παρέχει τη δυνατότητα ανάκτησης και παρουσίασης δεδομένων που είναι διασκορπισμένα σε διαφορετικούς εικονικούς οικισμούς και κοινοτικά δίκτυα.

Για την υλοποίηση του πρωτότυπου συστήματος, χρησιμοποιήθηκε η τεχνολογία Servlet, που παρέχεται από τη γλώσσα προγραμματισμού Java και εκτελείται σε οποιοδήποτε Application Server. Για την ανάπτυξη της βάσης δεδομένων χρησιμοποιήθηκε το σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων (DBMS) MySQL. Για τη δημιουργία της διεπαφής του χρήστη με το σύστημα χρησιμοποιήθηκαν εργαλεία της τρέχουσας τεχνολογικής στάθμισης όπως JSP, HTML, JavaScript, jQuery, JSON, και AJAX τεχνολογίες.

1.2. Κίνητρο για τη διεξαγωγή της εργασίας

Είναι συχνά δύσκολο έως ακατόρθωτο να μπορέσει κανείς να υποβάλλει ερωτήματα των οποίων η ανάκτηση δεδομένων διατρέχει τα όρια διαφορετικών κοινοτικών δικτύων και ηλεκτρονικών υπηρεσιών. Οι δυσκολίες ποικίλουν και συνίστανται κυρίως στην ανομοιογένεια που χαρακτηρίζει τα συστήματα αυτά, τον διαφορετικό τρόπο που συχνά αυτά δομούνται και τις ιδιαιτερότητες που σχετίζονται με το εύρος και τον τύπο των δεδομένων που χειρίζονται. Η παρούσα πτυχιακή εργασία στοχεύει στην καταγραφή των προβλημάτων αυτών και την μελέτη τους από την οπτική της διασυνοριακής σχεδίασης συστημάτων. Με τον όρο διασυνοριακή σχεδίαση σηματοδοτείται μια σχεδιαστική προσέγγιση που λαμβάνει υπόψη και αντιμετωπίζει τους περιορισμούς (όρια) που είναι σύμφυτοι με λειτουργικές και μη-λειτουργικές προδιαγραφές συστημάτων οι οποίες καθορίζουν τη δυνατότητα (και κυρίως την αδυναμία) τους να χειριστούν διάσπαρτα δεδομένα που φιλοξενούνται από ηλεκτρονικές υπηρεσίες και συστήματα του Διαδικτύου. Με αυτό το σκεπτικό σχεδιάστηκε και αναπτύχθηκε ένα εργαλείο υποβολής ερωτημάτων που συλλέγει και συναθροίζει δεδομένα από πολλαπλά (διαφορετικά) κοινοτικά δίκτυα.

1.3. Σκοπός και στόχοι εργασίας

Ο σκοπός της παρούσας εργασίας είναι διττός. Καταρχήν, επιχειρείται μια αναλυτική επισκόπηση του τρόπου λειτουργίας και των χαρακτηριστικών συμβατικών μηχανών αναζήτησης καθώς και αυτών που συναντώνται σε σύγχρονες υπηρεσίες κοινοτικής δικτύωσης. Ειδικότερα, μελετάται πως ακριβώς λειτουργεί ο μηχανισμός αναζήτησης σε υπηρεσίες κοινοτικής δικτύωσης όπως το Facebook και το YouTube, τι είδους αποτελέσματα παράγονται και πως αυτά αναπαριστώνται. Έμφαση δίδεται στα δομικά χαρακτηριστικά των στοιχείων που συνθέτουν το αποτέλεσμα της αναζήτησης και του τρόπου που αυτά τα

χαρακτηριστικά (π.χ., ιδιοκτήτης, κατηγορία, τύπος τεχνουργήματος που αναφέρεται, κλπ) αποκτούν κοινοτικά χαρακτηριστικά.

Στη συνέχεια παρουσιάζεται η σχεδίαση και υλοποίηση ενός συστατικού λογισμικού με αυξημένες δυνατότητες αλληλεπίδρασης που εμπλουτίζει το προσφερόμενο αποτέλεσμα μιας αναζήτησης. Ενδεικτικό παράδειγμα τέτοιου εμπλουτισμού είναι η διασύνδεση ενός τμήματος του αποτελέσματος μιας αναζήτησης με άλλα συστήματα ή περιβάλλοντα στα οποία υπάρχουν ίχνη του τεχνουργήματος που έχει επιλεγεί από τη λίστα των αποτελεσμάτων αναζήτησης. Άλλες δυνατότητες που μελετούνται περιλαμβάνουν την προσβασιμότητα του συστατικού με εναλλακτικές μεθόδους π.χ., φωνητικές εντολές, καθώς και η διασύνδεση του με άλλα κοινοτικά δίκτυα δηλ. αν η αναζήτηση έγινε στο Facebook να μπορεί να γίνει ιχνηλάτιση δεδομένων για ένα στοιχείο των αποτελεσμάτων σε άλλα δίκτυα όπως το YouTube.

1.4. Δομή εργασίας

Η αναφορά περιλαμβάνει έξι κεφάλαια. Στο **κεφάλαιο 2** παρουσιάζονται το αντικείμενο μελέτης της παρούσας εργασίας και τα θεωρητικά πλαίσια αναφοράς από τα οποία αντλεί τη θεματολογία του. Στο **κεφάλαιο 3** συνοψίζονται οι βασικές αρχές που αφορούν την αναζήτηση δεδομένων, εξειδικεύοντας τους επικρατέστερους αλγόριθμους αναζήτησης, καθώς και τα μοντέλα και τις θεωρίες που χρησιμοποιούνται μέχρι σήμερα στη συμβατική αναζήτηση. Στο **κεφάλαιο 4** μελετάται η αναζήτηση δεδομένων σε ευρέως γνωστά ηλεκτρονικά κοινωνικά δίκτυα και ιστοτόπους κοινωνικής διδκτύωσης. Στο **κεφάλαιο 5** συνοψίζονται οι βασικές απαιτήσεις ενός συστήματος που υποστηρίζει θεματική αναζήτηση διατρέχοντας ιστοτόπους κοινωνικής δικτύωσης, καθώς και η προσέγγιση που υιοθετήθηκε για την ανάπτυξη ενός πρωτοτύπου διασυνοριακής ανάκτησης δεδομένων από το Facebook και το YouTube. Για το σκοπό αυτό επιλέχτηκαν συγκεκριμένες περιπτώσεις χρήσης που αναλύονται και τεκμηριώνονται διεξοδικά. Τέλος, στο **κεφάλαιο 6** παρουσιάζονται τα συμπεράσματα που προκύπτουν κατά την υλοποίηση της εργασίας και παρατίθενται οι επεκτάσεις και οι μελλοντικές προτάσεις για την εξέλιξή της.

2. ΜΕΘΟΔΟΛΟΓΙΑ

2.1. Προσέγγιση και μέθοδος

Η μέθοδος στην οποία βασίζεται η παρούσα εργασία είναι η μελέτη περίπτωσης χρήσης (case study) όπου μια νέα υπηρεσία αναζήτησης αξιοποιείται για τις ανάγκες μιας ειδικής κατηγορίας πληροφοριακών συστημάτων που έχουν αμιγώς κοινοτικά χαρακτηριστικά. Μέχρι σήμερα, η συντριπτική πλειοψηφία των μηχανών αναζήτησης αξιοποιούν συγκεκριμένους μηχανισμούς και τεχνικές για τον προσδιορισμό κριτηρίων αναζήτησης (π.χ. λέξεις κλειδιά) και βάσει αυτών ιχνηλατούν διάσπαρτους εξυπηρετητές προκειμένου να ανακτηθεί το πλήθος των ιστοσελίδων με συναφή περιεχόμενα. Παρότι, τα εργαλεία αυτά έχουν τύχει ιδιαίτερης αναγνώρισης στη σχετική βιβλιογραφία, υπάρχουν αντικειμενικοί περιορισμοί που εξακολουθούν να υφίστανται και οι οποίοι καθορίζουν την ποιότητα των αποτελεσμάτων που επιστρέφονται στον χρήστη. Τέτοιοι περιορισμοί αφορούν την αξιοποίηση πληροφορίας που αφορά είτε τον ίδιο τον χρήστη είτε το γνωστικό πεδίο στο οποίο δραστηριοποιείται. Παραδείγματος χάριν, τα αποτελέσματα μέσω μιας μηχανής αναζήτησης θα μπορούσαν να βελτιωθούν αν υπήρχε δυνατότητα ο ίδιος ο χρήστης να προσδιορίζει (με την πάροδο του χρόνου) τυχόν εξειδικευμένα ενδιαφέροντα, τις πηγές που επιθυμεί να αναζητηθούν πρώτα ή ακόμη τυχόν άλλους χρήστες που έχουν παρόμοια ενδιαφέροντα ή εμπλέκονται στο ίδιο γνωστικό πεδίο και ανακτούν παρόμοιες πληροφορίες. Καθένα από αυτά τα κριτήρια θα μπορούσαν να εξειδικευτούν με συγκεκριμένους μηχανισμούς που είτε εμπλουτίζουν τη διαδικασία αναζήτησης πληροφοριών είτε την αναγάγουν σε μια ενεργητικότερη διαδικασία διαχείρισης κοινοτικής γνώσης και κεφαλαίου. Ωστόσο η διαδικασία εξειδίκευσης δεν είναι ακόμη σαφής και δεν μπορεί ακόμη να προσδιοριστεί σχεδιαστικά με μονοσήμαντο τρόπο.

Για το λόγο αυτό η εργασία αξιοποιεί την μέθοδο της περίπτωσης χρήσης που επιτρέπει αφενός την εστίαση της προσπάθειας (που διαφορετικά θα ήταν χαώδης) σε ένα πεδίο γνώσης και αφετέρου την πιλοτική ανάπτυξη ενός συστατικού λογισμικού (υπηρεσία τύπου μηχανής αναζήτησης) που θα επιτρέπει την συγκριτική διερεύνηση των παραγόντων που επηρεάζουν το αποτέλεσμα μιας αναζήτησης που διεξάγεται έχοντας κατά νου την κοινοτική συναναστροφή του χρήστη σε κοινωνικά δίκτυα.

2.2. Πεδίο αναφοράς

Ως πεδίο αναφοράς της συγκεκριμένης πτυχιακής εργασίας, επιλέχθηκε ο τομέας της βιολογικής καλλιέργειας (Organic Farming), με σκοπό τη χρήση των εργαλείων, που αναπτύσσονται σε αυτήν, σε μελλοντικές ερευνητικές και μη εργασίες. Χαρακτηριστικό παράδειγμα τέτοιας εργασίας αποτελεί το Biodrasis.

2.3. Ψηφιακά ίχνη και υπηρεσίες κοινωνικής δικτύωσης

Στην πτυχιακή εργασία επιλέχθηκαν τα συγκεκριμένα κοινωνικά δίκτυα (Facebook, YouTube, Twitter), αρχικά λόγω του ότι είναι τα δημοφιλέστερα κοινωνικά δίκτυα τη χρονική περίοδο συγγραφής της παρούσας εργασίας. Πέραν της δημοτικότητας των συγκεκριμένων κοινωνικών δικτύων κάθε ένα από αυτά επιλέχθηκε για ένα σύνολο γνωρισμάτων τους, τα οποία είναι χρήσιμα για μελλοντικές εργασίες.

Facebook

- Παρέχει ένα σύνολο από κοινωνικές διαδράσεις (interactions) ανάμεσα στους χρήστες (likes,comments,wall posts).
- Υποστηρίζει τη δυνατότητα δημιουργίας κοινοτήτων υπό τη μορφή ομάδων και γεγονότων (groups,events).
- Παρέχει τη δυνατότητα ανταλλαγής δεδομένων σε πολλές διαφορετικές μορφές (notes,photos,videos etc.).

YouTube

- Αποτελεί τη μεγαλύτερη δικτυακή πλατφόρμα παροχής βίντεο.
- Υποστηρίζει τη δυνατότητα διαχωρισμού των βίντεο σε διακριτές κατηγορίες.
- Παρέχει τη δυνατότητα δημιουργίας καναλιών των οποίων τα ενδιαφέροντα και τα περιεχόμενα είναι διαδραστικά και μπορούν να καθοριστούν από τους χρήστες.

Twitter

- Παρέχει τη δυνατότητα επικοινωνίας και ανταλλαγής πληροφοριών σε πραγματικό χρόνο. κυρίως μέσω του μεγέθους του tweet.

2.4. Περιορισμοί

Στο πλαίσιο της έρευνας που εκπονήθηκε κατά τη διάρκεια της πτυχιακής εργασίας παρατηρήσαμε ότι η επιλογή του πεδίου αναφοράς, όπου για την εργασία μας είναι η βιολογική καλλιέργεια, επηρεάζει άμεσα τόσο την ποσότητα όσο και την ποιότητα της πληροφορίας που εξάγουμε. Αυτό οφείλεται στο γεγονός ότι τα κοινωνικά δίκτυα που επιλέξαμε είναι γενικού ενδιαφέροντος και δεν είναι εξειδικευμένα σε κάποιον συγκεκριμένο τομέα, επομένως, η πληροφορία που ανταλλάσσεται ανάμεσα στους χρήστες είναι πληροφορία που ανταλλάσσεται ανάμεσα σε δύο μέσους ανθρώπους και όχι εξειδικευμένη επιστημονική ή τεχνική γνώση και πληροφορία.

Επίσης, παρατηρήθηκε ότι άλλα κοινωνικά δίκτυα πιο εξειδικευμένα σε διάφορους τομείς παρέχουν μεγαλύτερο όγκο πληροφορίας, στοχευμένης στο πεδίο αναφοράς που χρησιμοποιούμε σαν παράδειγμα. Παρόλα τα προβλήματα που παρατηρήθηκαν, επιλέξαμε τα συγκεκριμένα κοινωνικά δίκτυα λόγω των δυνατοτήτων που αναφέρθηκαν παραπάνω οι οποίες μας επιτρέπουν να χρησιμοποιήσουμε τα εργαλεία που προέκυψαν σε περισσότερα από ένα πεδία αναφοράς, χωρίς σημαντικές αναφορές στο λογισμικό και τη φιλοσοφία τους και επειδή μπορούν να προσαρμοστούν και να υποστηρίξουν πληροφορία που ενδιαφέρει εξειδικευμένους χρήστες χρησιμοποιώντας τις δυνατότητές τους.

3. ΤΕΧΝΟΛΟΓΙΚΉ ΣΤΑΘΜΙΣΗ

Η παρούσα πτυχιακή εργασία αποτελεί ένα συνδυασμό τεχνολογιών στα πεδία της αναζήτησης, της ανάκτησης δεδομένων και της τεχνολογίας των κοινωνικών δικτύων. Έτσι παρακάτω παρουσιάζεται το "State of the Art" για τις παραπάνω τεχνολογίες.

3.1. Ανάκτηση δεδομένων στο Διαδίκτυο [1]

Η ανάκτηση δεδομένων στο Διαδίκτυο είναι η λειτουργία που επιτελείται από μια εξειδικευμένη εφαρμογή που επιτρέπει την αναζήτηση κειμένων και αρχείων στο Διαδίκτυο. Αποτελείται από ένα πρόγραμμα υπολογιστή που βρίσκεται σε έναν ή περισσότερους υπολογιστές στους οποίους δημιουργεί μια βάση δεδομένων με τις πληροφορίες που συλλέγει από το Διαδίκτυο, και το διαδραστικό περιβάλλον που εμφανίζεται στον τελικό χρήστη ο οποίος χρησιμοποιεί την εφαρμογή από άλλον υπολογιστή συνδεμένο στο Διαδίκτυο.

3.1.1. Ιστορικό και τα πρώτα εργαλεία αναζήτησης [2]

Το πρώτο εργαλείο που χρησιμοποιήθηκε για αναζήτηση στο Διαδίκτυο ήταν ο Archie (http://archie.icm.edu.pl/archie-adv_eng.html). Το όνομα προέρχεται από τη λέξη "Archive" χωρίς το "v". Οι δημιουργοί του ήταν οι Alan Emtage, Bill Heelan και J.Peter Deutsch το 1990. Το πρόγραμμα κατέβαζε τα ονόματα των αρχείων από δημόσιες FTP σελίδες, δημιουργώντας έτσι μια βάση δεδομένων με ονόματα αρχείων. Το Archie δεν παρήγαγε ευρετήρια για τα περιεχόμενα, αφού ο όγκος των δεδομένων ήταν τόσο μικρός που η αναζήτηση μπορούσε να γίνει χειροκίνητα.

Η δημιουργία του Gopher οδήγησε σε δύο νέα προγράμματα αναζήτησης, τη Veronica και τον Jughead. Όπως και το Archie, τα προγράμματα αυτά αναζητούσαν ονόματα αρχείων που αποθηκεύονταν σε συστήματα ευρετηρίων Gopher. Η Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) παρείχε τη δυνατότητα αναζήτησης βάση μιας λέξης κλειδί σε όλους τους Gopher καταλόγους. Το Jughead (Jonzy's Universal Gopher Hierarchy Excavation And Display) ήταν ένα εργαλείο για την απόκτηση πληροφοριών από συγκεκριμένους Gopher servers.

Ο Oscar Nierstrasz δημιούργησε κάποια Perl Scripts, τα οποία αντέγραφαν ειδικούς καταλόγους σελίδων και τους τροποποιούσαν ώστε να αναπαριστώνται με συγκεκριμένο τρόπο. Αυτό έθεσε τις βάσεις για το W3Catalog, την πρώτη πρωτόγονη μηχανή αναζήτησης του Διαδικτύου, που παρουσιάστηκε στις 2 Σεπτεμβρίου 1993. Τον Ιούνιο του 1993, ο Matthew Gray δημιούργησε το πρώτο web robot, το World Wide Web Wanderer και το χρησιμοποίησε ως ένα ευρετήριο το οποίο ονόμασε 'Wandex'. Ο σκοπός του Wanderer ήταν ο υπολογισμός του μεγέθους του Διαδικτύου. Δεύτερη μηχανή αναζήτησης του Διαδικτύου ήταν το Aliweb που εμφανίστηκε το Νοέμβριο του 1993. Το Aliweb δεν χρησιμοποιούσε web robot αλλά βασιζόταν στις ιστοσελίδες να το ειδοποιήσουν για την ύπαρξή τους. Το JumpStation χρησιμοποιούσε ένα web robot για να εντοπίζει τις ιστοσελίδες και να δημιουργεί τα ευρετήρια του, και χρησιμοποιούσε μια φόρμα σαν διεπαφή. Έτσι ήταν το πρώτο εργαλείο που συνδύαζε τις τρεις βασικές λειτουργίες της μηχανής αναζήτησης[1].Λόγω των περιορισμένων πόρων της πλατφόρμας που βασιζόταν, το πρόγραμμα περιοριζόταν στην αποθήκευση των τίτλων των ιστοσελίδων που συναντούσε ο Crawler.

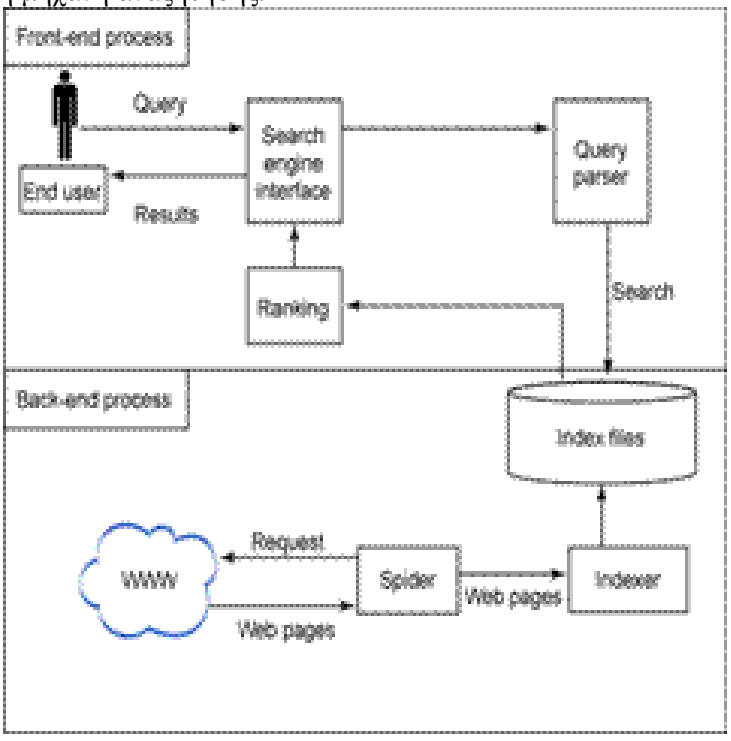
Η μηχανή αναζήτησης WebCrawler, σε αντίθεση με τους προκατόχους της, μπορούσε να αναζητήσει οποιαδήποτε λέξη σε οποιαδήποτε σελίδα, κάτι που υιοθετήθηκε από όλες τις μετέπειτα μηχανές αναζήτησης. Το 2000 η μηχανή αναζήτησης Google, ήταν η μηχανή αναζήτησης με τα καλύτερα αποτελέσματα λόγω της χρήσης του αλγορίθμου

[1] Crawling, Indexing, Searching

PageRank. Οι σύγχρονες μηχανές αναζήτησης χωρίζονται σε τρείς μεγάλες κατηγορίες: τις μηχανές αναζήτησης, τις συνεργατικές μηχανές αναζήτησης και τις Metasearch engines.

3.1.1.1. Μηχανές αναζήτησης [3]

Η αρχιτεκτονική μιας τυπικής μηχανής αναζήτησης παρουσιάζεται στην Εικόνα 1. Οι μηχανές αναζήτησης αποθηκεύουν πληροφορίες για τις ιστοσελίδες, τις οποίες συλλέγουν απευθείας από την HTML της σελίδας. Οι σελίδες αυτές συλλέγονται από έναν web Crawler, έναν αυτοματοποιημένο browser που ακολουθεί όλους τους συνδέσμους στον ιστότοπο. Έπειτα, τα περιεχόμενα κάθε σελίδας αναλύονται για να καθοριστεί το πώς πρέπει να τοποθετηθούν σε ευρετήρια. Δεδομένα για κάθε σελίδα αποθηκεύονται σε μια βάση δεδομένων για χρήση σε επόμενα ερωτήματα. Ο σκοπός των ευρετηρίων είναι να επιτρέπουν στις πληροφορίες να εντοπίζονται όσο το δυνατόν γρηγορότερα. Κάποιες μηχανές αναζήτησης (π.χ. Google), αποθηκεύουν όλη τη σελίδα ή τμήμα της, όπως επίσης και πληροφορίες σχετικά με τις σελίδες, ενώ κάποιες άλλες (π.χ. AltaVista), αποθηκεύουν κάθε λέξη, κάθε σελίδας που συναντούν. Κάθε αποθηκευμένη σελίδα διατηρεί πάντα το κείμενο με το οποίο έγινε η αναζήτηση, κάτι που είναι πολύ χρήσιμο όταν τα περιεχόμενα της σελίδας αλλάζουν και τα ορίσματα της αναζήτησης δεν υπάρχουν πια στη σελίδα. Η χρήση αυτής της μεθόδου εξασφαλίζει ότι ικανοποιείται η προσδοκία του χρήστη να περιέχονται τα δεδομένα της αναζήτησης στα αποτελέσματα. Αυτό ικανοποιεί την αρχή της ελάχιστης έκπληξης2, αφού ο χρήστης κανονικά περιμένει τα δεδομένα της αναζήτησης να υπάρχουν στις σελίδες που επιστρέφει η μηχανή αναζήτησης.



Εικόνα 1: Αρχιτεκτονική μιας συμβατικής μηχανής αναζήτησης

Όταν ο χρήστης διατυπώνει ένα ερώτημα στη μηχανή αναζήτησης, η μηχανή εξετάζει τους καταλόγους της και παρέχει μια λίστα με τις σελίδες που ταιριάζουν περισσότερο στα κριτήρια της αναζήτησης, συνήθως με μια μικρή περίληψη των περιεχομένων που περιέχει, τον τίτλο του εγγράφου και κάποιες φορές τμήματα της σελίδας.

Οι περισσότερες μηχανές αναζήτησης υποστηρίζουν τη χρήση των λογικών τελεστών AND, OR και NOT για την περαιτέρω εξειδίκευση των ερωτημάτων του χρήστη. Η μηχανή

[2] Principle of least astonishment.

ψάχνει για λέξεις ή φράσεις ακριβώς όπως εισάγονται. Κάποιες μηχανές αναζήτησης παρέχουν ένα προηγμένο χαρακτηριστικό, το οποίο ονομάζεται αναζήτηση εγγύτητας, που επιτρέπει στους χρήστες να ορίσουν την απόσταση ανάμεσα στους όρους της αναζήτησης. Επίσης παρέχεται η δυνατότητα η έρευνα να περιλαμβάνει τη χρήση στατιστικής ανάλυσης σε σελίδες που περιέχουν τις ζητούμενες λέξεις ή φράσεις. Τέλος οι αναζητήσεις σε φυσική γλώσσα επιτρέπουν στο χρήστη να διατυπώσει μία ερώτηση με τον ίδιο τρόπο που θα ρωτούσε κι έναν άνθρωπο (π.χ. ask.com).

ΤΡΟΠΟΠΟΙΗΜΕΝΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΦΥΣΑΛΙΔΕΣ [3] [4]

Πολλές μηχανές αναζήτησης όπως οι Google και Bing, παρέχουν εξατομικευμένα αποτελέσματα, βάση του ιστορικού του χρήστη. Αυτό οδηγεί σε ένα φαινόμενο το οποίο ονομάζεται φυσαλίδα (filter bubble). Ο όρος αναφέρεται στο φαινόμενο κατά το οποίο οι ιστοσελίδες χρησιμοποιούν αλγορίθμους για να υποθέσουν και να επιλέξουν τί πληροφορίες θα ήθελε να δει ο χρήστης βάση πληροφοριών σχετικά με το χρήστη (πληροφορίες όπως τοποθεσία, προηγούμενη συμπεριφορά και ιστορικό αναζήτησης). Σαν αποτέλεσμα, οι ιστοσελίδες τείνουν να προβάλουν μόνο πληροφορίες που συμπίπτουν με την προηγούμενη άποψη του χρήστη, απομονώνοντας το χρήστη σε μια φυσαλίδα που αποκλείει πληροφορίες αντίθετες με τις απόψεις του.

Από την στιγμή που εντοπίστηκε το πρόβλημα, πολλές ανταγωνιστικές μηχανές αναζήτησης προσπάθησαν να το αποφύγουν, με το να μην αποθηκεύουν πληροφορίες για το χρήστη. Ένα παράδειγμα είναι η μηχανή αναζήτησης DuckDuckGo.

3.1.1.2 Συνεργατικές μηχανές αναζήτησης [5]

Οι συνεργατικές μηχανές αναζήτησης (CSEs) αποτελούν μια αναδυόμενη τάση για την αναζήτηση στο Διαδίκτυο και επιχειρηματικές αναζητήσεις στα τοπικά intranets. Οι CSEs επιτρέπουν στους χρήστες να συντονίσουν τις προσπάθειές τους για την ανάκτηση πληροφοριών και να μοιράζονται πηγές πληροφοριών. Επίσης, αυτού του τύπου η αναζήτηση επιτρέπει σε πιο έμπειρους χρήστες να καθοδηγούν λιγότερο έμπειρους στις αναζητήσεις τους. Αυτό γίνεται με την παροχή ερωτημάτων, την πρόσθεση σχολίων και απόψεων, την βαθμολόγηση των αποτελεσμάτων αναζήτησης και την σύνδεση παλαιότερων (επιτυχημένων) αναζητήσεων με αναζητήσεις με ίδιο ή παρόμοιο πεδίο ορισμάτων και αποτελεσμάτων.

ΜΟΝΤΕΛΑ ΣΥΝΕΡΓΑΣΙΑΣ

Οι συνεργατικές μηχανές αναζήτησης μπορούν να ταξινομηθούν βάση διαφορετικών κριτηρίων όπως η μέθοδος συνεργασίας (προκύπτουσα ή σαφής), ο συγχρονισμός [6], το βάθος της διαμεσολάβησης [7], ο καταμερισμός της εργασίας και η ανταλλαγή γνώσεων.

ΠΡΟΚΥΠΤΟΥΣΑ ΚΑΙ ΣΑΦΗΣ ΣΥΝΕΡΓΑΣΙΑ

Η προκύπτουσα συνεργασία χρησιμοποιείται από συστήματα φιλτραρίσματος και συστάσεων, στα οποία το σύστημα ταξινομεί και αναγνωρίζει παρόμοιες ανάγκες πληροφόρησης. Συστήματα που εμπίπτουν σε αυτή την κατηγορία αναγνωρίζουν χρήστες με παρόμοιες ανάγκες και προτείνουν ερωτήματα και συνδέσμους προς τους χρήστες.

Η σαφής συνεργασία χρησιμοποιείται όταν οι χρήστες μοιράζονται προσυμφωνημένες ανάγκες πληροφόρησης και εργάζονται από κοινού για την επίτευξη του στόχου τους. Το πιο επιφανές παράδειγμα αυτής της κατηγορίας είναι το SearchTogether [8] που εκδόθηκε το 2007. Η μηχανή προσφέρει μία διεπαφή η οποία συνδυάζει αποτελέσματα

[3] Bubble Filtering

αναζήτησης από κλασσικές μηχανές αναζήτησης και μία εφαρμογή chat για την ανταλλαγή ερωτημάτων και συνδέσμων.

Όσον αφορά το βάθος της διαμεσολάβησης και τον καταμερισμό εργασίας, η μηχανή αναζήτησης Cerciamo [7] αποτελεί ενδεικτικό παράδειγμα. Η μηχανή υποστηρίζει σαφή συνεργασία με διακριτούς ρόλους όπου ένας χρήστης μπορεί να επικεντρωθεί στον εντοπισμό εγγράφων, ενώ κάποιος άλλος κρίνει την καταλληλόλητα των εγγράφων που εντοπίζονται από τον πρώτο χρήστη.

3.1.1.3 Μηχανές μετα-αναζήτησης [2]

Είναι αλήθεια πως οι μηχανές αναζήτησης επιστέφουν αρκετό υλικό ως αποτέλεσμα μιας αίτησης για αναζήτηση πληροφορίας στο Διαδίκτυο από το χρήστη. Εντούτοις, για απολύτως περιεκτικά αποτελέσματα στο κυνήγι της πληροφορίας θα πρέπει κανείς να λάβει υπόψη του τις λεγόμενες μηχανές μετα-αναζήτησης. Η μετα-αναζήτηση είναι ένας τύπος αναζήτησης που εδράζεται σε επιμέρους αποτελέσματα άλλων (συμβατικών) μηχανών αναζήτησης. Στην πραγματικότητα αυτό που υποστηρίζεται είναι ένα είδος δρομολόγησης ερωτημάτων σε πολλές μηχανές αναζήτησης ταυτόχρονα για ανάκτηση πληροφορίας.

ΒΑΣΙΚΗ ΛΕΙΤΟΥΡΓΙΑ [2]

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. Με το πάτημα του κουμπιού για την έναρξη της αναζήτησης, η μετα-μηχανή στέλνει το ερώτημα του χρήστη ταυτόχρονα σε πολλές, ξεχωριστές, απλές μηχανές αναζήτησης και συνεπώς στις βάσεις δεδομένων με web σελίδες αυτών. Μέσα σε λίγα δευτερόλεπτα, η μετα-μηχανή επιστρέφει στο χρήστη τα αποτελέσματα που συλλέγονται από όλες τις απλές μηχανές αναζήτησης στις οποίες διαβίβασε το ερώτημα του χρήστη.

Μια πιο πολύπλοκη μηχανή μετα-αναζήτησης επιτρέπει στον χρήστη να καθορίσει πολύπλοκες παραμέτρους με βάση τις οποίες επιθυμεί να γίνει η αναζήτηση πληροφορίας σχετικά με το συγκεκριμένο θέμα που τον ενδιαφέρει. Για παράδειγμα, ο χρήστης είναι δυνατό να καθορίσει το χρονικό διάστημα για το οποίο επιθυμεί να γίνει η αναζήτηση αυτή. Μια τέτοια λειτουργία υποστηρίζεται και από τις απλές μηχανές αναζήτησης. Επίσης, ακριβώς όπως και στις απλές μηχανές αναζήτησης, είναι δυνατή στις μηχανές μετα-αναζήτησης η χρήση των Boolean τελεστών AND, OR και NOT, καθώς και του τελεστή προσέγγισης NEAR, στη διατύπωση των ερωτημάτων από το χρήστη.

Οι μηχανές μετα-αναζήτησης δε διαθέτουν δικές τους βάσεις δεδομένων με web σελίδες, όπως συμβαίνει στις απλές μηχανές. Αυτό που κάνουν είναι να διαβιβάζουν τα ερωτήματα των χρηστών στις βάσεις δεδομένων των απλών μηχανών αναζήτησης. Μια μετα-μηχανή αναζήτησης απαιτεί περισσότερο χρόνο για την εκτέλεση ενός ερωτήματος καθώς θα πρέπει να πραγματοποιήσει ελέγχους σε πολλές άλλες μηχανές αναζήτησης για το ερώτημα αυτό. Το σημείο στο οποίο υπερέχουν οι μηχανές μετα-αναζήτησης έναντι των απλών μηχανών αναζήτησης είναι ότι συχνά επιστρέφουν απαντήσεις σε σχετικά ασαφείς ερωτήσεις του χρήστη που μια απλή μηχανή μπορεί να «χάσει».

Σήμερα υπάρχουν τρείς τύποι μηχανών μετα-αναζήτησης:

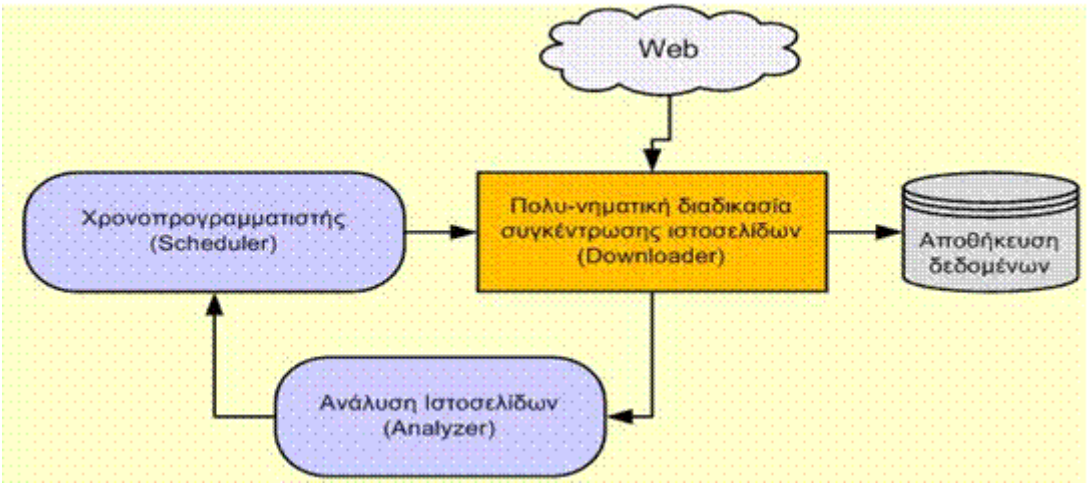
- Εργαλεία για ανάκτηση πληροφορίας(digging) σε πολλές πηγές, που προσφέρουν πολλές δυνατότητες για εύρεση αυτού που ζητά ο χρήστης μέσα σε αποτελέσματα αναζήτησης. Αυτά τα εργαλεία είναι κατάλληλα για ερευνητές που επιζητούν μια εις βάθος ανάκτηση πληροφοριών σχετικά με ένα θέμα (π.χ. Meta Crawler).
- Μηχανές μετα-αναζήτησης που πραγματοποιούν πολύπλοκες αναζητήσεις, ενοποιούν τα αποτελέσματα καλά, απαλείφουν τις διπλο-εμφανίσεις αποτελεσμάτων και προσφέρουν επιπρόσθετες επιλογές, όπως έξυπνη

ταξινόμηση ή ομαδοποίηση κατά θέματα των αποτελεσμάτων αναζήτησης(π.χ. SavvySearch, Clusty).

- Μηχανές μετα-αναζήτησης που «ψάχνουν» σε πολλά μέρη και επιστρέφουν αποτελέσματα χωρίς τις επιλογές που αναφέρθηκαν παραπάνω. Σε αυτή την κατηγορία ανήκουν πολλές μηχανές μετα-αναζήτησης (π.χ.Dogpile).

3.1.2. Η τεχνική crawling [9]

Ένας Web Crawler είναι ένα πρόγραμμα το οποίο περιπλανιέται στο Διαδίκτυο με ένα μεθοδικό, αυτοματοποιημένο τρόπο. Η διαδικασία αυτή ονομάζεται Web Crawling ή spidering. Πολλές ιστοσελίδες, και κυρίως μηχανές αναζήτησης, χρησιμοποιούν το spidering σαν μέθοδο για την παροχή ενημερωμένων δεδομένων. Οι Web Crawlers χρησιμοποιούνται κυρίως για να δημιουργούν ένα αντίγραφο όλων των σελίδων που έχουν επισκεφθεί για μετέπειτα επεξεργασία από τη μηχανή αναζήτησης. Η γενική αρχιτεκτονική ενός Crawler (βλέπε Εικόνα 2) αποτελείται από τέσσερις κύριες λειτουργικές μονάδες [10]: τον χρονοπρογραμματιστή, την ουρά, τη μέθοδο ανάκτησης και τη μονάδα αποθήκευσης. Η μέθοδος ανάκτησης περιλαμβάνει μια μονάδα αντιστοίχησης DNS, μια μονάδα ανάκτησης σελίδων μέσω του πρωτοκόλλου HTTP, και μία μονάδα συντακτικής ανάλυσης των HTML σελίδων για την εξαγωγή υπερσυνδέσμων και άλλων στατιστικών στοιχείων.



Εικόνα 2: Αρχιτεκτονική ενός Crawler

Αρχίζοντας με ένα seed set από URLs εκκίνησης, η μέθοδος ανάκτησης ανακτά τις αντίστοιχες σελίδες, εξάγει όλα τα URLs που περιέχονται σε αυτές με τη μορφή υπερσυνδέσεων και τα στέλνει στον χρονοπρογραμματιστή που τα προσθέτει στην ουρά, εφόσον δεν υπάρχουν ήδη (λειτουργία enqueue). Στη συνέχεια ο χρονοπρογραμματιστής επιλέγει (με κάποια σειρά) το επόμενο URL προς ανάκτηση (λειτουργία dequeue) και το στέλνει στη μέθοδο ανάκτησης. Το URL δεν διαγράφεται από την ουρά, αλλά μαρκάρεται ως “removed”. Μία άλλη μέθοδος είναι η διατήρηση μίας δεύτερης ουράς που περιλαμβάνει τα URLs που έχουν ήδη συναντηθεί, τα οποία διαγράφονται από την ουρά κατά το dequeue. Η διαδικασία επαναλαμβάνεται μέχρις ότου ο Crawler να συλλέξει έναν αρκετά μεγάλο αριθμό σελίδων. Το κριτήριο τερματισμού ποικίλλει ανάλογα με την περίπτωση και το σκοπό που ο Crawler καλείται να εκπληρώσει. Η μέθοδος ανάκτησης, εκτός από την ανάκτηση και την αποθήκευση των σελίδων, μπορεί παράλληλα να δημιουργεί ένα ευρετήριο από URLs ή να εξάγει αθροιστικά στατιστικά στοιχεία από το περιεχόμενο των σελίδων.

Η λειτουργία dequeue του χρονοπρογραμματιστή καθορίζει τη σειρά επιλογής των URLs που ανακτώνται, δηλαδή την πολιτική προτεραιότητας που υιοθετεί ο Crawler. Στην απλούστερη περίπτωση η επιλογή είναι η αναζήτηση ¨Πρώτα κατά πλάτος¨. Σε κάθε περίπτωση ο Crawler θα πρέπει να αποφεύγει να υπερφορτώνει έναν μοναδικό διακομιστή με αιτήσεις και θα πρέπει να περιμένει για ένα χρονικό διάστημα (συνήθως μεγαλύτερο των

30δευτερολέπτων) ανάμεσα στις διαδοχικές αιτήσεις προς τον ίδιο διακομιστή. Έτσι σε κάθε μέθοδο ανάκτησης ανατίθεται ένα πλήθος συνδέσεων προς διαφορετικούς διακομιστές στους οποίους μπορούν να γίνονται αιτήσεις ταυτόχρονα.

Σε πολλές περιπτώσεις ο Crawler χρησιμοποιεί πολλαπλές μεθόδους ανάκτησης ταυτόχρονα, με κάθε μία από αυτές να εκτελείται σε διαφορετικό σύστημα. Κατά αυτόν τον τρόπο μεγιστοποιείται ο ρυθμός συλλογής σελίδων και είναι δυνατή η κάλυψη μεγαλύτερου εύρους σελίδων του Ιστού σε μικρότερο χρονικό διάστημα. Ο χρονοπρογραμματιστής αποτελεί την κεντρική διεργασία που συντονίζει τις μεθόδους ανάκτησης: προωθεί νέα URLs στις κατάλληλες μεθόδους ανάκτησης οι οποίοι με τη σειρά τους επιστρέφουν στο χρονοπρογραμματιστή τα URLs που ανακαλύπτουν στις σελίδες που ανακτούν. Η παράλληλη αυτή αρχιτεκτονική εισάγει ένα πρόσθετο κόστος επικοινωνίας μεταξύ των πολλαπλών διαδικασιών και του χρονοπρογραμματιστή και για να ελαχιστοποιηθεί αυτό η ανταλλαγή δεδομένων γίνεται περιοδικά με μαζικό τρόπο (batch mode).

Οι λεπτομέρειες της υλοποίησης των Crawlers συνήθως παραμένουν κρυφές και αποτελούν επιχειρηματικά μυστικά, καθώς αποτελούν βασικά συστατικά των μηχανών αναζήτησης. Εφόσον οι Crawlers ακολουθούν κάποια στρατηγική που δίνει προτεραιότητα σε σελίδες που ικανοποιούν συγκεκριμένα κριτήρια, η δημοσίευση τέτοιων λεπτομερειών θα επέτρεπε σε κακόβουλους χρήστες να παραποιήσουν τις σελίδες τους ώστε να αυξήσουν την πιθανότητα επίσκεψης από τους Crawlers. Στη συνέχεια παρουσιάζονται συνοπτικά οι Crawlers των οποίων οι υλοποιήσεις έχουν δημοσιευθεί ως ένα επαρκές επίπεδο λεπτομέρειας.

3.1.3. SIMPLE Web Crawler [11]

Ένας Web Crawler είναι ένα πρόγραμμα που περιηγείται στο Διαδίκτυο με έναν μεθοδικό και αυτόματο τρόπο. Όταν πρωτοεμφανίστηκαν αυτού του είδους τα προγράμματα ονομάζονταν επίσης wanderers, robots, spiders, fishes και worms. Αρχικά, βασικό κίνητρο στη σχεδίαση των Web Crawlers ήταν η ανάκτηση ιστοσελίδων και η προσθήκη αυτών ή των απεικονίσεών τους σε μια τοπική μονάδα αποθήκευσης. Ένα τέτοιο repository μπορεί στη συνέχεια να εξυπηρετήσει συγκεκριμένες ανάγκες εφαρμογών όπως αυτών των Διαδικτυακών μηχανών αναζήτησης. Αν το Διαδίκτυο ήταν μια στατική συλλογή σελίδων, θα είχε μικρή μακροχρόνια χρήση του crawling. Από τη στιγμή που όλες οι σελίδες θα είχαν «φορτωθεί» σε ένα repository (όπως η βάση δεδομένων μιας μηχανής αναζήτησης), δε θα υπήρχε πλέον ανάγκη για crawling. Ωστόσο, το Διαδίκτυο είναι μια δυναμική οντότητα με υποχώρους που εξελίσσονται σε διαφορετικούς και συχνά ραγδαίους ρυθμούς. Επομένως, υπάρχει μια συνεχής ανάγκη για Crawlers προκειμένου να βοηθούν τις εφαρμογές να παραμένουν ενημερωμένες όσο νέες σελίδες προστίθενται και παλιές διαγράφονται, μετακινούνται ή τροποποιούνται.

Ο αριθμός των πιθανών ανιχνεύσιμων URLs που δημιουργούνται από το server-side λογισμικό δυσκολεύει το Web Crawler να αποφύγει να ανακτά αντίγραφα των σελίδων. Υπάρχουν ατέλειωτοι συνδυασμοί από HTTP-GET παραμέτρους, από τους οποίους μόνο ένα μικρό μέρος επιστρέφει μοναδικά δεδομένα. Για παράδειγμα μία μικρή έκθεση με φωτογραφίες στο Διαδίκτυο, προσφέρει τρεις επιλογές χρηστών, οι οποίες καθορίζονται μέσω HTTP-GET παραμέτρων στο URL. Αν υπάρχουν τέσσερεις τρόποι για την ταξινόμηση των φωτογραφιών, τρείς επιλογές για το μέγεθος της μικρογραφίας, δύο τύποι αρχείων και μία επιλογή για απενεργοποίηση δεδομένων που έστειλε ο χρήστης, τότε η ίδια πληροφορία παρέχεται από 48 διαφορετικά URLs, εκ των οποίων όλα μπορεί να είναι συνδεδεμένα με την ιστοσελίδα. Αυτός ο μαθηματικός συνδυασμός προκαλεί πρόβλημα στους Crawlers, αφού πρέπει να εντοπίσει και να ανακτήσει μοναδική πληροφορία.

Ο Crawler είναι ένα είδος πράκτορα λογισμικού, η λειτουργία του οποίου μπορεί να περιγραφεί περιληπτικά ως εξής:

- Ο Crawler έχει αρχικά μια λίστα URLs τα οποία επισκέπτεται.
- Από τα αρχικά URLs αποθηκεύει την πληροφορία που χρειάζεται.

- Εντοπίζει στα URLs τους εξωτερικούς συνδέσμους και τους αποθηκεύει σε μια λίστα που ονομάζεται frontier list.
- Επιλέγει από την frontier list τα URLs τα οποία θα επισκεφτεί στην συνέχεια.
- Συνεχίζει την ίδια λειτουργία μέχρι κάποιος τερματικός στόχος να επιτευχθεί.

3.1.3.1. Πολιτική επιλογής [9]

Αφού ένας Crawler μπορεί να αποθηκεύσει μόνο ένα μικρό μέρος από τις σελίδες του Διαδικτύου, είναι ιδιαίτερα επιθυμητό το αποθηκευμένο μέρος να περιέχει τις πιο σχετικές σελίδες και όχι απλά ένα τυχαίο δείγμα από το δίκτυο. Αυτό απαιτεί ένα μέτρο σημαντικότητας για να θέτει προτεραιότητες. Η σημαντικότητα μιας σελίδας είναι συνάρτηση της ποιότητάς της, της δημοτικότητάς της όσον αφορά συνδέσμους ή επισκέψεις, και άλλων παραγόντων.

Ένας Crawler μπορεί να αναζητά μόνο HTML σελίδες και να αποφεύγει όλους τους υπόλοιπους τύπους MIME. Προκειμένου να ζητήσει μόνο HTML σελίδες, ο Crawler μπορεί να στείλει μία HTTP HEAD αίτηση για να προσδιορίσει τους τύπους MIME μίας ιστοσελίδας προτού ζητήσει την σελίδα με μια GET αίτηση. Για να αποφύγει να στείλει πολυάριθμες HEAD αιτήσεις, ο Crawler μπορεί να εξετάσει το URL και να αποθηκεύει την ιστοσελίδα, μόνο αν έχει μια συγκεκριμένη κατάληξη όπως.html, .asp, .aspx, .php, .jsp, .jspx, ή κάθετος(/). Αυτή η μέθοδος μπορεί να προκαλέσει ακούσια παράλειψη πολλών HTML πόρων. Για παράδειγμα, ένας Crawler ο οποίος αναγνωρίζει μόνο τις παραπάνω καταλήξεις, θα προσπαθήσει να περιηγηθεί σε URLs όπως http://www.foo.com/something.html αλλά θα παραλήψει ένα URL με κατάληξη http://www.foo.com/bar.jpeg.

Οι Crawlers συνήθως εκτελούν κάποιο είδος τυποποίησης URL με σκοπό να αποφεύγουν την αποθήκευση των ίδιων πόρων παραπάνω από μία φορά. Ο όρος τυποποίησης URL, ή κανονικοποίησης URL αναφέρεται στη διαδικασία τροποποίησης και τυποποίησης ενός URL με κάποιο συγκεκριμένο τρόπο. Υπάρχουν διάφοροι τρόποι τυποποίησης που μπορούν να γίνουν, όπως η μετατροπή των χαρακτήρων σε πεζά, αφαίρεση των '.' και '..' τομέων, και προσθήκη καθέτων στις διαδρομές [12].

Κάποιοι Crawlers σκοπεύουν να αποθηκεύσουν όσο το δυνατόν περισσότερους πόρους από μία ιστοσελίδα. Έτσι δημιουργήθηκαν οι path ascending Crawlers, οι οποίοι μπορούν να ανέλθουν σε κάθε διαδρομή σε κάθε URL προς επίσκεψη. Για παράδειγμα, στο URL http://www.llama.org/hamster/monkey/parrot.html, ο Crawler θα προσπαθήσει να περιηγηθεί στα "hamster/monkey/", "/hamster/", και "/".

3.1.3.2. Πολιτική επανεξέτασης [9]

Το Διαδίκτυο έχει μία πολύ δυναμική φύση, και η περιήγηση σε ένα μέρος του, μπορεί να διαρκέσει εβδομάδες ή μήνες. Μέχρι να τελειώσει ο Crawler την περιήγησή του, μπορούν να προκύψουν πολλά γεγονότα, συμπεριλαμβανομένων δημιουργιών, ενημερώσεων και διαγραφών. Από την οπτική γωνία μίας μηχανής αναζήτησης, υπάρχει ένα κόστος που συνδέεται με τη μη ανίχνευση ενός γεγονότος, και επομένως να έχει μη ενημερωμένα αντίγραφα ενός πόρου. Οι πιο διαδεδομένες συναρτήσεις κόστους είναι η φρεσκότητα και η ηλικία [13].

Φρεσκότητα: Αποτελεί ένα δυαδικό μετρητή για την ακρίβεια ενός τοπικού αντίγραφου. Η φρεσκότητα μιας αποθηκευμένης σελίδας p που αποτελεί αντίγραφο σελίδας p' σε χρόνο t ορίζεται ως:

$$F_p(t) = \begin{cases} 1 & αν\ p = p'\ σε\ χρόνο\ t \\ 0 & σε\ αντίθετη\ περίπτωση \end{cases}$$

Ηλικία: Αποτελεί ένα μετρητή που καταδεικνύει πόσο ξεπερασμένο είναι το τοπικό αντίγραφο. Η ηλικία μίας αποθηκευμένης σελίδας p, σε χρόνο t ορίζεται ως:

$$A_p(t) = \begin{cases} 0 & \text{αν } p \text{ δεν τροποποιήθηκε στο χρόνο } t \\ t - \text{χρόνος τροποποίησης} & \text{σε διαφορετική περίπτωση} \end{cases}$$

3.1.3.3. Πολιτική ευγενείας [14]

Επειδή τα προγράμματα Web Crawling μπορούν να ανακτήσουν πολύ μεγάλο όγκο δεδομένων σε πολύ μικρό χρονικό διάστημα, μπορούν να αποτελέσουν μεγάλο φορτίο για έναν διακομιστή εις βάρος των κανονικών χρηστών του. Μεταξύ των προβλημάτων που μπορούν να δημιουργήσουν τα συστήματα Crawling στους διακομιστές ιστοσελίδων είναι:

- Αυξημένο κόστος πόρων δικτύου καθώς οι Crawlers απαιτούν αρκετό bandwidth
- Υπερφόρτωση διακομιστών
- Crawlers που διανέμονται ελεύθερα στο Διαδίκτυο για διάφορες χρήσεις μπορούν να δημιουργήσουν προβλήματα όταν πολλοί χρήστες στοχεύσουν στον ίδιο διακομιστή.

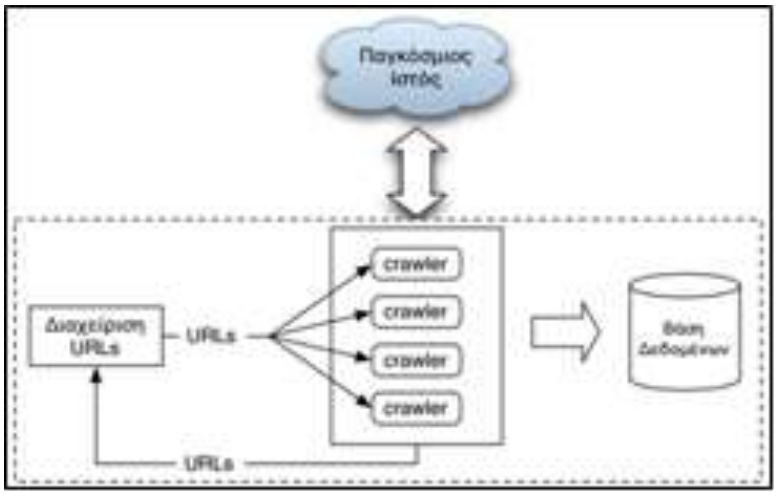
Μια μερική λύση σε αυτά τα προβλήματα είναι το πρωτόκολλο robots.txt (βλέπε Εικόνα 3) μέσω του οποίου οι διαχειριστές συστημάτων μπορούν να καθορίσουν ποια μέρη του διακομιστή δεν θα είναι προσβάσιμα από προγράμματα Crawling. Το πρωτόκολλο αυτό εφαρμόζεται με την ύπαρξη ενός αρχείου που ονομάζεται robots.txt και είναι αυτό το οποίο αναζητούν οι crawlers μόλις επισκέπτονται έναν ιστότοπο. Μέσα σε αυτό το αρχείο υπάρχουν οδηγίες για το ποιους καταλόγους απαγορεύεται να επισκεφθεί ένας Crawler. Οι οδηγίες μπορούν να αφορούν όλους τους Crawlers των μεγάλων εμπορικών μηχανών αναζήτησης, να τους κατονομάζουν και οι οδηγίες να αναφέρονται μόνο σε συγκεκριμένες μηχανές αναζήτησης.

```
User-agent: googlebot          # all services
Disallow: /private/            # disallow this directory

User-agent: googlebot-news     # only the news service
Disallow: /                    # on everything

User-agent: *                  # all robots
Disallow: /something/          # on this directory
```

Εικόνα 3: Παράδειγμα robot.txt για πολλαπλούς user agents



Εικόνα 4: Γενική αρχιτεκτονική ενός συστήματος παράλληλου Crawling

3.1.3.4. Πολιτική παραλληλοποίησης [14]

Ένα παράλληλο σύστημα Crawling (βλέπε Εικόνα 4) μπορεί να εκτελεί πολλές διεργασίες παράλληλα. Οι στόχοι ενός τέτοιου συστήματος είναι να μεγιστοποιήσει το ρυθμό ανακάλυψης και ανάγνωσης ιστοσελίδων ενώ ταυτόχρονα πρέπει να ελαχιστοποιήσει την επιβάρυνση από την παραλληλοποίηση και να αποφεύγει το επαναλαμβανόμενο κατέβασμα των ίδιων σελίδων. Οι στόχοι αυτοί επιτυγχάνονται με το διαχωρισμό του συστήματος σε

τμήματα καθένα από τα οποία είναι υπεύθυνο για μία συγκεκριμένη εργασία. Τα σχέδια και οι αρχιτεκτονικές των Crawlers θεωρούνται εταιρικά μυστικά και δε δημοσιοποιούνται από τις μηχανές αναζήτησης. Η πιο συνηθισμένη προσέγγιση όμως είναι να υπάρχει ένα σύστημα διαχείρισης των διευθύνσεων που πρέπει να επισκεφθεί ο Crawler και πολλές παράλληλες διεργασίες Crawling που παίρνουν τις διευθύνσεις από αυτό το σύστημα ως παραμέτρους.

3.1.4. Προβλήματα κατά το crawling [15]

Τα προβλήματα που δημιουργούνται κατά το crawling του παγκόσμιου ιστού είναι τα εξής:

ΑΠΟΔΟΣΗ

Ένας Crawler περιορίζεται από το δικτυακό εύρος ζώνης και το χώρο αποθήκευσης που έχει στη διάθεσή του τα οποία δεν είναι άπειρα αλλά ούτε και δωρεάν. Η διαδικασία του crawling είναι αρκετά χρονοβόρα και η συλλογή μερικών εκατομμυρίων σελίδων μπορεί να διαρκέσει μέρες ή βδομάδες. Γι' αυτόν τον λόγο, ο Crawler πρέπει να χρησιμοποιεί αποδοτικά τους υπολογιστικούς πόρους που διαθέτει έτσι ώστε να μεγιστοποιεί και να διατηρεί σταθερή απόδοση η οποία εκφράζεται με το ρυθμό ανάκτησης σελίδων. Η ταυτόχρονη ανάκτηση σελίδων από πολλαπλούς διαφορετικούς διακομιστές, εκτός από την αποφυγή υπερφόρτωσής τους, μεγιστοποιεί την απόδοση του Crawler και συνιστά μια αναγκαία τακτική.

ΔΙΑΘΕΣΙΜΟΤΗΤΑ ΤΩΝ ΔΙΑΚΟΜΙΣΤΩΝ

Η ποιότητα υπηρεσιών στο Διαδίκτυο πρέπει να θεωρείται δεδομένη. Σε πολλές περιπτώσεις συμβαίνει ο διακομιστής να μην είναι διαθέσιμος για ένα χρονικό διάστημα και να παράγει σφάλμα time-out κατά την προσπάθεια σύνδεσης. Συχνά ο διακομιστής επανέρχεται σε λειτουργία μετά από κάποιες ώρες ή μέρες. Ο Crawler θα πρέπει να ξαναεπισκέπτεται τις σελίδες ενός τέτοιου διακομιστή ανά διαστήματα (της τάξης των ωρών) ώστε να μπορέσει να τις ανακτήσει όταν και αν ξαναγίνει διαθέσιμος.

ΥΠΕΡΦΟΡΤΩΣΗ ΤΩΝ ΔΙΑΚΟΜΙΣΤΩΝ

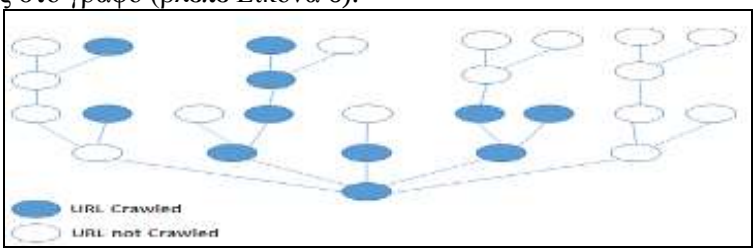
Η ύπαρξη πολλών Crawlers που σαρώνουν διαρκώς τον ιστό, οδηγεί σε κατανάλωση εύρους ζώνης που κανονικά προορίζεται για τους χρήστες. Αυτό μπορεί να έχει ως αποτέλεσμα την αντίδραση των διαχειριστών των διακομιστών που στη χειρότερη περίπτωση μπορούν να μπλοκάρουν την πρόσβαση στο περιεχόμενο του διακομιστή σε τέτοιου είδους προγράμματα. Για το λόγο αυτό προτείνονται να ακολουθούνται οι εξής δύο βασικές αρχές στην υλοποίηση ενός Crawler:

- Ο Crawler θα πρέπει να αποκαλύπτει την ταυτότητα του (με χρήση του HTTP πεδίου User-Agent) και να περιέχει μια διεύθυνση e-mail για επικοινωνία με τον υπεύθυνο χρήσης του Crawler.
- Ο Crawler θα πρέπει να περιμένει για ένα χρονικό διάστημα (τουλάχιστον 30 δευτερολέπτων) ανάμεσα στις διαδοχικές αιτήσεις στον ίδιο διακομιστή. Επειδή πολλά host names οδηγούν στην ίδια διεύθυνση IP, ο διακομιστής θα πρέπει να ταυτοποιείται από τη διεύθυνση IP και όχι μόνο από το host name.

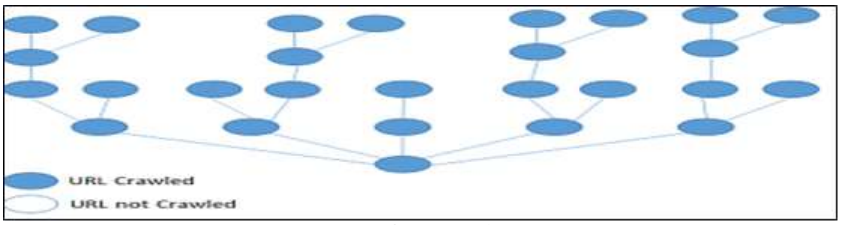
3.1.5. Εστιασμένος Web Crawler

Ένας εστιασμένος ή τοπικός web Crawler έχει σαν στόχο τη λήψη σελίδων οι οποίες σχετίζονται με ένα προκαθορισμένο θέμα ή σύνολο θεμάτων. Το τοπικό crawling γενικά υποθέτει πως δίνεται μόνο το θέμα, ενώ το εστιασμένο crawling υποθέτει πως δίνονται και κάποια παραδείγματα σχετικών και μη σελίδων [16] [17]. Ιδανικά ένας εστιασμένος Crawler αποθηκεύει μόνο σελίδες σχετικές με το δοθέν θέμα και αποφεύγει να αποθηκεύσει τις

υπόλοιπες. Ως εκ τούτου ένας εστιασμένος Crawler μπορεί να προβλέψει την πιθανότητα πως ένας σύνδεσμος προς μία συγκεκριμένη σελίδα να είναι σχετική, πριν την πραγματική λήψη της σελίδας. Επομένως, ένας εστιασμένος Crawler, ελέγχει και καταγράφει ένα περιορισμένο αριθμό σελίδων (βλέπε Εικόνα 5) σε αντίθεση με το συμβατικό Crawler, ο οποίος ελέγχει όλες τις σελίδες στο γράφο (βλέπε Εικόνα 6).



Εικόνα 5: Παράδειγμα Εστιασμένου Crawler



Εικόνα 6: Παράδειγμα απλού Crawler

Σε μια διαφορετική προσέγγιση, η σχετικότητα της σελίδας καθορίζεται μετά τη λήψη των περιεχομένων της. Οι σχετικές σελίδες στέλνονται προς ευρετηριοποίηση και τα URLs που περιέχονται στη σελίδα, τοποθετούνται στη λίστα με τα URLs προς επίσκεψη, ενώ σελίδες που δεν ξεπερνούν το όριο συνάφειας, απορρίπτονται. Η επίδοση ενός εστιασμένου Crawler εξαρτάται κυρίως από τον πλούτο των συνδέσμων για ένα συγκεκριμένο θέμα προς αναζήτηση, και ένας εστιασμένος Crawler συνήθως βασίζεται σε μία γενική μηχανή αναζήτησης για την παροχή του σημείου εκκίνησης. Στον παρακάτω πίνακα (Εικόνα 7) παρουσιάζονται τα κυριότερα χαρακτηριστικά των διάφορων τεχνικών ενός εστιασμένου Web Crawler.

Κύρια Χαρακτηριστικά Διαφόρων τεχνικών Focused Crawling		
Σύστημα		**Κύρια Χαρακτηριστικά**
Λεξιλογική Προσέγγιση	Fish Search System	Βασίζεται στην depth-first προσέγγιση. Ο crawler συνεχίζει την διάσχιση μόνο όταν βρει σχετικό link.
	Shark Search System	Επεκτείνει το Fish Search system δίνοντας μια ασαφή τιμή σχετικότητας σε μια σελίδα σε αντίθεση με την δυαδική τιμή του fish search.
	Focused Crawler βασισμένος σε ταξινόμηση	Βασίζεται σε ένα δέντρο ταξινόμησης το οποίο δημιουργείται από τα αρχικά links και στο οποίο κατατάσσονται τα εξερευνημένα links.
	Focused Crawler βασισμένος σε μηχανή υπολογισμού ομοιότητας	Δημιουργεί έναν πίνακα αναλύοντας λέξεις κλειδιά (keywords) από το θέμα που επιλέχθηκε από το χρήστη και κατατάσσει σε αυτόν τα links που αναλύονται. Χρησιμοποιεί TF.IDF σχήμα.
Προσέγγιση Συνδέσμων	Crawler βασισμένος στην ομοιότητα	Υπολογίζει το Page Rank που δημιουργείται από τις σελίδες που έχουν εξερευνηθεί ήδη και το χρησιμοποιεί σαν προτεραιότητα για τις μελλοντικές εξερευνήσεις.
	HITS	Χρησιμοποιεί το hub page score και το page authority score για να υπολογίσει την προτεραιότητα του link.
	ARC	Δημιουργεί μια λίστα πόρων για ένα θέμα που είναι ευρέως διαδεδομένο στο διαδίκτυο.
	Focused Crawler βασισμένος στο DOM	Χρησιμοποιεί το DOM δέντρο του link για να εντοπίσει περιοχές hub της σελίδας σχετικές με το θέμα και να δώσει μεγαλύτερη προτεραιότητα σ' αυτές.
	Context graph Focused Crawling	Δημιουργεί ένα μοντέλο από το περιεχόμενο των αρχικών links. Στη συνέχεια, σχηματίζονται ιεραρχίες link από τις οποίες προκύπτουν τα link με μεγαλύτερη αξία.

Εικόνα 7: Κύρια Χαρακτηριστικά Διαφόρων Τεχνικών του Εστιασμένου Web Crawling

3.1.6. Αλγόριθμοι

Στην ενότητα αυτή παρουσιάζονται συνοπτικά βασικά στοιχεία των επικρατέστερων αλγόριθμων που ενσωματώνονται σε δημοφιλής μηχανές αναζήτησης δεδομένων στο Διαδίκτυο.

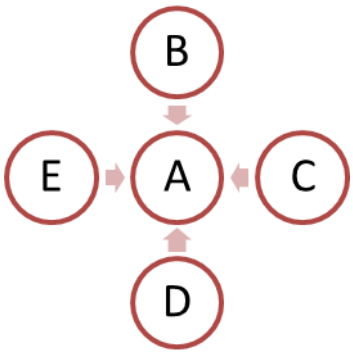
3.1.6.1 Αλγόριθμος PageRank [18] [19]

Ο αλγόριθμος PageRank είναι ένας αλγόριθμος ανάλυσης συνδέσμων. Ο αλγόριθμος πήρε το όνομα του από τον *Lawrence Page*, έναν εκ των δημιουργών του και χρησιμοποιήθηκε από τη μηχανή αναζήτησης Google. Ο αλγόριθμος θέτει ένα αριθμητικό βάρος σε κάθε στοιχείο ενός συνόλου από υπερσυνδεδεμένα έγγραφα, όπως ο Παγκόσμιος Ιστός, με σκοπό τον υπολογισμό της σχετικής "σημαντικότητας" μέσα στο σύνολο.

Από την πρώτη δημοσίευση του αλγόριθμου από τους *Page* και *Brin*, έχουν εκδοθεί αρκετές ερευνητικές εργασίες όσων αφορά το PageRank. Στην πραγματικότητα, ο αλγόριθμος είναι τρωτός σε χειραγώγηση.

ΕΠΕΞΗΓΗΣΗ ΑΛΓΟΡΙΘΜΟΥ

Ο αλγόριθμος PageRank αποτελεί μια κατανομή πιθανοτήτων που χρησιμοποιείται για να αναπαραστήσει την πιθανότητα ένα άτομο, το οποίο επιλέγει τυχαίους συνδέσμους[4], να καταλήξει σε κάποια συγκεκριμένη σελίδα. Η πιθανότητα αυτή εκφράζεται από μια αριθμητική τιμή μεταξύ των 0 και 1. Παραδείγματος χάριν η τιμή 0.5 σημαίνει ότι υπάρχει 50% πιθανότητα ένα άτομο που χρησιμοποιεί ένα τυχαίο υπερσύνδεσμο, να καταλήξει σε μια σελίδα με PageRank τιμή 0.5.

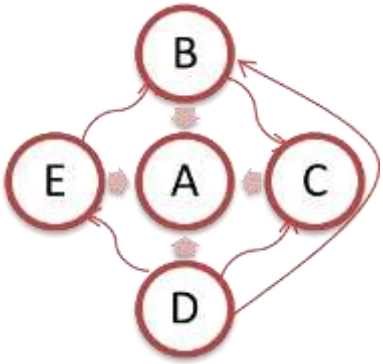


Εικόνα 8: Απλό Σύστημα 5 Ιστοσελίδων

Έστω ένα μικροσύστημα με πέντε ιστοσελίδες, όπως φαίνεται στην Εικόνα 8: τις **A**, **B**, **C**, **D** και **E.** Σύνδεσμοι από μια σελίδα προς την ίδια την σελίδα, ή πολλαπλοί σύνδεσμοι από μια μοναδική σελίδα σε μια άλλη μοναδική σελίδα, αγνοούνται. Η τιμή PageRank αρχικοποιείται στην ίδια τιμή για όλες τις σελίδες. Στην αρχική μορφή του αλγόριθμου, το άθροισμα της τιμής PageRank όλων των σελίδων ήταν ίση με τον συνολικό αριθμό των σελίδων εκείνη τη στιγμή, επομένως η αρχική τιμή των σελίδων στο παράδειγμα θα ήταν ίση με 1. Παρ'όλα αυτά μεταγενέστερες εκδόσεις του αλγορίθμου, προϋποθέτουν μια κατανομή πιθανοτήτων ανάμεσα στις τιμές 0 και 1. Επομένως η αρχική τιμη PageRank της κάθε σελίδας θα είναι 0.2. Εφ'όσων οι μόνοι σύνδεσμοι στο σύνολο προέρχονται από τις σελίδες **B**, **C**, **D** και **E** προς την **A**, κάθε σύνδεσμος θα μεταφέρει μια τιμή PageRank ίση με 0.20, με συνολική τιμή 0.80.

$$PR(A) = PR(B) + PR(C) + PR(D) + PR(E)$$

[4] Random Surfer Model.



Εικόνα 9: Απλό Σύστημα 5 Ιστοσελίδων(2)

Ας υποθέσουμε τώρα ότι η σελίδα **B** έχει συνδέσμους προς τις σελίδες **C** και **A,** ενώ η σελίδα **D** έχει συνδέσμους προς όλες τις σελίδες (βλέπε Εικόνα 9). Έτσι η σελίδα **B** θα μεταφέρει την μισή αρχική της τιμή ή 0.1, στην σελίδα **A** και την υπόλοιπη μισή ή 0.1, στην σελίδα **C.** Εφ'όσων η σελίδα **D** έχει τέσσερις εξωτερικούς συνδέσμους, θα μεταφέρει ένα τέταρτο της αρχικής της τιμής, ή 0.05 στην **A.**

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{4} + \frac{PR(E)}{1}$$

Δηλαδή, η τιμή PageRank που παρέχεται σε έναν εξωτερικό σύνδεσμο ισούται με την τιμή PageRank του εγγράφου δια τον αριθμό των εξωτερικών συνδέσμων L.

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \frac{PR(E)}{L(E)}$$

Η γενική μορφή υπολογισμού της PageRank τιμής οποιασδήποτε σελίδας δίνεται από τον τύπο:

$$PR(u) = \sum_{u \in B} \frac{PR(u)}{L(u)}$$

$$\frac{PR(B)}{2}\frac{PR(C)}{1}\frac{PR(D)}{4}\ \frac{PR(E)}{1}\frac{PR(B)}{L(B)}\frac{PR(C)}{L(C)}\frac{PR(D)}{L(D)}\frac{PR(E)}{L(E)}\sum_{u \in B}\frac{PR(u)}{L(u)}$$

ΣΥΝΤΕΛΕΣΤΗΣ ΑΠΟΣΒΕΣΗΣ

Ο αλγόριθμος PageRank θεωρεί ότι ακόμη και ένας φανταστικός χρήστης ο οποίος επιλέγει τυχαία υπερσυνδέσμους προς άλλες ιστοσελίδες τελικά θα σταματήσει. Η πιθανότητα, σε οποιοδήποτε βήμα, ο χρήστης να συνεχίσει, είναι ο συντελεστής απόσβεσης *d.* Διάφορες ερευνητικές εργασίες έχουν πειραματιστεί με διαφορετικούς συντελεστές απόσβεσης, αλλά η γενική αποδεκτή τιμή του συντελεστή απόσβεσης είναι η τιμή 0.85.

Ο συντελεστής απόσβεσης αφαιρείται από τη μονάδα και διαιρείται από τον αριθμό των εγγράφων μέσα στο σύνολο *N,* και ο όρος αυτός προστίθεται στο γινόμενο του συντελεστή απόσβεσης και του αθροίσματος των εισερχόμενων PageRank τιμών:

$$PR(u) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \frac{PR(E)}{L(E)}\right)$$

Επομένως η τιμή PageRank οποιασδήποτε σελίδας έχει άμεση σχέση με τις τιμές PageRank των υπόλοιπων σελίδων. Ο συντελεστής απόσβεσης προσαρμόζει την παραγώμενη τιμή. Στην αρχική εργασία, όμως, παρουσιαζόταν η ακόλουθη εξίσωση:

$$PR(u) = 1\text{-}d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \frac{PR(E)}{L(E)}\right)$$

Η διαφορά ανάμεσα στις δύο εξισώσεις είναι πως οι PageRank τιμές που παράγονται από την πρώτη εξίσωση έχουν άθροισμα 1, ενώ οι PageRank τιμές που παράγονται από την δεύτερη εξίσωση ξεπερνούν σε άθροισμα το 1. Η αναφορά στην εργασία των Page και Brinn πως «το άθροισμα όλων των τιμών PageRabk είναι 1[5]» υποστηρίζουν την πρώτη εκδοχή της παραπάνω εξίσωσης.

Όταν υπολογίζεται η τιμή PageRank σε ένα σύνολο, οι σελίδες χωρίς εξωτερικούς συνδέσμους, θεωρείται πως έχουν εξωτερικούς συνδέσμους προς όλες τις υπόλοιπες σελίδες μέσα στο σύνολο. Η PageRank τιμή τους επομένως μοιράζεται ισομερώς σε όλες τις υπόλοιπες σελίδες.

Έτσι, η εξίσωση φαίνεται παρακάτω:

$$PR(p_i) = \frac{1-d}{N} + d\sum_{Pj \in M(Pi)}^{j=1,2,\ldots,N} \frac{PR(Pj)}{L(Pj)}$$

$$\frac{1-d}{N}\frac{PR(B)}{L(B)}\frac{PR(C)}{L(C)}\frac{PR(D)}{L(D)}\frac{PR(E)}{L(E)}\frac{PR(B)}{L(B)}\frac{PR(C)}{L(C)}\frac{PR(D)}{L(D)}\frac{PR(E)}{L(E)}\frac{1-d}{N}$$

όπου $p_1, p_2, \ldots, p_N$ είναι οι σελίδες υπό εξέταση, $M(p_i)$ το σύνολο των σελίδων με εξωτερικούς συνδέσμους προς την p_i, $L(p_j)$ ο αριθμός των εξωτερικών συνδέσμων στην σελίδα p_j, και N ο συνολικός αριθμός των σελίδων.

Η PageRank τιμή μπορεί να υπολογιστεί από έναν απλό αναδρομικό αλγόριθμο και αποτελεί το κυρίαρχο ιδιοδιάνυσμα ενός ειδικά κανονικοποιημένου πίνακα συνδέσμων του Διαδικτύου. [20]

Οι PageRank τιμές είναι είσοδοι του κυρίαρχου ιδιοδιανύσματος του τροποποιημένου πίνακα γειτνίασης. Το ιδιοδιάνυσμα είναι:

$$R = \begin{matrix} PR(p1) \\ \vdots \\ PR(pN) \end{matrix}$$

όπου **R** η λύση της εξίσωσης :

$$R = \begin{matrix} \frac{1-d}{N} \\ \vdots \\ \frac{1-d}{N} \end{matrix} + d\begin{bmatrix} \ell(p1,p1) & \cdots & \ell(p1,pN) \\ \vdots & \ddots & \vdots \\ \ell(pN,p1) & \cdots & \ell(pN,pN) \end{bmatrix}$$

όπου η εξίσωση γειτνίασης $\ell(Pi,Pj)$ είναι 0 αν η σελίδα p_j δεν έχει εξωτερικό σύνδεσμο προς το p_i, και κανονικοποιείται έτσι, για κάθε j

$$\sum_{i=1}^{N} \ell(Pi,Pj) = 1$$

Τα στοιχεία κάθε στήλης έχουν άθροισμα μέχρι 1, έτσι ο πίνακας είναι στοχαστικός. Άρα, αυτή είναι μια παραλλαγή του μέτρου του ιδιοδιανύσματος κεντρικότητας που χρησιμοποιείται στην ανάλυση δικτύων.

[5] the sum of all PageRanks is one

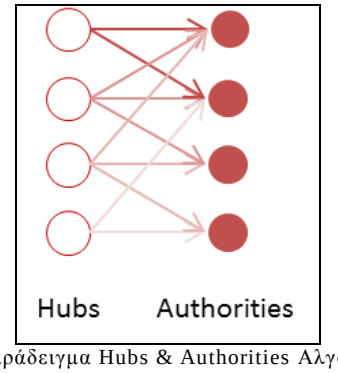
$$PR(p1)\frac{1-d}{N}\begin{bmatrix}\ell(p1,p1) & \cdots & \ell(p1,pN)\\ \vdots & \ddots & \vdots\\ \ell(pN,p1) & \cdots & \ell(pN,pN)\end{bmatrix}\ell(Pi,Pj)\sum_{i=1}^{N}\ell(Pi,Pj)$$
$$PR(pN)\frac{1-d}{N}$$

3.1.6.2 Αλγόριθμος HITS [21]

Ένας εναλλακτικός αλλά εξίσου σημαντικός αλγόριθμος είναι ο αλγόριθμος HITS (Hyperlink-Induced Topic Search) που δημιουργήθηκε από τον Jon Kleinberg και χρησιμοποιήθηκε πρώτη φορά από τη μηχανή αναζήτησης CLEVER της IBM [20].

Ο αλγόριθμος βασίζεται στην θεωρία κατά την οποία στο Διαδίκτυο υπάρχουν σελίδες με ενδιαφέρον ή αξιόλογο υλικό για το χρήστη (*authorities*) και σελίδες οι οποίες χρησιμοποιούνται σαν κατάλογοι, που δεν περιέχουν ιδιαίτερα σημαντική πληροφορία αλλά περιέχουν ένα σύνολο από συνδέσμους προς σημαντικές σελίδες (*hubs*) [22].

Το σύστημα, επομένως, αποδίδει δύο τιμές σε κάθε ιστοσελίδα: την τιμή της **αρχής**, η οποία εκτιμά την ποιότητα του περιεχομένου της σελίδας, και την τιμή **διανομέα**, η οποία εκτιμά την ποιότητα των συνδέσμων της προς άλλες σελίδες [23].



Εικόνα 10: Παράδειγμα Hubs & Authorities Αλγορίθμου HITS

Ο αλγόριθμος HITS παράγει αποτελέσματα ως εξής: Αρχικά δημιουργεί έναν εστιασμένο υπογράφο ο οποίος περιέχει τις περισσότερες (ή τουλάχιστον πολλές) από τις authority σελίδες (βλέπε Εικόνα 10). Έπειτα ο αλγόριθμος εκτελεί μία σειρά από επαναλήψεις στις οποίες ενημερώνεται η τιμή αρχής και διανομέα της κάθε σελίδας. Τέλος κανονικοποιούνται οι τιμές αρχής και διανομέα για να παραχθούν τα αποτελέσματα.

ΔΗΜΙΟΥΡΓΙΑ ΕΝΟΣ ΕΣΤΙΑΣΜΕΝΟΥ ΥΠΟΓΡΑΦΟΥ

Ο εστιασμένος υπογράφος χρησιμοποιείται για να συλλέξουμε τα πιο σχετικά αποτελέσματα. Έστω μια συλλογή V από συνδεδεμένες σελίδες. Μπορούμε να αναπαραστήσουμε τη συλλογή αυτή σαν ένα κατευθυνόμενο γράφο G = (V,E): οι κόμβοι αντιπροσωπεύουν τις σελίδες και μία κατευθυνόμενη ακμή (p,q) $\in E$ αντιπροσωπεύει ένα σύμβολο ανάμεσα στο p και το q. Από τον γράφο G, μπορούμε να απομονώσουμε υπογράφους με τον εξής τρόπο. Αν $W \subseteq V$ είναι ένα υποσύνολό από σελίδες, λέμε ότι G[W] είναι ένας γράφος που επάγεται στον W. Για μία παράμετρο t (συνήθως με τιμή 200), συλλέγονται οι *t* υψηλότερα σε κατάταξη σελίδες από ένα ερώτημα *σ* από κάποια text-based μηχανή αναζήτησης. Αυτές οι *t* σελίδες αναφέρονται σαν *root σύνολο* R_σ. Χρησιμοποιώντας αυτό το σύνολο για τη δημιουργία ενός συνόλου σελίδων S_σ, που ικανοποιούν τις συνθήκες του αλγόριθμου. Έστω μια σελίδα με υψηλή authority τιμή, η οποία δεν ανήκει στο σύνολο R_σ, είναι πιθανό να υπάρχει σύνδεσμος ανάμεσα σε αυτή και κάποια σελίδα στο σύνολο R_σ. Μπορούμε να αυξήσουμε τον αριθμό των σελίδων με μεγάλη τιμή αρχής στον υπογράφο μας

προσθέτοντας τη σελίδα μαζί με τους συνδέσμους της. Στην Εικόνα 11 παρατίθεται ο αλγόριθμος της δημιουργίας ενός εστιασμένου υπογράφου για τον αλγόριθμο HITS.

```
Subgraph(σ,ℰ,t,d)
  σ: a query string.
  ℰ: a text-based search engine.
  t, d: natural numbers.
  Let R_σ denote the top t results of ℰ on σ.
Set S_σ := R_σ
For each page p ∈ R_σ
  Let Γ+(p) denote the set of all pages p points to.
  Let Γ−(p) denote the set of all pages pointing to p.
  Add all pages in Γ+(p) to S_σ.
  If |Γ−(p)| ≤ d, then
    Add all pages in Γ−(p) to S_σ.
  Else
    Add an arbitrary set of d pages from Γ−(p) to S_σ.
End
Return S_σ
```

Εικόνα 11: Ψευδοκώδικας Δημιουργίας Εστιασμένου Υπογράφου

Υπολογισμός των τιμών Αρχής των Διανομέων κάθε Σελίδας

Οι τιμές αρχής και διανομέα κάθε εγγράφου, υπολογίζονται χρησιμοποιώντας τη μία την άλλη σε μία κοινή αναδρομή. Η τιμή αρχής ισούται με το άθροισμα των τιμών διανομέα των σελίδων που οδηγούν σε αυτήν. Η τιμή διανομέα ισούται με το άθροισμα των τιμών αρχής των σελίδων που οδηγεί.
Οι τιμές αρχής (Authority Score) και διανομέα (Hub Score) υπολογίζονται από τις παρακάτω εξισώσεις:

$$AuthorityScore(p) = \sum_{\forall q \, linking \, to \, p}^{n} HubScore(q)$$

$$HubScore(p) = \sum_{\forall r \, linking \, from \, p}^{m} AuthorityScore(r)$$

Σε αντίθεση με τον αλγόριθμο PageRank, ο αλγόριθμος HITS προσφέρει μία βαθμολόγηση η οποία εξαρτάται από το ερώτημα, το οποίο σημαίνει ότι υπολογίζει την ποιότητα και την σχετικότητα μίας σελίδας σε σχέση με το δοθέν ερώτημα. Έτσι δεν απαιτείται καθολική βαθμολόγηση των σελίδων, κάτι που κάνει τον αλγόριθμο πιο κατάλληλο για εργασίες όπως το web crawling.

Αδυναμίες Αλγόριθμου HITS και αλγόριθμοι επίλυσης

Ο αλγόριθμος HITS παρουσιάζει κάποια προβλήματα τα οποία αναφέρονται παρακάτω:

i. Δίνει μεγάλη αξία σε δημοφιλείς ιστοσελίδες, οι οποίες όμως, δεν είναι σχετικές με το δοθέν ερώτημα.
ii. Παρουσιάζει απόκλιση από το θέμα στην περίπτωση που μία σελίδα με υψηλή τιμή διανομέα, αναφέρεται σε διαφορετικά θέματα αφού δίνει το ίδιο βάρος σε όλους τους εξωτερικούς συνδέσμους της σελίδας.

Για να ελαχιστοποιήσει τα προβλήματα του αλγόριθμου HITS προτάθηκε ένας αλγόριθμος από τους S.Chakrabarti et al. [24]. Αποτελεί τροποποίηση του αλγορίθμου HITS.

Κατά τον αλγόριθμο αυτόν, αποδίδεται ένα αριθμητικό βάρος σε κάθε σύνδεσμο, το οποίο εξαρτάται από τις παραμέτρους του ερωτήματος. Επίσης μία σελίδα με μεγάλη τιμή διανομέα και πολλά διαφορετικά θέματα χωρίζεται σε τμήματα τέτοια ώστε το καθένα να αναφέρεται σε ένα και μόνο θέμα.

Ένας ακόμη περιορισμός του αλγορίθμου HITS είναι ότι υποθέτει ίσα βάρη σε όλους τους συνδέσμους που κατευθύνονται σε μία ιστοσελίδα και έτσι αποτυγχάνει να αναγνωρίσει το γεγονός ότι κάποιοι σύνδεσμοι μπορεί να είναι πιο σημαντικοί από άλλους. Για να επιλύσει αυτό το πρόβλημα προτάθηκε από τους D.Cohn et al.ένα πιθανοτικό ανάλογο του HITS (PHITS-Probabilistic HITS) [25]. Ο PHITS παρέχει μία πιθανολογική εξήγηση της σχέσης ανάμεσα στις παραμέτρους του ερωτήματος και το έγγραφο, και σύμφωνα με το συγγραφέα μπορεί να παρέχει έγκυρα αποτελέσματα.

3.1.6.3 Google Panda [26]

Το Google Panda είναι ένας διαφορετικός αλγόριθμος αναζήτησης και κατάταξης αποτελεσμάτων της Google που κυκλοφόρησε για πρώτη φορά το Φεβρουάριο του 2011. Η αλλαγή είχε ως στόχο να μειώσει την κατάταξη των χαμηλών ποιοτικά ιστοσελίδων και να επιστρέψει υψηλής ποιότητας θέσεις κοντά στην κορυφή των αποτελεσμάτων αναζήτησης. Η CNET αναφέρει μία αύξηση στην κατάταξη των ειδησεογραφικών σελίδων και σελίδων κοινωνικής δικτύωσης, καθώς και μία πτώση στις ιστοσελίδες οι οποίες περιέχουν μεγάλες ποσότητες διαφήμισης. Αυτή η αλλαγή φέρεται να επηρέασε την κατάταξη σχεδόν το 12 τοις εκατό όλων των αποτελεσμάτων αναζήτησης. Λίγο μετά την εγκατάσταση του Panda, πολλές σελίδες συμπεριλαμβανομένων φόρουμ webmaster της Google γέμισε με καταγγελίες παράβασης πνευματικών δικαιωμάτων (copyright) για να πάρουν καλύτερη κατάταξη από sites τα οποία είχαν πρωτότυπο περιεχόμενο και άρα καλύτερη κατάταξη. Κάποια στιγμή η Google ζήτησε σημεία δεδομένων (data points) για να εντοπίσει τις υποκλοπές πιο εύκολα.

Το Google Panda έχει ενημερωθεί μετά την εγκατάστασή του τον Φεβρουάριο του 2011 και το αποτέλεσμα δόθηκε τον Απρίλη του 2011. Για να βοηθήσει τους εκδότες η Google έκανε μία ενημερωτική δημοσίευση στο blog της δίνοντας έτσι κάποια κατεύθυνση για την αυτό-αξιολόγηση της ποιότητας ενός Διαδικτυακού τόπου.

Τέλος, η Google δημοσίευσε μία λίστα με 13 bullets στο blog της που απαντούσε στην ερώτηση «ποια σελίδα μπορεί να μετρηθεί ως υψηλής ποιότητας σελίδα» που υποτίθεται ότι θα βοηθήσει τους webmasters να οδηγηθούν ένα βήμα προς την νοοτροπία της Google.

Τα Χαρακτηριστικά του Google Panda

Ένα έγγραφο ταξινόμησης (document classifier) είναι ένα υπολογιστικό πρόγραμμα το οποίο εκτελεί μια συγκεκριμένη λειτουργία ή ένα σύνολο λειτουργιών για την αξιολόγηση των εγγράφων σε ένα ευρετήριο αναζητήσεων. Μερικά παραδείγματα των δυνατοτήτων του είναι: Οι ταξινομητές μπορούν να προσδιορίζουν, να επισημαίνουν και να ταξινομούν έγγραφα. Επισημαίνουν έγγραφα για περαιτέρω επεξεργασία, ταξινομούν τα έγγραφα σε ομάδες, σημειώνουν έγγραφα έναντι κάποιων μοντέλων μέτρησης. Μπορούν να μειώσουν ή να τροποποιήσουν τα δεδομένα του εγγράφου, να σαρώσουν έγγραφα για συγκεκριμένη αναζήτηση.

Ένας απλός ορισμός για τον όρο «ταξινομητής εγγράφου»: Είναι ένα ειδικό πρόγραμμα το οποίο λύνει ένα συγκεκριμένο πρόβλημα για ένα χώρο με έγγραφα ή ένα σύνολο κατηγοριοποιημένων εγγράφων. Παραδείγματος χάριν ένα πρόβλημα θα μπορούσε να είναι: «Πώς μπορώ να διαιρέσω το σύνολο των εγγράφων σε «Κατηγορία 1»και «Κατηγορία 2».

Το Google Panda εισήχθη για να απομακρύνει τις εκμεταλλεύσεις περιεχομένου και τις περιοχές χαμηλής ποιότητας που προσπαθούν να ζουν από την παραγωγή χαμηλής ποιότητας περιεχομένου. Νέες ενημερώσεις του Google Panda το βοηθούν να αναλύει δικτυακούς τόπους οι οποίοι βασίζονται στον αλγόριθμο αυτό-διδασκαλίας Al, ο οποίος

αναζητά ομοιότητες μεταξύ των ανθρώπων που βρέθηκαν να είναι χαμηλής και υψηλής ποιότητας βασιζόμενος στον τρόπο πλοήγησης. Υπάρχει επίσης μια μεγάλη πιθανότητα η αυξημένη χρήση της ανάλυσης εικόνας με ανάλυση κειμένου να προσδιορίζει το αρχικό περιεχόμενο.

3.1.6.4. Google Penguin [27]

Ο αλγόριθμος Google Penguin είναι ένας νέος αλγόριθμος κατάταξης, που ανακοινώθηκε από την εταιρία στις 24 Απριλίου, 2012 [28]. Σκοπός του αλγορίθμου είναι η μείωση της κατάταξης των ιστοσελίδων που παραβιάζουν τους κανονισμούς της Google για υπεύθυνους ιστοσελίδων6 [29], χρησιμοποιώντας κακόβουλες SEO τεχνικές όπως η υπερβολική χρήση λέξεων-κλειδιών, το cloacking, η συμμετοχή σε "παράνομους" σχηματισμούς υπερσυνδέσμων7, σκόπιμη αντιγραφή περιεχομένου, κ.α

Σύμφωνα με την εταιρία, ο αλγόριθμος επηρεάζει το 3.1% των αναζητήσεων στην Αγγλική γλώσσα, 3% σε γλώσσες όπως τα Γερμανικά, τα Αραβικά και τα Κινέζικα, ενώ σε άλλες γλώσσες τα ποσοστά είναι ακόμη μεγαλύτερα. [30].

3.1.6.5. Άλλοι Σημαντικοί Αλγόριθμοι Κατάταξης Ιστοσελίδων [31]

Πέρα των αλγορίθμων που παρουσιάσαμε και οι οποίοι έχουν ενσωματωθεί σε ευρέως διαθέσιμες και δημοφιλείς μηχανές αναζήτησης, έχουν υπάρξει κατά καιρούς, πολλές άλλες προσπάθειες είτε επέκτασης υφιστάμενων αλγορίθμων είτε ανάπτυξης νέων που να αντιμετωπίζουν εξειδικευμένα θέματα. Οι σημαντικότερες αυτών των προσπαθειών συνοψίζονται παρακάτω.

WEIGHTED PAGE RANK ALGORITHM [32]

Ο αλγόριθμος Weighted Page Rank (WPR) είναι μια τροποποίηση του αλγόριθμου PageRank. Ο WPR θεωρεί πως κάποιες σελίδες είναι πιο δημοφιλείς από άλλες, και κάθε σύνδεσμος παίρνει τιμή ανάλογη της δημοτικότητάς της, σε αντίθεση με τον αλγόριθμο PageRank που μοιράζει την τιμή μίας σελίδας ισότιμα σε κάθε εξωτερικό σύνδεσμο. Η δημοτικότητα από τον αριθμό εσωτερικών και εξωτερικών συνδέσμων συμβολίζονται ως $W^{in}_{(v,u)}$ και $W^{out}_{(u,v)}$ και υπολογίζεται βάση των συνδέσμων που κατευθύνονται από τη σελίδα και προς τη σελίδα αντίστοιχα.

WEIGHTED LINKS RANK ALGORITHM [33]

Ο αλγόριθμος Weighted Links Rank (WLRank) είναι μια τροποποίηση του αλγόριθμου PageRank. Αποδίδει αριθμητικές τιμές βάρους σε ένα σύνδεσμο βασιζόμενος σε τρις παραμέτρους. Το HTML tag που χρησιμοποιείται από τη σελίδα για το συγκεκριμένο σύνδεσμο, τη θέση του συνδέσμου μέσα στη σελίδα και το μέγεθος του κειμένου του συνδέσμου. Το μέγεθος του κειμένου του συνδέσμου, σύμφωνα με τις παρατηρήσεις φαίνεται να παράγει τα καλύτερα αποτελέσματα.

EIGENRUMOR ALGORITHM [34]

Οι τιμές που αποδίδονται από τους αλγόριθμους PageRank και HITS σε blog ιστοσελίδες είναι πολύ μικρός και έτσι αποτρέπει την κατάταξή τους βάση της σημαντικότητάς τους. Ο αλγόριθμος αυτός δημιουργήθηκε για τη βαθμολόγηση blog ιστοσελίδων, ώστε να αποφεύγεται το παραπάνω πρόβλημα. Αποδίδει μία τιμή σε κάθε σελίδα υπολογίζοντας τις τιμές αρχής και διανομέα κάθε blogger χρησιμοποιώντας ιδιοδιανύσματα.

[6] Google's Webmaster GuideLines
[7] Βλ. Link Schemes και Link Wheels

DISTANCE RANK ALGORITHM [35]

Σε αυτό τον αλγόριθμο η κατάταξη των σελίδων γίνεται βάση της μικρότερης λογαριθμικής απόστασης ανάμεσα σε δύο ιστοσελίδες. Βασικό πλεονέκτημα αυτού του αλγορίθμου είναι πως μπορεί να εντοπίσει σελίδες πολύ γρήγορα με τη χρήση της μεταξύ τους απόστασης. Το βασικό μειονέκτημα είναι πως ο Crawler πρέπει να κάνει περίπλοκους υπολογισμούς για να υπολογίζει το διάνυσμα απόστασης όταν εμφανίζεται μια νέα σελίδα.

TIME RANK ALGORITHM [36]

Στον αλγόριθμο αυτό, υπολογίζεται ο χρόνος που ένας χρήστης περνά σε μια ιστοσελίδα, και προστίθεται στην τιμή που αποδίδεται στη σελίδα από τον αλγόριθμο PageRank. Βασικό μειονέκτημα του αλγορίθμου είναι πως αυξάνει την κατάταξη σελίδων που μένουν ανοιχτές για μεγάλα χρονικά διαστήματα.

3.1.7. Ευρετηριοποίηση

Η ευρετηριοποίηση των μηχανών αναζήτησης αποτελεί ιδιαίτερο χαρακτηριστικό τους που απαιτεί στοιχειώδη επεξήγηση για την κατανόηση της λειτουργίας των μηχανών αναζήτησης. Η ευρετηριοποίηση είναι το συστατικό που συλλέγει, αναλύει και αποθηκεύει δεδομένα για την διευκόλυνση της γρήγορης και ακριβούς ανάκτησης πληροφοριών. Η σχεδίαση ευρετηρίου ενσωματώνει διεπιστημονικές έννοιες από επιστήμες όπως η γλωσσολογία, η γνωστική ψυχολογία, τα μαθηματικά, η πληροφορική, η φυσική και η επιστήμη υπολογιστών.

Οι δημοφιλείς μηχανές αναζήτησης επικεντρώνονται στην ευρετηριακή ταξινόμηση πλήρους κειμένου φυσικής γλώσσας [37]. Τύποι πολυμέσων, όπως βίντεο, ήχος και γραφικά, είναι επίσης αναζητήσιμα. Οι meta-μηχανές αναζήτησης επαναχρησιμοποιούν τους δείκτες των άλλων υπηρεσιών και δεν αποθηκεύουν ένα τοπικό ευρετήριο, ενώ cache-based μηχανές αναζήτησης αποθηκεύουν μόνιμα το δείκτη μαζί με το σώμα κειμένων. Σε αντίθεση με τους δείκτες πλήρους κειμένου, οι υπηρεσίες επιμέρους-κειμένου περιορίζουν το βάθος ευρετηρίου ώστε να μειωθεί μέγεθος του. Μεγαλύτερες υπηρεσίες επιτελούν κατά κανόνα ευρετηριοποίηση σε ένα προκαθορισμένο χρονικό διάστημα, λόγω του απαιτούμενου χρόνου και κόστους επεξεργασίας.

Ο σκοπός της αποθήκευσης ενός ευρετηρίου είναι η βελτιστοποίηση της ταχύτητας και απόδοσης στην ανεύρεση σχετικών εγγράφων για ένα ερώτημα αναζήτησης. Χωρίς ένα ευρετήριο, η μηχανή αναζήτησης θα σαρώσει κάθε έγγραφο στο σώμα κειμένων, το οποίο θα απαιτήσει σημαντικό χρόνο και υπολογιστική ισχύ. Για παράδειγμα, ενώ ένα ευρετήριο 10.000 εγγράφων μπορεί να αναζητηθεί σε κλάσματα του δευτερολέπτου, μια διαδοχική σάρωση της κάθε λέξης σε 10.000 μεγάλων εγγράφων θα μπορούσε να πάρει ώρες. Ο πρόσθετος αποθηκευτικός χώρος που απαιτείται για την αποθήκευση του ευρετηρίου, καθώς και τη σημαντική αύξηση του χρόνου που απαιτείται για μια ενημέρωση που θα πραγματοποιηθεί, αντισταθμίζονται από την εξοικονόμηση χρόνου κατά τη διάρκεια της ανάκτησης πληροφοριών.

ΑΝΕΣΤΡΑΜΜΈΝΟ ΕΥΡΕΤΉΡΙΟ

Πολλές μηχανές αναζήτησης ενσωματώνουν ένα ανεστραμμένο ευρετήριο κατά την αξιολόγηση ενός ερωτήματος αναζήτησης για να εντοπίζονται γρήγορα τα έγγραφα που περιέχουν τις λέξεις σε ένα ερώτημα και στη συνέχεια κατατάσουν αυτά τα έγγραφα βάση συνάφειας. Επειδή το ανεστραμμένο ευρετήριο αποθηκεύει μια λίστα εγγράφων που περιέχουν κάθε λέξη, η μηχανή αναζήτησης μπορεί να χρησιμοποιήσει απευθείας πρόσβαση για να βρει τα έγγραφα που σχετίζονται με κάθε λέξη στο ερώτημα για τη γρήγορη ανάκτηση των εγγράφων που ταιριάζουν. Στον πίνακα 1, παρουσιάζεται μία απλοποιημένη απεικόνιση ενός ανεστραμμένου ευρετηρίου:

Λέξη	Έγγραφα
the	Έγγραφο 1, Έγγραφο 3, Έγγραφο 4, Έγγραφο 5
cow	Έγγραφο 2, Έγγραφο 3, Έγγραφο 4
says	Έγγραφο 5
moo	Έγγραφο 7

Πίνακας 1: Παράδειγμα Ανεστραμένου Ευρετηρίου

Το ευρετήριο αυτό μπορεί μόνο να καθορίσει αν μια λέξη υπάρχει σε ένα συγκεκριμένο έγγραφο, δεδομένου ότι δεν αποθηκεύει πληροφορίες σχετικά με τη συχνότητα και τη θέση της λέξης, ως εκ τούτου θεωρείται ένα λογικό ευρετήριο. Ένα τέτοιο ευρετήριο καθορίζει ποια έγγραφα ταιριάζουν με ένα ερώτημα, αλλά δεν ταξινομεί που ταιριάζουν τα έγγραφα. Σε ορισμένα σχέδια ο δείκτης περιλαμβάνει πρόσθετες πληροφορίες όπως η συχνότητα της κάθε λέξης σε κάθε έγγραφο ή τις θέσεις της λέξης σε κάθε έγγραφο [38]. Οι πληροφορίες θέσης επιτρέπουν στον αλγόριθμο αναζήτησης να εντοπίσει την εγγύτητα της λέξης για να υποστηρίξει την αναζήτηση φράσεων. Η συχνότητα μπορεί να χρησιμοποιηθεί για να βοηθήσει στην ιεράρχηση της συνάφειας των εγγράφων στο ερώτημα.

Το ανεστραμμένο ευρετήριο είναι ένας αραιός πίνακας, δεδομένου ότι δεν είναι όλες οι λέξεις παρούσες σε κάθε έγγραφο. Το ευρετήριο είναι παρόμοιο με τον πίνακα εγγράφων, όρος που χρησιμοποιείται από τη σημασιολογική ανάλυση. Το ανεστραμμένο ευρετήριο μπορεί να θεωρηθεί μια μορφή πίνακα κατακερματισμού. Σε ορισμένες περιπτώσεις, το ευρετήριο είναι ένα είδος δυαδικού δέντρου, το οποίο απαιτεί επιπλέον αποθηκευτικό χώρο, αλλά μπορεί να μειώσει τον χρόνο αναζήτησης. Σε μεγαλύτερα ευρετήρια η αρχιτεκτονική που χρησιμοποιείται είναι κατά κανόνα ένας κατανεμημένος πίνακας κατακερματισμού [39].

Ορθό Ευρετήριο

Το ορθό ευρετήριο αποθηκεύει μία λίστα λέξεων για κάθε έγγραφο. Ο πίνακας 2, αποτελεί μία απλοποιημένη μορφή ορθού ευρετηρίου:

Έγγραφο	Λέξεις
Έγγραφο 1	the, cow, says, moo
Έγγραφο 2	the, cat, and, the, hat
Έγγραφο 3	the, dish, ran, away, with, the, spoon

Πίνακας 2: Παράδειγμα Ορθού Ευρετηρίου

Η λογική πίσω από την ανάπτυξη ενός ορθού ευρετηρίου είναι ότι καθώς τα έγγραφα αναλύονται, είναι προτιμότερο να αποθηκεύονται αμέσως οι λέξεις ανά έγγραφο. Η οριοθέτηση επιτρέπει την ασύγχρονη επεξεργασία στο σύστημα, το οποίο παρακάμπτει μερικώς την συμφόρηση ενημέρωσης του ανεστραμμένου ευρετηρίου [18]. Το ορθό ευρετήριο είναι ουσιαστικά μια λίστα, τα ζεύγη της οποίας αποτελούνται από ένα έγγραφο και μια λέξη, τα οποία συγκεντρώνονται από το έγγραφο. Μετατρέποντας το ορθό ευρετήριο σε ανεστραμμένο είναι απλώς θέμα ταξινόμησης των ζευγών σύμφωνα με τις λέξεις.

Ανάλυση Εγγράφων

Η ανάλυση του εγγράφου χωρίζει τα στοιχεία του εγγράφου, ή κάποιου άλλου είδους πολυμέσου, για την καταχώρηση σε κάποιο ορθό ή ανεστραμμένο ευρετήριο. Οι λέξεις που βρέθηκαν ονομάζονται τεκμήρια, και έτσι, στο πλαίσιο της ευρετηριοποίησης και επεξεργασίας της φυσικής γλώσσας, η ανάλυση αναφέρεται πιο συχνά ως αναγνώριση λέξεων. Η αναγνώριση λέξεων ευρετηρίων περιλαμβάνει πολλές τεχνολογίες, η εφαρμογή των οποίων είναι συνήθως εταιρικό μυστικό.

Αναγνώριση Λέξεων

Κατά την αναγνώριση λέξεων, ο μεταγλωττιστής αναγνωρίζει αλληλουχίες χαρακτήρων που αναπαριστούν λέξεις και άλλα στοιχεία, όπως σημεία στίξης, που

αναπαριστώνται από αριθμητικούς κώδικες, κάποιοι από τους οποίους είναι μη-εκτυπώσιμοι χαρακτήρες ελέγχου. Ο μεταγλωττιστής μπορεί επίσης να αναγνωρίσει οντότητες όπως διευθύνσεις ηλεκτρονικού ταχυδρομείου, τηλεφωνικούς αριθμούς, και URLs. Όταν αναγνωρίζεται κάθε λέξη, διάφορα χαρακτηριστικά μπορεί να αποθηκευτούν, όπως γλώσσα, κωδικοποίηση, λεξιλογική κατηγορία (για παράδειγμα 'επίθετο' ή 'ρήμα'), θέση, αριθμός πρότασης, θέση πρότασης, μέγεθος, και αριθμός γραμμής.

ΑΝΑΓΝΩΡΙΣΗ ΤΜΗΜΑΤΩΝ

Κάποιες μηχανές αναζήτησης περιλαμβάνουν αναγνώριση τμημάτων, πριν την αναγνώριση λέξεων. Πολλά έγγραφα στο Διαδίκτυο, όπως εταιρικές αναφορές, περιέχουν τμήματα που δεν περιέχουν υλικό σχετικό με το κείμενο. Παρόλο που το περιεχόμενο προβάλλεται σε διαφορετικές θέσεις, η πληροφορία μπορεί να αποθηκεύεται διαδοχικά. Αν η μηχανή αναζήτησης τοποθετήσει τα δεδομένα αυτά με τα δεδομένα του εγγράφου, η ποιότητα του ευρετηρίου και η ποιότητα της αναζήτησης, υποβαθμίζεται σημαντικά.

3.1.8. Μέθοδοι Χειραγώγησης

3.1.8.1. Υπερβολική Χρήση Λέξεων Κλειδιών [40] [41]

Η υπερβολική χρήση λέξεων κλειδιών επιτυγχάνεται με την υπερφόρτωση λέξεων κλειδιών είτε στις ετικέτες <meta>, είτε στο περιεχόμενο της σελίδας. Στην Εικόνα 12, παρουσιάζεται ένα παράδειγμα υπερφόρτωσης λέξεων-κλειδιών σε μια ιστοσελίδα.


no hands seo review, download No Hands SEO, no hands seo, auto approved list for no hands seo, "no hands seo" download, No Hands SEO download, no hands seo software download, linxbot alternative, no hands seo download, no hand seo tutorial, everquest link bot, free LinxBot, TweeterNaire , auto seo backlink software, No hands SEO review, scrapebox nohandsseo, download linkbot backlink, 0h, what does autobacklink bomb do, [GET] no hands seo, no hands seo backlinksforum, tweeternaire, backlink bot scrapebox, tutoriel no hands seo, best backlinks sofftware, No Hands SEO, Auto Backlink Bomb dl, nohandsseo review, which one is better scrapebox or nohandsseo, "no hands seo" forum, software for SEO link building, backlinks software download, no hands backlinking software, no hands seo results, nohandseo.blogspot, No Hands SEO software, get no hands seo, auto backlink bomb index fast?, no hands seo mediafire, software backlinks building, linxbot, auto backlink bomb tutorials, no hand seo, no hand seo in mediafire, Download Auto backlink Bomb, download no hands seo , has anyone used linkbot#sclient=psy, nohandseo price, LinxBot â, LinxBot â, LinxBot â, no hands seo opinions, tweeternaire review, nohandseo, inurl:forum no hand seo, LinxBot megaupload , auto backlink bomb download, top backlink building software reviews, nohandseo software, NOHANDSSEO vs scrapebox, no hands seo free, banned no hands seo, no hand seo softwear, f, auto backlink software, tutorial auto blacklink, How to use No Hands SEO software, Tweeternaire review, No Hands SEO megaupload, free squidoo linkposting software#q=free automatic high pr link building software, "no hands seo" use approve , no hands seo blogspot, no hands seo#pq=no hands seo, SEO applications, seo software forum, nohandsseo tutorial, Tweeternaire, linxbot tutorial, download smf forum txt backlink, "free linxbot", scrape High PR websites software, no hands seo tutorial, What is "No Hands SEO"?, Submit and Share your sites, news and stories, Submit and Share your sites, news and stories, Submit and Share your sites, news and stories, Submit and Share your sites, news and stories, no hands seo rapidshare, seo rapidshare, nohand seo review, [Get] NO HANDS SEO, get tweeternaire download, auto backlink bomb review, forum links for no hands seo, "Auto Backlink Bomb" hotfile, tweeternaire filesonic, [get]no hands seo mediafile, back link bot, mediafire backlink software, No_Hands_SEO.rar, linxbot vs scrapebox, "No hands seo", No Hands SEO rar, no hands seo rapidshare download, TweeterNaire#sclient=psy-ab, earn money with tweeternaire, tweeternaire backlinks, TweeterNaire, buy linxbot, I used tweeternaire and my account was banned, backlink software rar, no handsseo video tutoials, TweeterNaire download, no hands seo download blogspot, high PR SEO Forum list, seo software auto backlink bomb rar, Auto Backlink Bomb rar, auto backlink bomb mediafire, seo software link building -directory, NOHANDSEO, best link building hands free, Backlink Building And Pinging Software mediafire, no hands seo filesonic, AutoBacklinkBomb .rar, get No Hands SEO, auto seo free, filesonic seo software, linxbot negative reviews, No hands SEO software, mp hands seo revioew#sclient=psy-ab, no hands seo megaupload, no hands seo forum, the best link building software that actually works, download:nohandseo +.rar, linxbot download, auto approve list no hands seo, no hands seo vs scrapebox, auto comment bomb hotfile, Free no hands SEO, No Hands SEO filesonic rapidshare megaupload, No Hands SEO rapidshare, mediafire seo link building software, linx bot download#sclient=psy-ab, LinxBot.rar -filestube, forum hands no seo, no hand seo review , Auto Backlink Bomb rapidshare, auto backlink bomb rapidshare, autobacklinkbomb download, download no hands seo, no hand seo opinion, LinxBot latest version rapidshare, LinxBot free download, SEO hand on tutorial, download backlinks spftware.rar, seo backlink bomb.rar, AUTO BACK LINKS SOFTWARE mediafire, yahoo, how to use no hands seo, auto backlink bomb medIAFIRE LINKS, autobacklink bomb mediafire links, No Hands SEO â, autobacklinkbomb warez, backlink bomb hotfile, auto backlink bomb hotfile, best auto backlink program, best wordpress themes, SEO auto link bot, download nohandseo, seo softwares, backlink filesonic, best backlink software, best seo software, Auto backlink Bomb, review nohandseo, "auto link bot", , autolinkbot review, "no hands seo.rar", backlink+megaupload, "autobacklinkbomb.rar", backlink mediafire, auto backlink, smf backlink, backlink megaupload, seo filesonic, autolink bot, profile multithread seo, tweeternaire mediafire, the best automated backlink seo software, "backlinksoftware.rar", index/of autolinkbot.zip, autobacklinkbomb no hand seo, free high pr backlinks list, tweeternaire warez, backlink + rapidshare download, i want link building free software, auto "twitter marketing software", link building mediafire, "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", "Submit and Share your sites, news and stories", blackhatworld pagerank backlinks, get scrapebox mediafire, warez backlinkuri, anything, tweeternaire download, free auto link building software, auto seo backlinks mediafire, scrapebox.rar seo, autolinkbot reviews, software backlink warez, free backlink bot, TOP SITE AUTO BACKLINKS, autolinkbot software reviews, keywords no hands seo, best seo software auto backlinks blog posts, tweeternaire reviews, high pr backlinks anchor text software free download, backlinks building software rapidshare, seo backlink mf, autolink bot download, bomb top softwere, seo or software engineer? which is best, tutorial no hands seo, seo software in hotfile, high pr cheap mobile blog list mediafire, no hand no seo backlink tutorial, seo, auto backlinks bomb, linxbot free download, how fast is no hand seo, seo link building bot download, no hands seo.rar, high pr links software, scrapebox.rar mediafire, software.mediafire, how to index backlinks from rss, linxbot rapidshare, high pr comments software, free

Εικόνα 12: Παράδειγμα Spamdexing

Υπάρχουν πολλές τεχνικές στην απόκρυψη των λέξεων κλειδιών μέσα στο περιεχόμενο της ιστοσελίδας, ώστε να μην γίνονται αντιληπτές από τους χρήστες. Το κείμενο συνήθως χρωματίζεται μέσω CSS σε χρώμα πανομοιότυπο με το χρώμα υποβάθρου, ή τοποθετείται πίσω από εικόνες με τη χρήση της ιδιότητας Z-index. Ακόμη το κείμενο μπορεί να βγει από το πεδίο που βλέπει ο χρήστης, κάνοντας το έτσι αόρατο.

Μία άλλη τεχνική είναι η χρήση λέξεων-κλειδιών μέσα στο κείμενο, οι οποίες έχουν μικρή ή καθόλου συνάφεια με το περιεχόμενο, και είναι συχνά αναζητήσιμες, με επικρατέστερη τη λέξη sex. Άλλα παραδείγματα keyword stuffing είναι η συχνή επανάληψη λέξεων, όπως για παράδειγμα: We sell custom cigar humidors. Our custom cigar humidors are handmade. If you're thinking of buying a custom cigar humidor, please contact our custom cigar humidor specialists at custom.cigar.humidors@example.com. Η απαρίθμηση πόλεων και χωρών στις οποίες η σελίδα θέλει να επιτύχει υψηλή κατάταξη και η απαρίθμηση τηλεφωνικών αριθμών που δεν έχουν σχέση με τη σελίδα.

3.1.8.2. Cloaking

Το Cloaking, είναι μια τεχνική βελτιστοποίησης της θέσης κατάταξης στα αποτελέσματα των μηχανών αναζήτησης, με την οποία το περιεχόμενο που παρουσιάζεται στη μηχανή αναζήτησης είναι διαφορετικό από εκείνο που παρουσιάζεται στην απλή επίσκεψη των web surfers. Όταν ένας χρήστης έχει χαρακτηρισθεί ως μηχανή αναζήτησης αράχνη ένα script παραδίδει μια διαφορετική εκδοχή της ιστοσελίδας, η οποία περιέχει ένα περιεχόμενο που δεν είναι ορατό στη μορφή της σελίδας που βλέπουν οι επισκέπτες. Στις μηχανές αναζήτησης δεν αρέσει η τεχνική της απόκρυψης επειδή ο σκοπός της είναι να εξαπατήσει μηχανές αναζήτησης. Αν το Google εντοπίσει ότι ένα website χρησιμοποιεί απόκρυψη, θα καταργήσει οριστικά την ιστοσελίδα από το ευρετήριο.

3.1.8.3. Link Bombing

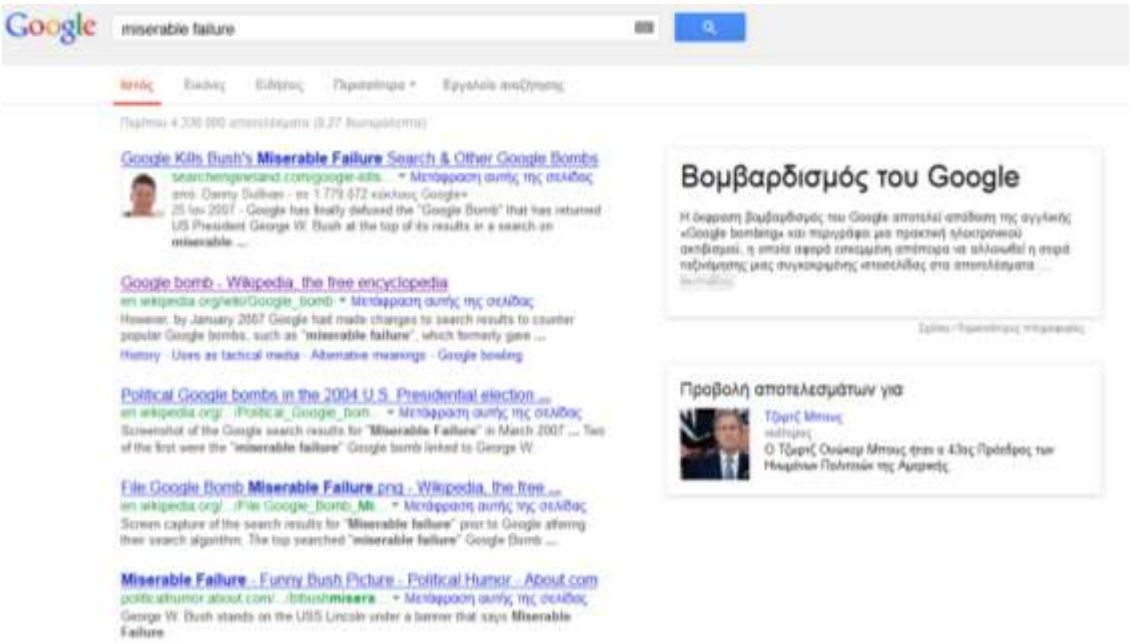
Το Link Bombing περιγράφει μια πρακτική ηλεκτρονικού ακτιβισμού η οποία αφορά εσκεμμένη απόπειρα να αλλοιωθεί η σειρά ταξινόμησης μιας συγκεκριμένης ιστοσελίδας στα αποτελέσματα που παράγονται από τη μηχανή αναζήτησης Google.

Η ύπαρξη του αλγόριθμου PageRank δημιούργησε την ιδέα ότι θα μπορούσε κανείς να «οδηγήσει» την αναζήτηση, με βάση έναν όρο προσβλητικό, σε κάποια ιστοσελίδα. Αυτός ο τρόπος «προσβολής» ονομάστηκε βόμβα Google. Μια βόμβα Google κατασκευάζεται όταν ένα μεγάλο πλήθος ιστοχώρων συνδέουν στη σελίδα αυτή με αυτό τον τρόπο, όχι τυχαία, αλλά με σκοπό να επηρεάσουν τα αποτελέσματα της μηχανής αναζήτησης.

Τα google bombs οργανώνονται ανεπίσημα μεταξύ κατόχων ιστολογιών (blogs) ή άλλων ιστότοπων, με συμφωνία και εθελοντική τοποθέτηση τέτοιων συνδέσμων με το ίδιο κείμενο και προορισμό τον ίδιο ιστότοπο. Συνήθως πραγματοποιούνται είτε ως αστείο, είτε για τη διαμαρτυρία ή την προώθηση ενός μηνύματος με κοινωνικό ή πολιτικό περιεχόμενο. Χρησιμοποιούνται επίσης από εμπορικούς ιστοτόπους, συνήθως ενσωματώνοντας τους συνδέσμους σε ιστότοπους τρίτων όπου επιτρέπεται κάποιο είδος καταχώρησης όπως βιβλία επισκεπτών, ή ανοικτά wiki, κάτι που χαρακτηρίζεται ως spam, και για την καταπολέμηση αυτού του φαινομένου έχουν δημιουργηθεί διάφοροι τρόποι φιλτραρίσματος των καταχωρήσεων ή ακύρωσης των συνδέσμων.

Η Google, από την πλευρά της, προσπαθεί να καταπολεμήσει τέτοιου είδους προσπάθειες. Έτσι, τα αποτελέσματα της αναζήτησης διορθώνονται αμέσως μόλις εντοπιστεί κάποια προσπάθεια εξαπάτησης της μηχανής αναζήτησης.Τα παραδείγματα που ακολουθούν έμειναν στο Διαδίκτυο από μία με δύο βδομάδες μέχρι 2 μήνες το μέγιστο.

Το πρώτο Google Bomb ξέσπασε το 1999, που προκάλεσε την αρχική σελίδα του Google να κατακτήσει την πρώτη θέση με τις λέξεις-κλειδιά: «more evil than Satan himself» («περισσότερο κακό απ'ότι ο ίδιος ο σατανάς»). Ένα άλλο επίσης γνωστό Google Bomb ήταν το 2004, όπου στις λέξεις-κλειδιά: «παταγώδης αποτυχία» εμφανιζόταν πρώτο πρώτο το βιογραφικό του George W. Bush (βλέπε Εικόνα 13)!



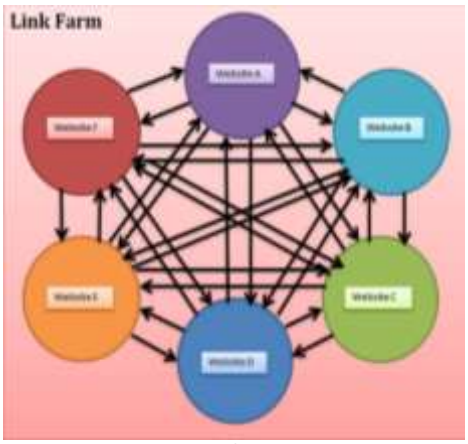
Εικόνα 13: Παράδειγμα Link Bombing

3.1.8.4. Scraper Sites

Ένα scraper site είναι μια ιστοσελίδα η οποία τραβά το περιεχόμενο από άλλες πηγές και το αναδημοσιεύει. Τα περισσότερα scraper sites παραβιάζουν δικαιώματα πνευματικής ιδιοκτησίας νόμου από την επανεκτύπωση του περιεχομένου χωρίς τη συγκατάθεση του συγγραφέα, και προκαλούν επίσης τον όλεθρο στα αποτελέσματα των μηχανών αναζήτησης και τις βαθμολογίες των site, πράγμα το οποίο δυσκολεύει τους χρήστες του Διαδικτύου να βρούν τις τοποθεσίες που πραγματικά θέλουν να δουν.

3.1.8.5. Link Farms

Στο World Wide Web, ένα link farm είναι οποιαδήποτε ομάδα ιστοσελίδων οι οποίες έχουν υπερσυνδέσμους προς κάθε άλλη τοποθεσία στην ομάδα (βλέπε Εικόνα 14). Παρά το γεγονός ότι ορισμένα link farms μπορούν να δημιουργηθούν με το χέρι, οι περισσότερες δημιουργούνται μέσω αυτοματοποιημένων προγραμμάτων και υπηρεσιών.



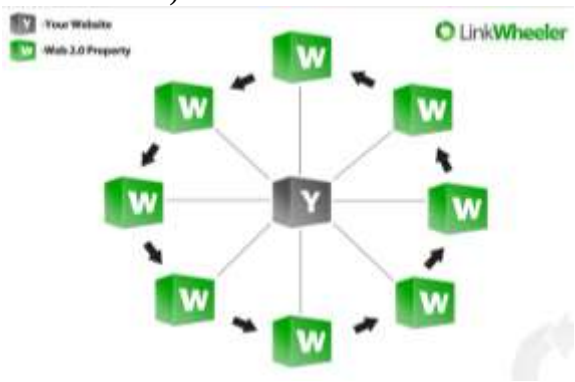
Εικόνα 14: Παράδειγμα Link Farm

Ένα link farm είναι μια μορφή spamming του ευρετηρίου μιας μηχανής αναζήτησης (μερικές φορές ονομάζεται spamdexing ή spamexing). Οι μηχανές αναζήτησης απαιτούν τρόπους να επιβεβαιώσουν τη σχετικότητα της σελίδας. Μία γνωστή μέθοδος είναι η εξέταση

για μονόδρομες συνδέσεις που προέρχονται απευθείας από σχετικές ιστοσελίδες. Η διαδικασία της οικοδόμησης συνδέσμων δεν πρέπει να συγχέεται με την καταχώρησή τους σε link farms, καθώς το τελευταίο απαιτεί αμοιβαίες συνδέσεις επιστροφής, που συχνά καθιστά το συνολικό όφελος backlink άχρηστο.

3.1.8.6. Link Wheels

Το Link Wheel είναι μια ευρέως χρησιμοποιούμενη στρατηγική στο Διαδικτυακό μάρκετινγκ. Η βασική ιδέα του Link Wheel είναι να δημιουργηθεί ένα μοτίβο συνδέσμων το οποίο ανακατευθύνει το χρήστη από τη μια ιστοσελίδα στην άλλη και τελικά τον οδηγεί στην ιστοσελίδα στόχου (βλέπε Εικόνα 15).



Εικόνα 15: : Παράδειγμα Link Wheel

Το Link Wheel εμπίπτει στην κατηγορία της βελτιστοποίησης των μηχανών αναζήτησης, ωστόσο, όταν οι ιστοσελίδες κοινωνικής δικτύωσης συνηθίζουν να δημιουργούν τη ροή συνδέσμων, η στρατηγική του Link Wheel κατηγοριοποιείται ως μάρκετινγκ κοινωνικών δικτύων. Σύμφωνα με τους ειδικούς, τα Link Wheels έχουν τη μέγιστη δυνατή αποτελεσματικότητα όταν χρησιμοποιούνται οι ιστοσελίδες κοινωνικών δικτύων.

Το μοτίβο του Link Wheel

Στην πραγματικότητα δεν υπάρχει κάποιο συγκεκριμένο πρότυπο ή κανόνας όταν πρόκειται για τη διαμόρφωση ενός Link Wheel. Η αποτελεσματικότητα ενός Link Wheel εξαρτάται από το πόσο οργανικό είναι το μοτίβο. Οι ειδικοί του SEO πιστεύουν ότι τα Link Wheels δεν πρέπει να σχεδιάζονται για να ξεγελάσουν τις μηχανές αναζήτησης, αλλά θα πρέπει να εξασφαλίζουν μια φυσική ροή από τη μία ιστοσελίδα στην άλλη.

3.2. Παραδείγματα μηχανών αναζήτησης

3.2.1. Η Google [2]

Η Google είναι μια από τις μεγαλύτερες εταιρείες Διαδικτυακών υπηρεσιών. Η λειτουργία της ξεκίνησε το Σεπτέμβρη του 1998. Ο στόχος της είναι να οργανώσει όλες τις πληροφορίες του κόσμου και να τις κάνει παγκόσμια διαθέσιμες. Το Google ξεκίνησε σαν μια κολεγιακή εργασία από τον Λάρρυ Πέιτζ και τον Σεργκέι Μπρίν το 1996 για μια μηχανή αναζήτησης. Χρησιμοποιεί έναν αλγόριθμο ανάλυσης συνδέσμων (PageRank) ο οποίος ορίζει μια αριθμητική στάθμιση σε κάθε σημείο ενός συνόλου εγγράφων, όπως είναι το World Wide Web, με σκοπό να μετρήσει την ανάλογη σημασία του μέσα στο σύνολο. Με άλλα λόγια τα αποτελέσματα του PageRank προκύπτουν από το πόσο σημαντική είναι μια σελίδα στο World Wide Web. Ένας σύνδεσμος υπερκειμένου σε μια σελίδα προσμετράται σαν ψήφος εμπιστοσύνης. Το PageRank μιας ιστοσελίδας καθορίζεται κατ' επανάληψη και εξαρτάται από τον αριθμό και την τιμή του PageRank όλων των σελίδων που δείχνουν σε αυτήν. Μια σελίδα που συνδέεται με πολλές σελίδες με υψηλό PageRank λαμβάνει η ίδια ένα υψηλό

PageRank. Εάν δεν υπάρχουν σύνδεσμοι προς μια ιστοσελίδα δεν υπάρχει τιμή PageRank γι' αυτήν τη σελίδα.

3.2.2. Το Yahoo! [2]

Το Yahoo είναι μια εταιρεία Διαδικτυακών υπηρεσιών. Είναι ένας από τους πιο γνωστούς και παλιούς θεματικούς καταλόγους του Διαδικτύου. Αν και ξεκίνησε ως θεματικός κατάλογος, αργότερα εξελίχθηκε και σε μια πανίσχυρη μηχανή αναζήτησης. Προσφέρει στους χρήστες του ένα μεγάλο αριθμό υπηρεσιών που περιλαμβάνουν ηλεκτρονικό ταχυδρομείο, μηχανή αναζήτησης, ομάδες χρηστών, νέα, παιχνίδια, διαφημίσεις και επίσης ένα πρόγραμμα για άμεσα ηλεκτρονικά μηνύματα, το Yahoo! Messenger.

Ιδρύθηκε τον Ιανουάριο του 1994 από τους τότε τελειόφοιτους του πανεπιστημίου Stanford, David Filo και Jerry Yang, όταν θέλανε να ομαδοποιήσουν και να καταγράψουν Διαδικτυακές τοποθεσίες μείζονος ενδιαφέροντος και να τις ταξινομήσουν σε θεματικές ενότητες. Αρχικά ο θεματικός κατάλογος ήταν δημοσιευμένος στο δικτυακό τόπο του Stanford.

4. ΑΝΑΔΥΟΜΕΝΕΣ ΤΑΣΕΙΣ: ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΑΝΑΖΗΤΗΣΗ ΚΑΙ ΑΝΑΖΗΤΗΣΗ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

4.1. Εισαγωγή στη Σημασιολογική Έρευνα (Semantic Search Introduction)

Η σημασιολογική έρευνα (semantic search) αποτελεί μια εφαρμογή έρευνας του σημασιολογικού ιστού (semantic web). Η έρευνα μέσα στο Διαδίκτυο είναι μια από τις πιο δημοφιλείς εφαρμογές με μεγάλες προοπτικές βελτίωσης. Η προσθήκη ρητής σημασιολογίας είναι δυνατό να βελτιώσει τα αποτελέσματα της τρέχουσας έρευνας στο παραδοσιακό παγκόσμιο ιστό, μέσω της χρησιμοποίησης δεδομένων από το σημασιολογικό ιστό. Η παραδοσιακή τεχνολογία ανάκτησης της πληροφορίας (Information Retrieval technology) βασίζεται σχεδόν αποκλειστικά στη συχνότητα εμφάνισης συγκεκριμένων λέξεων μέσα σε έγγραφα. Οι μηχανές αναζήτησης, όπως το Google, αυξάνουν αυτή τη δυνατότητα στα πλαίσια του παγκόσμιου ιστού παρέχοντας πληροφορίες σχετικές με τη δομή των υπερσυνδέσμων (hyperlinks) του παγκόσμιου ιστού. Η διαθεσιμότητα μεγάλων ποσών δομημένης, μηχανικά αναγνώσιμης πληροφορίας γύρω από ένα ευρύ φάσμα αντικειμένων του σημασιολογικού ιστού προσφέρει δυνατότητες βελτίωσης της παραδοσιακής έρευνας.

4.2. Επαύξηση Παραδοσιακής Αναζήτησης Λέξεων-Κλειδιών με Σημασιολογικές Τεχνικές

Πολλές εφαρμογές επέκτασης ερωτημάτων που χρησιμοποιούνται στην αναζήτηση λέξεων- κλειδιών κάνουν χρήση της πλοήγησης ενός οντολογικού λεξικού για την επέκταση των ερωτημάτων. Ειδικότερα χρησιμοποιείται η οντολογία WorldNet, η οποία προσδιορίζει συνώνυμα και μερώνυμα[8] σύνολα λέξεων. Τα συστήματα αυτά λειτουργούν κάτω από την ίδια βασικά αρχή: Αρχικά, οι λέξεις κλειδιά εντοπίζονται μέσα στην οντολογία, έπειτα, διάφορες άλλες έννοιες εντοπίζονται μέσω διάσχισης γραφήματος, μετά την οποία οι όροι που σχετίζονται με αυτές τις έννοιες αξιοποιούνται είτε για να διερευνηθεί, είτε για να περιοριστεί η αναζήτηση.

Σε κάποια συστήματα([42], [43]), οι όροι της αναζήτησης διευρύνονται μέσω των συνωνύμων και των μερώνυμων με τη χρήση του λογικού τελεστή OR που υποστηρίζονται από τις περισσότερες μηχανές αναζήτησης. Το Clever Search [44], δίνει τη δυνατότητα να επιλεχθεί κάποια συγκεκριμένη έννοια μιας λέξης μέσα από το WordNet, δημιουργώντας έτσι ένα επεξηγηματικό κείμενο της έννοιας, το οποίο προστίθεται στους όρους αναζήτησης με τη χρήση του λογικού τελεστή OR.

Στο έργο των R.Guca et.al. [45], εκτός από την παραδοσιακή αναζήτηση με λέξεις-κλειδιά σε μια βάση δεδομένων εγγραφών, οι λέξεις κλειδιά αντιστοιχούνται με εννοιολογικές ετικέτες μέσα σε ένα RDF αποθετήριο. Επίσης αν πολλαπλές έννοιες αντιστοιχούνται σε κάποιο όρο, ο χρήστης μπορεί να επιλέξει την έννοια που επιθυμεί για να περιορίσει την αναζήτηση.

Σε μία υβριδική προσέγγιση [46],παρουσιάζεται ένας αλγόριθμος για τον εντοπισμό επιπλέον πληροφοριών, σχετικές με ένα ερώτημα, όταν δίνεται ένα σύνολο δεδομένων. Αρχικά μία παραδοσιακή αναζήτηση, γίνεται σε μία συλλογή εγγραφών. Έπειτα μία διεργασία διάσχισης γραφήματος RDF ξεκινά από τα σχόλια των εγγραφών. Σκοπός είναι να βρεθούν σχετικές έννοιες όπως ο συγγραφέας του εγγράφου, ο εκδότης κλπ. με γενικό τρόπο.

[8] Μερώνυμο: Ένα μερώνυμο υποδηλώνει μέρος ενός συνόλου, ή συστατικό μέρος.

4.2.1. Εντοπισμός Βασικών Εννοιών

Συνήθως τα δεδομένα στο σημασιολογικό ιστό χωρίζονται σε δύο κλάσεις: οντολογικά και instance δεδομένα. Τα δεδομένα για τα οποία ενδιαφέρεται ο χρήστης ανήκουν σε μία κλάση, αλλά το πεδίο γνώσης και οι σχέσεις περιγράφονται σαν σχέσεις της κλάσης μέσα στην οντολογία.

Στο σύστημα αναζήτησης SHOE [47], αρχικά δίνεται στο χρήστη μία απεικόνιση του επαγωγικού δέντρου των κλάσεων μέσα στην οντολογία, από το οποίο μπορεί να επιλέξει τα στιγμιότυπα που ψάχνει. Έπειτα, αναζητώνται οι πιθανές σχέσεις ή properties που σχετίζονται με την κλάση και παρουσιάζεται μία φόρμα που επιτρέπει στο χρήστη να περιορίσει το στόχο των στιγμιότυπων, εφαρμόζοντας φίλτρα λέξεων-κλειδιών σε διάφορα properties των στιγμιότυπων. Όταν τα properties δείχνουν προς αντικείμενα, ο στόχος του φιλτραρίσματος θα είναι η ετικέτα του αναφερόμενου αντικειμένου. Τα ερωτήματα που παράγονται από αυτό το σύστημα είναι της μορφής "find all publications with a particular author name, from a particular project". Παρόμοια προσέγγιση χρησιμοποιείται από κάποιες εκδόσεις του SEAL portal tool [48]. Μια άλλη προσέγγιση, είναι η πολύπλευρη αναζήτηση [49]. Αυτή είναι η προσέγγιση που χρησιμοποιείται από τις αναζητήσεις των portals που βασίζονται στο OntoViews [50] και στο SWED [51] directory portal. Στην πολύπλευρη αναζήτηση, παρέχονται στα δεδομένα πολλαπλές διακριτές όψεις. Ένα παράδειγμα χρήσης είναι το OntoViews-based portal Museum Finland [52], όπου τα αντικείμενα είναι μουσειακά κομμάτια, και στον χρήστη παρουσιάζονται όψεις όπως υλικό κατασκευής, τόπος κατασκευής και τρόπος χρήσης. Σε κάποιες εκδόσεις του OntoViews χρησιμοποιείται μία έννοια που ονομάζεται σημασιολογική αυτόματη συμπλήρωση [53] , η οποία κάνει χρήση της αναζήτησης λέξεων-κλειδιών ως προοίμιο της οντολογικής πλοήγησης. Η ιδέα χρησιμοποιείται στο μέγιστο βαθμό, από το portal Veturi [54].

4.2.2. Σύνθετοι Περιορισμοί Ερωτημάτων

Πολλά είδη πολύπλοκων ερωτημάτων μπορούν να διαμορφωθούν ως εντοπισμός μιας ομάδας αντικειμένων ορισμένου τύπου που συνδέονται με ορισμένες σχέσεις. Στο σημασιολογικό ιστό, αυτό μεταφράζεται σε μοτίβα γραφήματος με περιορισμένους κόμβους αντικειμένου και τόξα τύπων ιδιοτήτων. Ένα παράδειγμα θα μπορούσε να είναι "Εντόπισε όλα τα παιχνίδια που κατασκευάστηκαν στην Ευρώπη το 19ο αιώνα και χρησιμοποιήθηκε από κάποιον τον 20ο αιώνα", όπου "παιχνίδια","Ευρώπη","18ος αιώνας"," κάποιος" και "20ος αιώνας" είναι περιορισμοί οντολογικών κλάσεων στους κόμβους και "κατασκευάστηκε στην"," χρησιμοποιήθηκε από" και "ημερομηνία γέννησης" είναι τα απαιτούμενα τόξα σύνδεσης στο μοτίβο. Ενώ τέτοια μοτίβα είναι εύκολο να επικυρωθούν και να γίνουν ερωτήματα στο πλαίσιο του σημασιολογικού ιστού, παραμένουν προβληματικά επειδή δεν είναι εύκολο για τους χρήστες να τα διατυπώσουν άμεσα.

Το έργο των Athanasis N. et. al. [55], παρουσιάζει την GRQL, μία γραφική διεπαφή χρήστη υπολογιστή για τη δημιουργία ερωτημάτων γραφικών μοτίβων που βασίζεται στην πλοήγηση της οντολογίας. Αρχικά, κάποια κλάση στην οντολογία επιλέγεται σαν αφετηρία. Όλες οι ιδιότητες που ορίζονται ως ισχύουσες σε μία κλάση της οντολογίας προωθούνται προς επέκταση. Η επιλογή μίας ιδιότητας επεκτείνει το γραφικό μοτίβο για να περιέχει την ιδιότητα. Εκτός από την επιμήκυνση της διαδρομής, μπορούν να εκτελεστούν και άλλες λειτουργίες στο μοτίβο ερωτήματος. Το μοτίβο μπορεί να περιοριστεί σε μερικές μόνο υποκλάσεις. Με παρόμοιο τρόπο, οι ιδιότητες μπορούν να περιοριστούν σε υπο-ιδιότητες.

4.2.3. Επίλυση Προβλημάτων

Η περιγραφή ενός προβλήματος, και η αναζήτηση λύσης μέσα από οντολογική γνώση είναι μία βασική περίπτωση χρήσης του σημασιολογικού ιστού. Ωστόσο, τέτοιες πραγματικές εφαρμογές είναι σπάνιες και πολύ απλοϊκές.

Οι Fikes et. al. [56] περιγράφουν μία γλώσσα ερωτημάτων για τον σημασιολογικό ιστό, η οποία παρά το γεγονός ότι προορίζονται ως επί το πλείστον για απλούστερα ερωτήματα της μορφής SQL, βασίζεται σε ένα DL-reasoner, και επιτρέπει ερωτήματα της μορφής "if-then". Αυτή η λειτουργικότητα με τη σειρά της χρησιμοποιείται από το Wine Agent demonstration portal . Εκεί ο χρήστης εισάγει πληροφορίες για τις γεύσεις ενός πιάτου, και το σύστημα, βασιζόμενο σε οντολογικές γνώσεις παράγει μία πρόταση για το ποιό κρασί ταιριάζει με το πιάτο.

4.2.4. Ανακάλυψη Συνδεόμενων Μονοπατιών

Ενώ συνήθως οι σχέσεις ιδιοτήτων χρησιμοποιούνται για τη μετακίνηση από μία ενδιαφέρουσα πηγή στην επόμενη, κάποιες φορές αυτό που είναι ενδιαφέρον είναι τα ίδια τα μονοπάτια μέσα στο γράφο. Στο σημασιολογικό ιστό μία τεράστια ποσότητα από ποικίλα σημασιολογικά δεδομένα είναι διαθέσιμα μέσα στις σημασιολογικές συνδέσεις.

Ένα μείζων πρόβλημα είναι ο καθορισμός του πόσο ενδιαφέρον είναι ένας σύνδεσμος, έτσι ώστε να εξαλείφονται αδιάφορες σχέσεις, αλλά και να είναι και αρκετά γενικές ώστε να είναι χρήσιμες στον εντοπισμό περίπλοκων, κρυφών σχέσεων ανάμεσα στα δεδομένα.

4.3. Κοινωνικά Δίκτυα

4.3.1. Εισαγωγή

Τα κοινωνικά δίκτυα είναι Διαδικτυακές υπηρεσίες που στην ουσία αποτελούν «προσωπικούς χώρους» για επικοινωνία και διαμοιρασμό περιεχομένου. Χαρακτηρίζονται από ευκολία στη χρήση, από γρήγορη προσαρμογή στις αλλαγές της καθημερινής ζωής, διευκολύνουν τη δημιουργία αυθόρμητων σχέσεων, και διευρύνουν την αλληλεπίδραση και την επικοινωνία σύμφωνα με τους Ajjan & Hartshorne (2008).

Οι χρήστες αυτών των δικτύων, μπορούν να δημιουργούν ένα δημόσιο ή ημι-δημόσιο προφίλ σε ένα συγκεκριμένο σύστημα και να το καθιστούν ορατό στον κόσμο του Διαδικτύου. Στο προφίλ περιγράφουν προσωπικές πληροφορίες, ενδιαφέροντα, παρέχουν φωτογραφίες κλπ. Με απλά λόγια, επιτρέπουν σε κάποιον να δει την προσωπική τους ατζέντα και να αλληλεπιδράσει άμεσα με τα στοιχεία που περιέχει.

Μια άλλη βασική δυνατότητα των κοινωνικών δικτύων είναι ότι επιτρέπονται συνδέσεις οι οποίες δίνουν στους χρήστες τη δυνατότητα να συνάπτουν «Διαδικτυακή σχέση». Έτσι δημιουργούνται από ένα χρήστη, λίστες άλλων χρηστών με τους οποίους μοιράζεται μια σύνδεση (λίστα φίλων- friendslist). Ανάλογα με τον τύπο του δικτύου, εφαρμόζονται διαφορετικές πολιτικές σχετικά με τη σύναψη σχέσεων μεταξύ των χρηστών. Δηλαδή, δύο χρήστες μπορούν είτε να συνάψουν ένα σύνδεσμο διπλής κατεύθυνσης «Friends» (απαιτείται συναίνεση και από τους δυο) είτε ένα σύνδεσμο μονής κατεύθυνσης «Follower», «Fan» (ο ένας μπορεί να ακολουθήσει τον άλλο, χωρίς να είναι απαραίτητο να συμβεί και το αντίστροφο).

Επιπροσθέτως, θα ήταν σημαντικό να τονιστεί ότι οι χρήστες μπορούν να έχουν πρόσβαση στις λίστες των φίλων, να βλέπουν και να πλοηγούνται στη λίστα των συνδέσμων τους και των δραστηριοτήτων που δημοσιεύουν και ακόμη να αφήνουν δημόσια μηνύματα στο προφίλ τους. Ταυτόχρονα παρέχεται και μηχανισμός ιδιωτικής επικοινωνίας συνήθως με μορφή μηνυμάτων, παραπλήσια αυτής του ηλεκτρονικού ταχυδρομείου.

Παράλληλα με τα αυτά που αναφέρθηκαν παραπάνω, παρέχονται στους χρήστες κάποιες άλλες χρήσιμες υπηρεσίες όπως ο διαμοιρασμός περιεχομένου (φωτογραφιών, video, ανακοινώσεων), η δημιουργία συζητήσεων, η δημιουργία ομάδων, η παροχή ιστολογίων, η χρήση σύγχρονης επικοινωνίας, η πρόσβαση μέσω κινητού κ.α.

Τα περισσότερα δίκτυα υποστηρίζουν τη δημιουργία και επέκταση προϋπαρχόντων δικτύων φιλίας. Υπάρχουν δίκτυα που εστιάζονται στο να βοηθήσουν τα άτομα να βρουν

αγνώστους με τους οποίους μοιράζονται ενδιαφέροντα, πολιτικές πεποιθήσεις ή χόμπι. Κάποια δίκτυα προσπαθούν να διαφοροποιηθούν βάσει γλώσσας, θρησκείας, εθνότητας, ενώ άλλα δίκτυα διαφοροποιούνται σε σχέση με τις υπηρεσίες που παρέχουν, π.χ. διαμοιρασμός φωτογραφιών, βίντεο κτλ.

Εικόνα 16: Ιστορική Αναδρομή Κοινοτικών Δικτύων

4.3.2. Ιστορική Αναδρομή

Το Facebook, αν και αυτή τη στιγμή είναι μία από τις πιο δημοφιλείς σελίδες κοινωνικής δικτύωσης που υπάρχουν, δεν είναι και το πρώτο που εμφανίστηκε.

Η ιστορία των κοινωνικών δικτύων ξεκινάει από τα μέσα της δεκαετίας του ’90 (βλέπε Εικόνα 16), όπου τα πρώτα κοινωνικά δίκτυα κάνουν την εμφάνισή τους. Ξεκίνησαν με τη μορφή γενικών κοινοτήτων και μερικά παραδείγματα αυτών είναι το “The WELL” (1985), το “Theglobe.com” (1994), GeoCities (1994) και το “Tripod.com” (1995). Στην ουσία αυτό που προσπάθησαν να κάνουν οι κοινότητες αυτές ήταν να φέρουν κοντά τους χρήστες, να μοιραστούν προσωπικές πληροφορίες και ιδέες μέσω εργαλείων και προσωπικών δημοσιεύσεων. Ουσιαστικά αποτελούσαν προγόνους των ιστολογίων.

Από το 1997 έως το 2001, ένας αριθμός από εργαλεία δημιουργίας κοινοτήτων με συνδυασμό δημιουργίας προφίλ και δημόσιας λίστας φίλων επέτρεπαν στους χρήστες να δημιουργούν προσωπικά, επαγγελματικά και αισθηματικά προφίλ και να δημιουργούν φίλους

χωρίς να απαιτείται η έγκριση της σύνδεσης. Παραδείγματα αυτών είναι το "AsianAvenue.com", το "blackplanet.com " και το "MiGente.com".

Η νέα γενιά κοινωνικών δικτύων εμφανίστηκε το 2001 με το Ryze.com που είχε σκοπό να βοηθήσει τα άτομα να αξιοποιούν τα επιχειρηματικά τους δίκτυα το οποίο ποτέ δεν απέκτησε μεγάλη δημοσιότητα, ενώ από το 2003, αναπτύχθηκαν πολλές νέες υπηρεσίες κοινωνικής δικτύωσης και εμφανίστηκε ο όρος YASNS: «Yet Another Social Networking Service». Χαρακτηριστικά μπορούμε να αναφέρουμε τα δίκτυα LinkedIn, Visible Path, and Xing τα οποία αποτάθηκαν στον επιχειρηματικό κόσμο, ενώ κάποια άλλα όπως τα: Dogster (φιλίες μεταξύ ατόμων βάσει ενδιαφέροντος για τους σκύλους), Care2 (συναντήσεις ακτιβιστών), Couchsurfing (συνδέσεις ταξιδιωτών), MyChurch (σύνδεση χριστιανικών εκκλησιών και των μελών τους) αποτέλεσαν προσπάθειες για δημιουργία κοινοτήτων κοινών ενδιαφερόντων.

Καθώς όλο και περισσότεροι άνθρωποι αποκτούσαν πρόσβαση και ταχύτερες συνδέσεις στο Διαδίκτυο με ταυτόχρονη μείωση του κόστους, κοινωνικά δίκτυα όπως το MySpace, ή το HiFive, άρχισαν να προσελκύουν το παγκόσμιο ενδιαφέρον και να γίνονται δημοφιλή και ευρέως γνωστά και αναγνωρίσιμα. Το γεγονός αυτό αύξησε σημαντικά τον όγκο του περιεχομένου που δημιουργούνταν από τους χρήστες και γινόταν διαθέσιμο στο Διαδίκτυο και σαν αποτέλεσμα οι ιστοσελίδες που παρείχαν πλατφόρμες για τη δημοσίευση φωτογραφιών(Flickr.com), video (YouTube.com) ή μουσικής (Last.FM), καθώς και υπηρεσίες instant messaging, συζητήσεων, ιστολογίων, άρχισαν και αυτές να αποκτούν χαρακτηριστικά SNS.

4.3.3. Facebook

Το Facebook ξεκίνησε στις αρχές του 2004 ως ένα κοινωνικό δίκτυο μόνο για τους φοιτητές του Harvard. Ο χρήστης για να εισαχθεί έπρεπε να είχε email της μορφής harvard.edu. Καθώς το Facebook άρχισε να υποστηρίζει άλλες σχολές, απαιτούσε αντίστοιχες διευθύνσεις email. Το χαρακτηριστικό αυτό κράτησε αρχικά τον ιστοχώρο σχετικά κλειστό και δημιούργησε την εντύπωση μιας κλειστής, φιλικής και προνομιακής κοινότητας. Από το Σεπτέμβριο του 2005, άρχισε να ανοίγει στο κοινό. Αυτό όμως δε σήμαινε ότι ένας νέος χρήστης μπορούσε εύκολα να αποκτήσει πρόσβαση σε υπάρχοντα δίκτυα χρηστών. Σε αντίθεση με τα άλλα δίκτυα, το προφίλ των χρηστών του Facebook δεν είναι δημόσιο και κάποιος μπορεί να δει περισσότερα στοιχεία μόνο εφόσον εγκριθεί η σύνδεση και από τους δυο.

Είναι ο δημοφιλέστερος Ιστοχώρος Κοινωνικής Δικτύωσης και ιδρύθηκε από τον Mark Zuckerberg. Η κύρια λειτουργία του σήμερα, προσανατολίζεται στην κοινωνική προσέγγιση χρηστών Διαδικτυακά, μέσω της δημιουργίας ενός προσωπικού προφίλ.

Το Facebook καταμετράει περισσότερα από 400.000.000 ενεργά μέλη παγκοσμίως, που κατά μέσο όρο διασυνδέονται με 130 φίλους και ξοδεύουν παραπάνω από 55 λεπτά ημερησίως. Εκτιμάται ότι το 1/3 του πληθυσμού των ΗΠΑ διατηρεί προφίλ στην υπηρεσία, ενώ το αντίστοιχο ποσοστό για την Ελλάδα ανέρχεται στο 22,9% του πληθυσμού (περί τα 2.515.220 μέλη), αριθμός που αυξάνεται με ρυθμούς γεωμετρικής προόδου. Με αφορμή τη δημοτικότητά του, το Facebook έχει υποστεί κριτική και κατηγορείται για θέματα που αφορούν στα προσωπικά δεδομένα των χρηστών και τις πολιτικές απόψεις των ιδρυτών του. Ωστόσο, η συγκεκριμένη ιστοσελίδα παραμένει η πιο διάσημη εφαρμογή κοινωνικής δικτύωσης.

4.3.4. YouTube

Το YouTube είναι ένας δημοφιλής διαδικτυακός τόπος, που επιτρέπει αποθήκευση, αναζήτηση κι αναπαραγωγή ψηφιακών ταινιών. Ιδρύθηκε το 2005, ενώ το 2006 η εταιρεία εξαγοράστηκε από την Google έναντι του αστρονομικού ποσού των 1,65 δις δολαρίων. Η υπηρεσία παρέχει τη δυνατότητα σε όλους του επισκέπτες να προβάλλουν τα αποθηκευμένα βίντεο, ενώ τα εγγεγραμμένα μέλη μπορούν να αποθηκεύουν απεριόριστο αριθμό ταινιών με

χρονικό όριο δέκα λεπτών το καθένα. Για κάθε εγγραφή βίντεο παρέχεται ο αριθμός των επισκεπτών που το έχουν προβάλλει, καθώς και σχόλια χρηστών προκειμένου να μπορεί να αξιολογηθεί. Ενδεικτικά, κάθε λεπτό της ώρας υπολογίζονται 24 νέες ώρες βίντεο που προστίθενται από χρήστες, αριθμός ασύλληπτος αφού για να τα παρακολουθήσει κανείς χρειάζεται 4 συνεχόμενα χρόνια.

4.3.5. Twitter

Το Twitter αντιπροσωπεύει μια νέα μορφή διαδραστικότητας, όπου ολόκληρη η επικοινωνία διεξάγεται με μόλις 140 χαρακτήρες, βασιζόμενο στη λογική του λεγόμενου microblogging. Όπως όλα τα sites κοινωνικής δικτύωσης, έτσι και το Twitter, στα πρώτα βήματά του, «αφοσιώθηκε» κυρίως στην προσέλκυση των περισσότερων δυνατών χρηστών, παρά στην παραγωγή και εξασφάλιση κερδών, γεγονός που έκανε τους αναλυτές να αναρωτιούνται, ως προς το πώς θα μπορούσε να «μεταμορφωθεί» σε προσοδοφόρα επιχείρηση. Η απάντηση ήρθε με τη λέξη "Twitter" να χαρακτηρίζεται ως η κορυφαία του 2009 (σύμφωνα με έρευνα του Global Language Monitor), ενώ εταιρείες που επιθυμούν να τη χρησιμοποιούν ως διαφημιστικό μέσο οφείλουν να πληρώνουν χρηματικό αντίτιμο. Από συμφωνίες που υπογράφηκαν με την Microsoft και τη Google, οι ιδρυτές της υπηρεσίας εξασφάλισαν κέρδη της τάξης των 25 εκατομμυρίων δολαρίων, ενώ από το 2006 οπότε και ιδρύθηκε υπολογίζονται περίπου έσοδα που ανέρχονται στα 155 εκατομμύρια δολάρια.

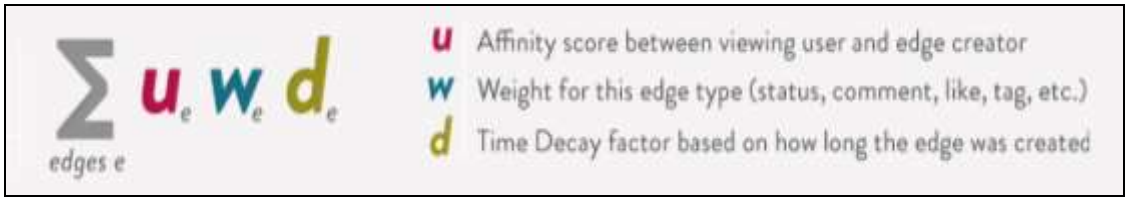
4.4. Αλγόριθμοι Αναζήτησης Κοινωνικών Δικτύων

4.4.1. Διαφοροποίηση αναζήτησης στο Διαδίκτυο και σε κοινωνικά δίκτυα

Η αναζήτηση σε κοινοτικά δίκτυα διαφέρει κατα πολύ από την αναζήτηση σε συμβατικές μηχανές. Η διαφορά έγκειται στο γεγονός ότι η συμβατική μηχανή αναζήτησης αναλύει και μετατρέπει τα δεδομένα που ανακτά σε λέξεις. Έπειτα υπολογίζει τη σχετικότητα των λέξεων στο ευρετήριο (βλέπε ευρετηριοποίηση) βάση διανυσμάτων. Αντίθετα, οι μηχανές αναζήτησης των κοινοτικών δικτύων, αναζητούν πολύπλοκα αντικείμενα τα οποία είναι δύσκολο ή συχνά αδύνατο να ταξινομηθούν για λόγους πολυπλοκότητας και όγκου. Επομένως στις περισσότερες μηχανές αναζήτησης των κοινοτικών δικτύων, τα πεδία που ελέγχονται είναι περιορισμένα και δεν ανταποκρίνονται στο ευρύτερο φάσμα δεδομένων που περιέχονται στο αντικείμενο που τα αποτελούν.

4.4.2. EdgeRank

Ο μέσος χρήστης του Facebook ξοδεύει περισσότερο από το ένα τέταρτο του χρόνου που περνάει στην ιστοσελίδα, κοιτάζοντας απλά το News Feed. Για τους χρήστες αυτό μπορεί να σημαίνει ατελείωτες φωτογραφίες από μωρά, ταξίδια ή κατοικίδια και αμέτρητες δημοσιεύσεις από ατάκες και τραγούδια. Για τις εταιρίες όμως αυτός ο χρόνος που ξοδεύουμε αποτελεί μία τεράστια εμπορική ευκαιρία. Όπως έχει παρατηρηθεί ο μέσος χρήστης δε θα επισκεφθεί ποτέ μία σελίδα εταιρίας στο Facebook. Επομένως ποιό είναι το καλύτερο μέρος για να προσεγγίσει η εκάστοτε εταιρία τους χρήστες; Φυσικά δεν είναι άλλο από το News Feed μας. Οι εταιρίες φαίνεται πως τελικά βρήκαν τον πιο αποτελεσματικό τρόπο για κοινωνικό μάρκετινγκ, αλλά έχουν ακόμα πολλά να μάθουν για τον τρόπο που αυτό λειτουργεί. Και κάπου εδώ έρχεται το EdgeRank. Είτε είστε εταιρία είτε ένα απλός χρήστης, είναι χρήσιμο να κατανοήσουμε τι εμφανίζεται στο News Feed μας και γιατί. Ο συγκεκριμένος αλγόριθμος κάνει ακριβώς αυτό. Αναγνωρίζει τι είδους δημοσιεύσεις εμφανίζονται στο News Feed του κάθε χρήστη ξεχωριστά, λαμβάνοντας υπόψη τρεις μεταβλητές, τις Affinity, Weight και Time Delay (βλέπε Εικόνα 17).



Εικόνα 17: Μαθηματικός Τύπος EdgeRank

Η μεταβλητή Affinity προσπαθεί να περιγράψει τη σχέση μεταξύ του χρήστη και του δημιουργού της δημοσίευσης, με σκοπό να δώσει προτεραιότητα στις δημοσιεύσεις από άτομα που έχουμε μεγαλύτερη σχέση. Το Weight μετράει το μέγεθος των δημοσιεύσεων, δίνοντας προτεραιότητα σε αυτές με το μεγαλύτερο σκορ. Πρώτα θα εμφανιστούν οι φωτογραφίες και τα βίντεο, στη συνέχεια οι σύνδεσμοι και τέλος τα γραπτά κείμενα. Τέλος το Time Delay αναφέρεται στην "ηλικία" της δημοσίευσης, όσο πιο παλιά είναι μία δημοσίευση τόσο μεγαλώνει η πιθανότητα να μην κάνει την εμφάνισή της στο News Feed σας, έτσι διασφαλίζεται το γεγονός πως θα έχετε ένα πάντα ενημερωμένο με τα πιο πρόσφατα νέα New Feed.

4.4.3. Ανοιχτά ζητήματα ανάκτησης δεδομένων σε κοινωνικά δίκτυα

Συχνά οι αλγόριθμοι ταξινόμησης και αναζήτησης τόσο σε συμβατικές μηχανές αναζήτησης, όσο και σε κοινοτικά δίκτυα αποτελούν επιχειρησιακά μυστικά. Επομένως δεν παρέχεται από την εταιρεία καμία λεπτομέρεια σχετικά με την αρχιτεκτονική ή τους αλγορίθμους που χρησιμοποιούνται. Στην περίπτωση του Youtube ο αλγόριθμος αναζήτησης αποτελεί εταιρικό μυστικό και κατά τη διάρκεια συγγραφής αυτής της πτυχιακής εργασίας δεν παρέχεται καμία επίσημη λεπτομέρεια σχετικά με τους μηχανισμούς αναζήτησης.

Από ανεπίσημες πηγές, όπως SEO Experts και άλλες ιστοσελίδες, παρέχονται πληροφορίες οι οποίες είναι συχνά αντικρουόμενες και βασίζονται περισσότερο στη διαίσθηση του κάθε παρατηρητή παρά σε έγκυρη επιστημονική παρατήρηση. Οι πληροφορίες οι οποίες είναι κοινές σε όλους τους ιστοτόπους, περιγράφουν πως ο αλγόριθμος εξετάζει τον τίτλο του video, την περιγραφή, την προτίμηση του χρήστη σε συγκεκριμένα κανάλια και στο ιστορικό αναζήτησης.

5. ΑΝΑΛΥΣΗ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΣΧΕΔΙΟ ΔΡΑΣΗΣ

5.1. Δήλωση προβλήματος

Είναι γνωστό πως δεν υπάρχει μηχανισμός που να επιτρέπει διασυνοριακά ερωτήματα σε διάφορα κοινοτικά δίκτυα. Δηλαδή δεν παρέχεται η δυνατότητα σε ένα χρήστη να πραγματοποιεί ερωτήματα που να περιέχουν πληροφορία από περισσότερό του ενός κοινοτικά δίκτυα. Σε αυτή την εργασία λύσαμε αυτό το πρόβλημα με τη χρήση μιας βάσης δεδομένων η οποία περιέχει πληροφορίες από πολλά κοινοτικά δίκτυα. Η πληροφορία αυτή λαμβάνεται και αποθηκεύεται μέσω ειδικών μηχανισμών οι οποίοι παρουσιάζονται παρακάτω. Επίσης, παρέχεται και μία διεπαφή υποβολής ερωτημάτων.

5.2. Σχεδιασμός

5.2.1. Αρχιτεκτονική

Η εφαρμογή βασίζεται σε μια πολυεπίπεδη αρχιτεκτονική. Το κατώτατο επίπεδο της αρχιτεκτονικής αυτής, αποτελεί το επίπεδο δεδομένων (Data Layer). Στο επίπεδο δεδομένων βρίσκεται η βάση δεδομένων της εφαρμογής η οποία περιγράφεται στην επόμενη ενότητα. Στο ακριβώς επόμενο επίπεδο της εφαρμογής βρίσκεται το επίπεδο προσπέλασης δεδομένων (Data Access Layer). Το επίπεδο αυτό είναι ο ενδιάμεσος ανάμεσα στο επίπεδο της βάσης δεδομένων και στη λογική της εφαρμογή. Το επόμενο επίπεδο είναι το λογικό επίπεδο της εφαρμογής (Logic Layer). Στο επίπεδο αυτό βρίσκονται οι λογικές μονάδες οι οποίες είναι υπεύθυνες για την αποκωδικοποίηση των ερωτημάτων και οι υπηρεσίες της εφαρμογής. Τέλος, στην κορυφή του συστήματος, βρίσκεται το επίπεδο παρουσίασης (Presentation Layer), το οποίο περιλαμβάνει τη διεπαφή τοης εφαρμογής με το χρήστη.

5.2.2. Σχεδίαση Βάσης Δεδομένων

Η λογική του σχεδιασμού της βάσης δεδομένων, βρίσκεται στις αρχές του αντικειμενοστραφή προγραμματισμού. Αρχικά εντοπίστηκαν και δημιουργήθηκαν έννοιες οι οποίες μπορούν να αντιπροσωπεύσουν συστατικά του κοινωνικού δικτύου. Παράλληλα οι έννοιες αυτές θα πρέπει να είναι αρκετά αφηρημένες, έτσι ώστε να μπορούν να ανταποκριθούν τόσο σε μελλοντικές προσθήκες στο κοινωνικό δίκτυο, όσο και σε συστατικά κοινωνικών δικτύων που δεν μελετήθηκαν στα πλαίσια αυτής της πτυχιακής εργασίας. Οι έννοιες αυτές παρουσιάζονται παρακάτω:

- **Ισχυρές Οντότητες (Strong Entities):** Αποτελούν τις κυρίαρχες οντότητες του συστήματος, οι οποίες αντιπροσωπεύουν το χρήστη μέσα στο κοινωνικό δίκτυο, και έχουν ένα πλήρες σύνολο δυνατοτήτων.
- **Ασθενείς Οντότητες (Weak Entities):** Αποτελούν τις δευτερεύουσες οντότητες του συστήματος. Οι οντότητες αυτές αντιπροσωπεύουν κεντρικές έννοιες του κοινωνικού δικτύου ή αφηρημένες έννοιες του πραγματικού κόσμου.
- **Συστατικά (Components):** Τα συστατικά είναι αντικείμενα που αποτελούν συστατικό τμήμα των οντοτήτων. Ουσιαστικά τα συστατικά αποτελούν τον κορμό των οντοτήτων.

- **Περιέκτες (Containers):** Οι περιέκτες αποτελούν συναθροίσεις αντικειμένων του κοινοτικού δικτύου. Τα αντικείμενα που μπορούν να τοποθετηθούν μέσα στους περιέκτες παρουσιάζονται παρακάτω.
- **Πολυμέσα (Multimedia):** Τα πολυμέσα αποτελούν πολυμεσικά αντικείμενα με τα οποία αλληλεπιδρούν και επικοινωνούν οι οντότητες.
- **Έγγραφα (Documents):** Τα έγγραφα αποτελούν αντικείμενα τα οποία περιέχουν συνήθως μεγάλο όγκο γραπτού λόγου, με τα οποία αλληλεπιδρούν και επικοινωνούν οι οντότητες.
- **Κοινωνικές Αλληλεπιδράσεις (Social Interactions):**Οι κοινωνικές αλληλεπιδράσεις αποτελούν το κύριο μέσο επικοινωνίας ανάμεσα στις οντότητες και κάθε κοινωνικό δίκτυο χρησιμοποιεί αλληλεπιδράσεις οι οποίες είναι κοινές σε όλα τα δίκτυα, αλλά και μοναδικές για το κάθε ένα.

Οι έννοιες αυτές αποτελούν τα «υπερ-αντικείμενα» ή «υπερ-κλάσεις», οι οποίες είναι και ο κορμός της βάσης δεδομένων. Κάθε συστατικό στοιχείο του κοινωνικού δικτύου αποτυπώνεται στη βάση σαν ένας πίνακας ο οποίος ¨επεκτείνει¨ κάποιο «υπερ-αντικείμενο». Τα «υπερ-αντικείμενα» αναπαριστώνται στη βάση δεδομένων από τους παρακάτω πίνακες:

- Strong Entity (SE_ID, Entity Type)
- Weak Entity (WE_ID, Entity Type)
- Component (Component ID, Component Type)
- Container (Container ID, Container Type)
- Multimedia (Multimedia ID, Multimedia Type)
- Document (Document ID, Document Type)
- Social Interaction (Social ID, Social Type)
- Community (Community ID, Community Type)
- Classifier (Classifier ID, Classifier Name, Classifier Popularity)

ΣΥΣΧΕΤΊΣΕΙΣ ΑΝΆΜΕΣΑ ΣΤΑ ΑΝΤΙΚΕΊΜΕΝΑ

- Ένα StrongEntity περιέχει ένα Component
  - StrongEntity (SE ID, Entity Type, Component ID)
- Ένα WeakEntity περιέχει ένα Component
  - WeakEntity (WE ID, Entity Type, Component ID)
- Ένα Community περιέχει ένα Component
  - Community (Community ID, Community Type, Component ID)
- Ένα StrongEntity παράγει πολλά Multimedia Objects
  - Multimedia (Multimedia ID, Multimedia Type, SE ID)
- Ένα WeakEntity παράγει πολλά Multimedia Objects
  - Multimedia (Multimedia ID, Multimedia Type, SE ID, WE ID)
- Ένα StrongEntity παράγει πολλά Documents
  - Document (Document ID, Document Type, SE ID)
- Ένα WeakEntity παράγει πολλά Documents
  - Document (Document ID, Document Type, SE ID, WE ID)
- Ένα StrongEntity παράγει πολλά Social Interactions
  - Social Interactions (Social ID, Social Type, Creator ID)
- Ένα StrongEntity διευθύνει πολλά Communities
  - Community (Community ID, Community Type, Component ID, SE ID)
- Ένα WeakEntity παράγει πολλά Communities
  - Community (Community ID, Community Type, Component ID, SE ID, WE ID)

- Ένα Community παράγει πολλά Communities
  - Community (Community ID, Community Type, Component ID, SE ID, WE ID, Community ID)
- Πολλά StrongEntities σχετίζονται με πολλά άλλα StrongEntities
  - Relation (SE ID1, SE ID2)
- Πολλά StrongEntities σχετίζονται με πολλά WeakEntities
  - Fans (SE ID, WE ID)
- Ένα Social Interaction μπορεί να είναι εμφωλευμένο σε ένα άλλο Social Interaction
  - Social Interactions (Social ID, Social Type, Parent ID, Creator ID)
- Πολλά Social Interactions σχετίζονται με ένα Document
  - Document Interactions (Document ID, SE ID, WE ID, Social ID)
- Πολλά Social Interactions σχετίζονται με ένα Multimedia Object
  - Multimedia Interactions (Multimedia ID, SE ID, WE ID, Social ID)
- Πολλά StrongEntities είναι μέλη σε πολλά Communities
  - Community Members (Community ID, Community Type, Component ID, SE ID)
- Ένας Classifier μπορεί να είναι παιδί ενός άλλου Classifier
  - Classifier (Classifier ID, Classifier Name, Classifier Popularity, Parent ID)
- Πολλοί Classifiers μπορούν να ταξινομήσουν πολλά Documents
  - Document Classification(Document ID, SE ID, WE ID, Classifier ID, Classifier Parent, Popularity)
- Πολλοί Classifiers μπορούν να ταξινομήσουν πολλά Multimedia Objects
  - Multimedia Classification (Multimedia ID, SE ID, WE ID, Classifier ID, Classifier Parent, Popularity)
- Πολλοί Classifiers μπορούν να ταξινομήσουν πολλά WeakEntities
  - Document Classification(WE ID, Classifier ID, Classifier Parent, Popularity)

Συγκεντρωτικά, η βάση δεδομένων συνοψίζεται στον Πίνακας 3.

ΑΝΤΙΚΕΙΜΕΝΟ	ΠΕΔΙΑ
Strong Entity	(SE ID, Entity Type, Component ID)
Weak Entity	(WE ID, Entity Type, Component ID)
Community	(Community ID, Community Type, Component ID, SE ID, WE ID, Community ID)
Multimedia	(Multimedia ID, Multimedia Type, SE ID, WE ID)
Document	(Document ID, Document Type, SE ID, WE ID)
Social Interactions	(Social ID, Social Type, Parent ID, Creator ID)
Container	(Container ID, SE ID, WE ID, Document ID, Multimedia ID, Container Type, Component ID)
Component	(Component ID, Component Type, Owner ID)
Relation	(SE ID1, SE ID2, Type of Relation)
Fan	(SE ID, WE ID)
Container Social	(Container ID, Social ID, Parent ID)
Classifier	(Classifier ID, Classifier Name, Classifier Popularity, Parent ID)
Weak Entity Classification	(WE ID, Classifier ID, Classifier Parent)

Multimedia Classification	(Multimedia ID, Classifier ID, Classifier Parent, Creator ID, Creator Type)
Document Classification	(Document ID, Classifier ID, Classifier Parent, Creator ID, Creator Type)
Community Members	(Community ID, Community Type, Component ID, SE ID)

Πίνακας 3: Συνοπτική αποτύπωση της βάσης δεδομένων

ΑΝΤΙΚΕΊΜΕΝΑ ΤΟΥ FACEBOOK ΠΟΥ ΕΝΔΙΑΦΈΡΟΥΝ

Όσων αφορά το κοινοτικό δίκτυο Facebook το σύστημα αποθηκεύει δεδομένα που αφορούν τις δομές Comment, Event, Group, Link, Note, Page, Photo, Post, User, Video. Η σχεσιακή δομή που καταγράφει τα δεδομένα αυτά συνοψίζεται παρακάτω:

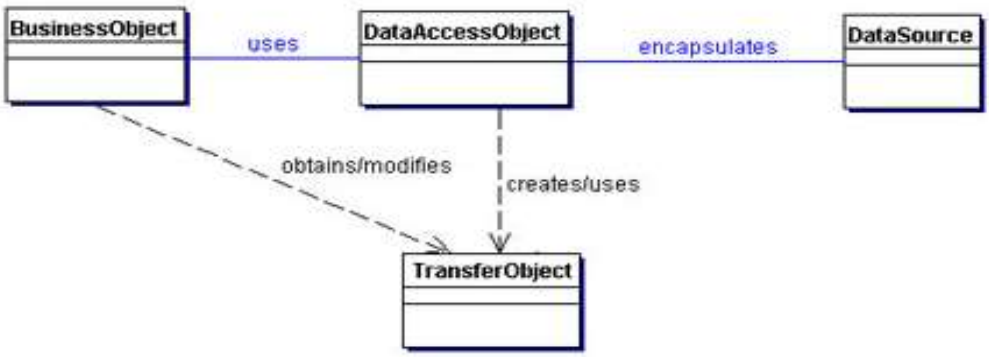
- Comments (Comment ID, Creator ID, Message, Created Time, Like Count, Parent ID)
- Event (Event ID, User ID, Page ID, Group ID, Wall ID, name, description, start time, end time, location)
- Group (Group ID, User ID, Wall ID, name, description, link)
- Link (Link ID, User ID, name, description, message)
- Note(Note ID, User ID, subject, message, created time, updated time)
- Page(Page ID, name, link, categories, likes, picture, talking about, Wall ID)
- Photo(Photo ID, User ID, Page ID, picture, source, link, created time)
- Post(Post ID, User ID, Page ID, type, created time, updated time)
- User(User ID, name. gender, link, username, location, picture, Wall ID,)
- Video(Video ID, User ID, Page ID, name, description, picture, embed html, source, created time, updated time)

5.2.3. Αρχιτεκτονική Επιπέδου Προσπέλασης Δεδομένων

Το επίπεδο προσπέλασης δεδομένων, χρησιμοποιεί αντικείμενα πρόσβασης δεδομένων (Data Transfer Objects – DAOs) τα οποία ενθυλακώνουν την πρόσβαση στη βάση δεδομένων. Ένα DAO υλοποιεί το μηχανισμό που απαιτείται για τη χρήση της βάσης δεδομένων. Η εφαρμογή χρησιμοποιεί την απλούστερη διεπαφή που παρέχεται από το DAO. Έτσι το αντικείμενο αποκρύπτει εντελώς την αρχιτεκτονική της βάσης από την εφαρμογή. Επειδή η διεπαφή που παρέχεται από το DAO, δεν αλλάζει, το πρότυπο αυτό επιτρέπει αλλαγές στη βάση δεδομένων, χωρίς να επηρεάσει την εφαρμογή.

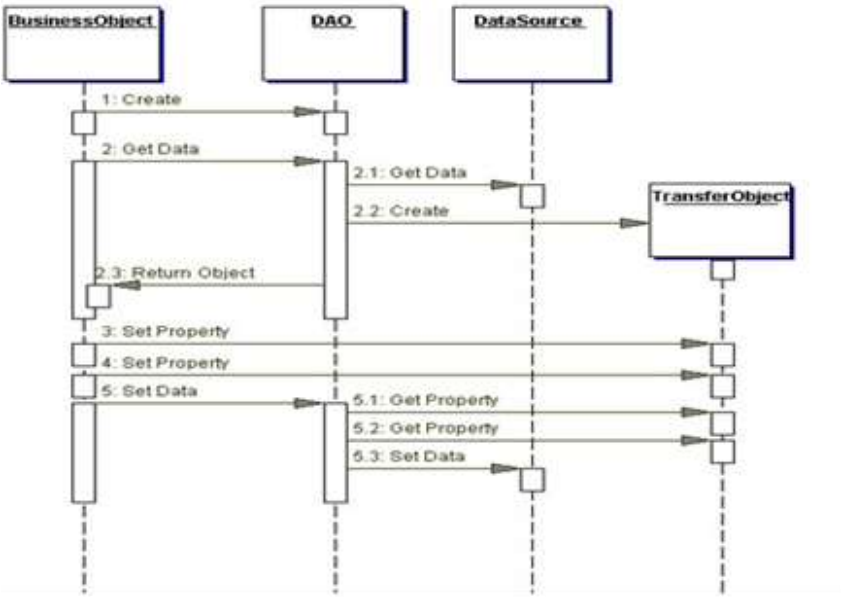
ΔΟΜΗ ΤΟΥ ΠΡΟΤΥΠΟΥ

Στην Εικόνα 18 παρουσιάζεται το διάγραμμα που αναπαριστά τις συσχετίσεις στο πρότυπο ΑΠΔ.



Εικόνα 18: Πρότυπο Αντικειμένου Πρόσβασης Δεδομένων(ΑΠΔ)

Η Εικόνα 19 παρουσιάζει το διάγραμμα ακολουθίας που αναπαριστά τις αλληλεπιδράσεις ανάμεσα στους συμμετέχοντες του προτύπου.



Εικόνα 19: Διάγραμμα Ακολουθίας του προτύπου ΑΠΔ

Αντικείμενο Εφαρμογής – Business Object

Το αντικείμενο εφαρμογής αναπαριστά το αντικείμενο το οποίο απαιτεί πρόσβαση στην πηγή δεδομένων για να ανακτήσει και να αποθηκεύσει δεδομένα. Το αντικείμενο αυτό μπορεί να είναι οποιοδήποτε αντικείμενο της εφαρμογής.

Αντικείμενο Πρόσβασης Δεδομένων – Data Transfer Object

Το Αντικείμενο Πρόσβασης Δεδομένων – Data Transfer Object DAO (Εικόνα 20) αποτελεί το κεντρικό αντικείμενο του συγκεκριμένου προτύπου. Το αντικείμενο αυτό, αποκρύπτει την δομή της πηγής δεδομένων στην οποία συνδέεται έτσι ώστε να παρέχει διαφανή πρόσβαση στην πηγή δεδομένων. Το αντικείμενο εφαρμογής χρησιμοποιεί τις μεθόδους του αντικειμένου για εγγραφή και ανάκτηση δεδομένων στη βάση.

```
import istl.sociomine.storage.DAO.interfaces.DAO;
import istl.sociomine.storage.transfer_objects.superobjects.ComponentTransferObject;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import java.sql.SQLException;

public class ComponentDAO implements DAO {


    private static Connection getConnection() {
//Creates a Connection with the database in order to perform operations

    public boolean create(ComponentTransferObject object) {
//Creates a new entry in Component Table

    public ComponentTransferObject read(ComponentTransferObject object) {
//Reads a tuple from the Component Table

    public boolean update(ComponentTransferObject object) {
//Alters the data in a tuple of the Component Table

    public boolean delete(ComponentTransferObject object) {
//Deletes a tuple of the Component Table
    }
```

Εικόνα 20: Δείγμα Κώδικα αντικειμένου DAO

Πηγή Δεδομένων – Data Source

Η βάση δεδομένων της εφαρμογής, όπως περιγράφθηκε στην προηγούμενη ενότητα.

Αντικείμενο Μεταφοράς – Transfer Object

Αναπαριστά ένα αντικείμενο μεταφοράς δεδομένων (Εικόνα 21). Το DAO μπορεί να χρησιμοποιεί ένα αντικείμενο μεταφοράς για να επιστρέφει δεδομένα στο αντικείμενο-πελάτη. Επίσης μπορεί να χρησιμοποιεί το αντικείμενο αυτό, για να λαμβάνει δεδομένα από το αντικείμενο-πελάτη, ώστε να πραγματοποιεί ενέργειες στην πηγή δεδομένων.

```
package istl.sociomine.storage.transfer_objects.superobjects;

import istl.sociomine.storage.transfer_objects.TransferObject;

public class ComponentTransferObject implements TransferObject{

    private String componentID;
    private String componentType;
    private String OwnerType;

    public String getComponentID() {

    public void setComponentID(String componentID) {

    public String getComponentType() {

    public void setComponentType(String componentType) {

    public String getOwnerType() {

    public void setOwnerType(String OwnerType) {
}
```

Εικόνα 21: Δείγμα Κώδικα του Αντικειμένου Μεταφοράς

ΛΗΨΗ ΔΕΔΟΜΕΝΩΝ

Η λήψη δεδομένων γίνεται μέσω ενός μηχανισμού, ο οποίος αποτελείται από πολλές κλάσεις οι οποίες αναλαμβάνουν τη λήψη δεδομένων για κάθε διαφορετικό αντικείμενο του συστήματος. Στην Εικόνα 22 παρουσιάζεται το αντικείμενο ανάκτησης πληροφοριών σελίδας και στην Εικόνα 23 παρουσιάζονται τα αποτελέσματα.

```
package istl.sociomine.crawler.fb.retriever.leafs;

import com.restfb.DefaultFacebookClient;
import com.restfb.Facebook;
import com.restfb.FacebookClient;
import com.restfb.types.Page;
import istl.sociomine.crawler.fb.retriever.beans.PageBean;
import java.util.List;

public class PageRetriever implements Runnable {


    public PageRetriever(String accessToken, String pageID) {
    //Default constructor for page Retriever
    @Override
    public void run() {

    private void fetch() {
}
//Fetches information about the page with id: pageID

class PagePic {

//Support class for storing the page profile picture.
```

Εικόνα 22: Δείγμα κώδικα αντικειμένου ανάκτησης πληροφοριών σελίδας

```
run:
Retrieved Results
Page ID: 9991232322
Page Name: Supernatural
Page Link: https://www.facebook.com/Supernatural
Page Categories: Tv show
Page Likes Count: 11722117
Page Picture: https://fbcdn-profile-a.akamaihd.net/hprofile-ak-ash4/203549_9991232322_1722944795_q.jpg
Page Talking About Count: 153292
```

Εικόνα 23: Αποτελέσματα ανάκτησης πληροφοριών σελίδας

5.2.4. Αρχιτεκτονική Διεπαφών

5.2.4.1. Search

Η εφαρμογή της αναζήτησης χωρίζεται σε δύο μέρη / τύπους ανάλογα με τον στόχο της αναζήτησης και τα κριτήρια που απαιτούνται. Στην απλή αναζήτηση, ο χρήστης πληκτρολογεί τη λέξη ή τη φράση που επιθυμεί, ενώ στην προχωρημένη αναζήτηση, δίνεται στο χρήστη η δυνατότητα να διαμορφώσει ερωτήματα που περιλαμβάνουν περισσότερο πολύπλοκα κριτήρια. Και τις δύο περιπτώσεις τα αποτελέσματα παρουσιάζονται με μορφή λίστας.

ΑΠΛΗ ΑΝΑΖΗΤΗΣΗ

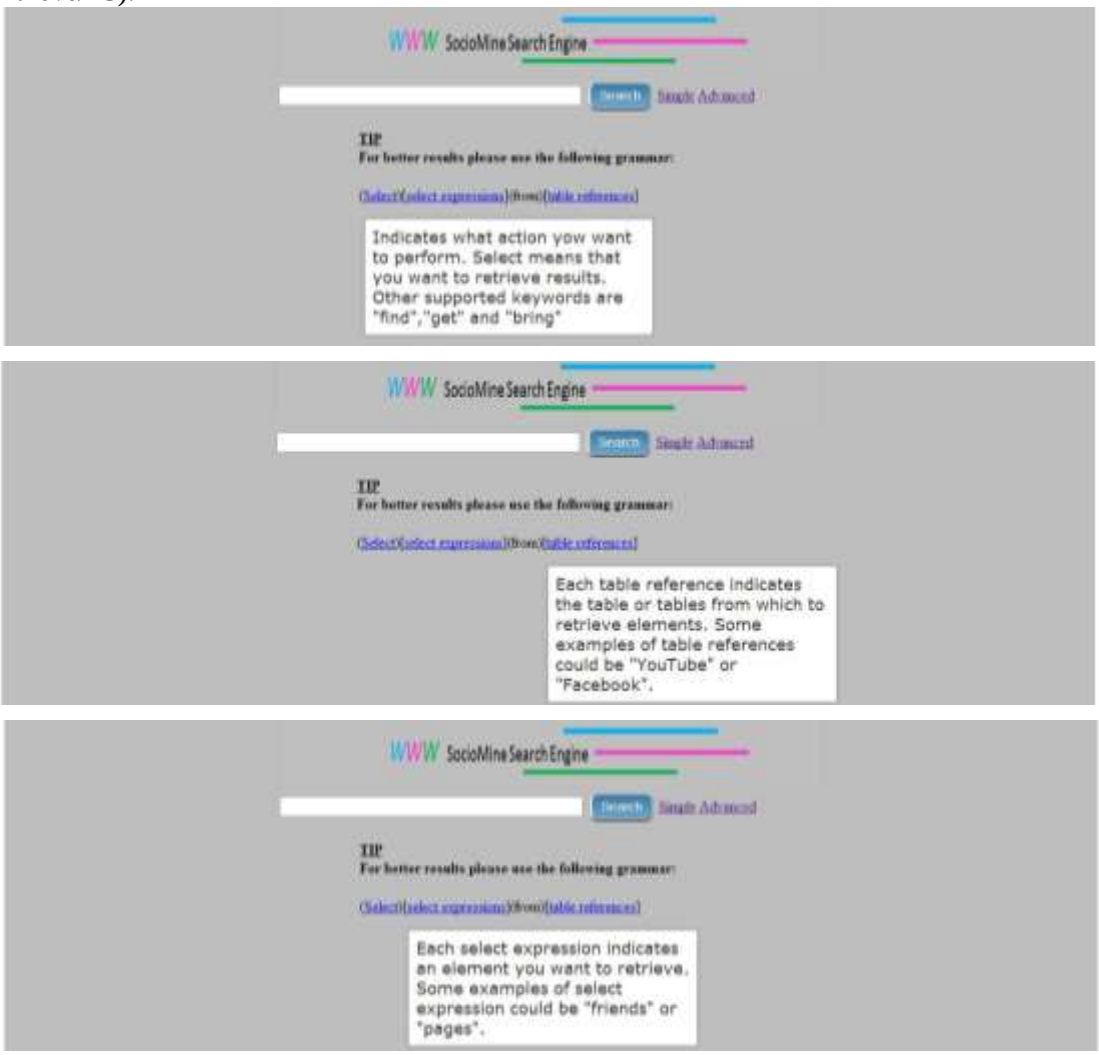
Η απλή αναζήτηση (Εικόνα 24) προσφέρει στο χρήστη τη δυνατότητα να λάβει αποτελέσματα σε απλά ερωτήματα όπως «Θέλω να δω όλους τους φίλους μου στο facebook». Ο χρήστης εισάγει το ερώτημα χρησιμοποιώντας μια απλή γραμματική της μορφής:

(Select) [select expressions] (from) [table references]

Εικόνα 24: Διεπαφή χρήστη για την απλή αναζήτηση

Στο σύστημα μας επιλέξαμε τη συγκεκριμένη γραμματική διατύπωσης ερωτημάτων, επειδή καθορίζει διακριτά τους τρεις νοηματικούς άξονες του ερωτήματος. Επίσης με τον τρόπο αυτό είναι ευκολότερη η κωδικοποίηση του ερωτήματος από το σύστημα. Οι τρεις νοηματικοί άξονες του ερωτήματος είναι ο άξονας της εντολής, ο άξονας των πεδίων και ο άξονας των αναφορών. Ειδικότερα, ο *άξονας της εντολής* καθορίζει την ενέργεια που θέλει να πραγματοποιήσει ο χρήστης. Στο παρών σύστημα το οποίο σχετίζεται με την αναζήτηση δεδομένων, η εντολή πάντα θα είναι εντολή αναζήτησης. Σε διαφορετικά συστήματα θα μπορούσαν να υποστηριχθούν και άλλου τύπου εντολές. Ο *άξονας των πεδίων* καθορίζει ποιά πεδία θέλει ο χρήστης να αναζητήσει. Τα πεδία μπορούν να είναι οποιαδήποτε αντικείμενα του συστήματος, όπως χρήστες, σχόλια ή σημειώσεις. Ο *άξονας των αναφορών* καθορίζει το πού αναφέρεται το ερώτημα. Ένα παράδειγμα αναφοράς θα μπορούσε να είναι οι φίλοι στο Facebook ή τα βίντεο στο YouTube. Ακόμη θα μπορούσε να είναι μια πιο αφηρημένη αναφορά όπως για παράδειγμα Facebook ή YouTube. Η παραπάνω λογική αποτυπώθηκε και

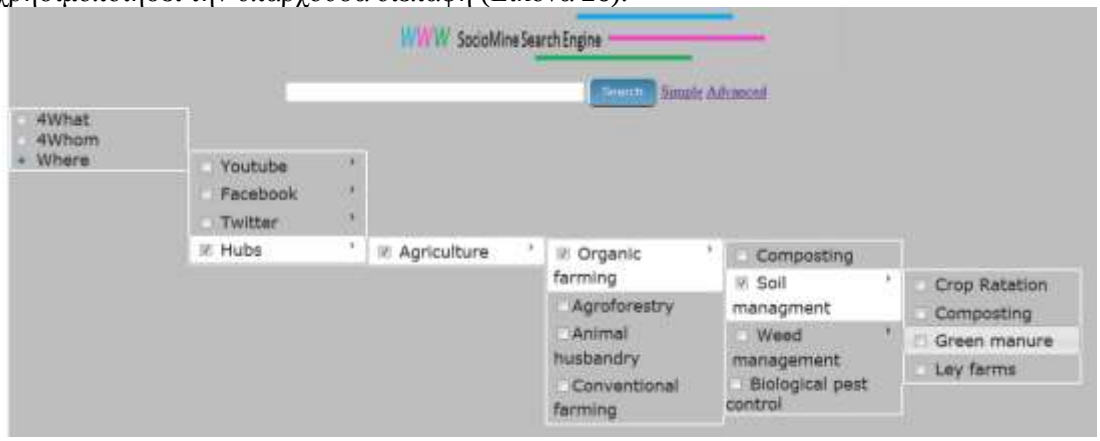
στην χρήση των tooltips που καθοδηγούν το χρήστη στη δημιουργία ενός ερωτήματος (βλέπε Εικόνα 25).



Εικόνα 25: Tooltips

ΣΥΝΘΕΤΗ ΑΝΑΖΗΤΗΣΗ

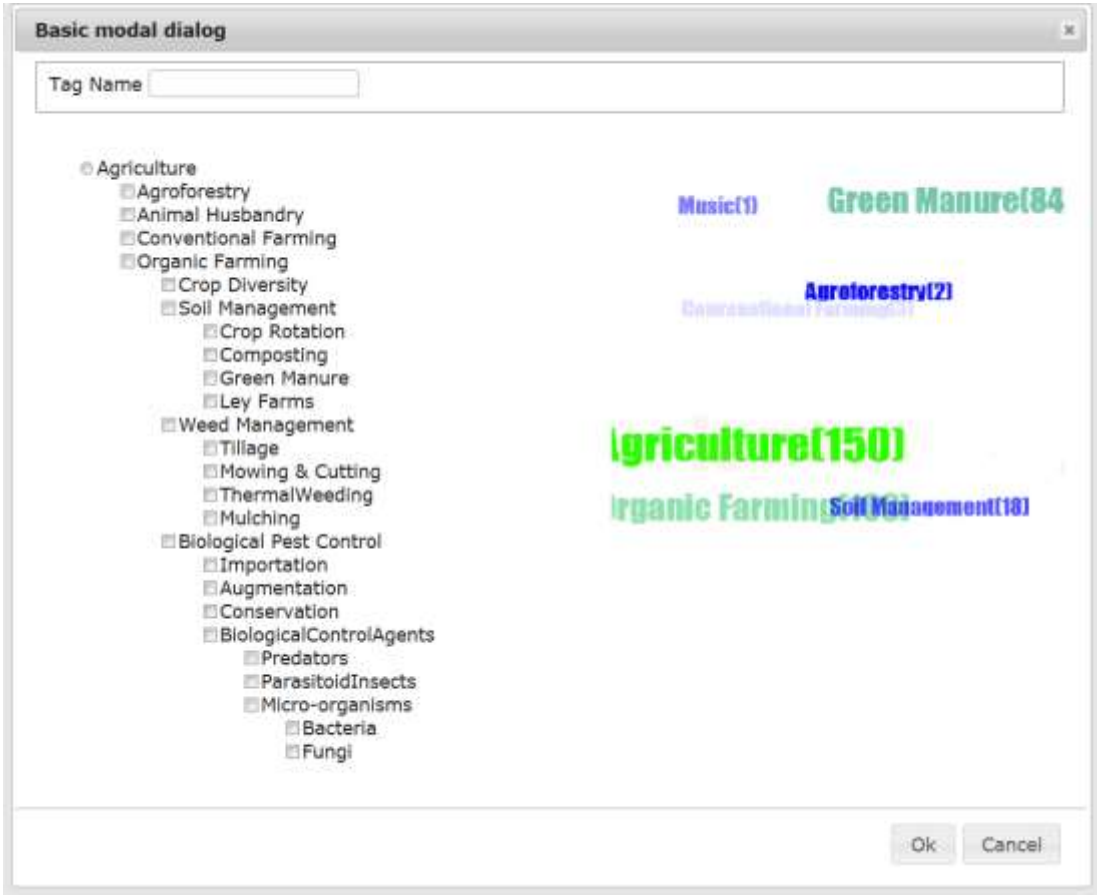
Η σύνθετη αναζήτηση επιτρέπει στο χρήστη να παράξει περίπλοκα ερωτήματα όπως «Θέλω να δω τα σχόλια των φίλων μου στο facebook οι οποίοι έχουν ανεβάσει το βίντεό μου από το YouTube στον τοίχο τους». Ο χρήστης μπορεί είτε να εισάγει το ερώτημα χειροκίνητα είτε να χρησιμοποιήσει την υπάρχουσα διεπαφή (Εικόνα 26).



Εικόνα 26: Διεπαφή χρήστη για τη Σύνθετη Αναζήτηση

5.2.4.3. Κοινότητες Ενδιαφέροντος (Hubs of interest)

Εκτός από τη δυνατότητα αναζήτησης το σύστημα επιτρέπει στο χρήστη να προσθέσει κάποιες ειδικές ετικέτες οι οποίες αντιπροσωπεύουν αντικείμενα που δεν υπάρχουν στα κοινοτικά δίκτυα. Στο σύστημά μας επιλέξαμε τον τομέα της βιολογικής καλλιέργειας. Έτσι προσφέρουμε τη δυνατότητα στο χρήστη να ταξινομήσει αντικείμενα των κοινοτικών δικτύων όπως: σελίδες, ομάδες και βίντεο κάτω από ετικέτες που αναπαριστούν έννοιες ενός πεδίου ενδιαφέροντος όπως παραδείγματος χάριν η βιολογική καλλιέργεια. Ενδεικτικά η δυνατότητα αυτή λειτουργεί ως εξής. Αν κάποιος χρήστης που περιηγείται στο Διαδίκτυο διαπιστώσει ότι ένα τεχνούργημα (π.χ. μια ιστοσελίδα, ένα video) εμπίπτει σε ενδιαφέροντα είτε του ίδιου του χρήστη είτε της ομάδας στην οποία ο χρήστης ανήκει, τότε του προσφέρεται η δυνατότητα να ταξινομήσει το τεχνούργημα αυτό κατάλληλα και να το καταστήσει διαθέσιμο για τον ίδιο ή τους συνεργάτες του σε μελλοντικές αναζητήσεις. Με τον τρόπο αυτό γίνονται δυνατές δυναμικές ταξινομήσεις τεχνουργημάτων κάτω από ετικέτες που επιλέγουν οι ίδιοι οι χρήστες και που αποδίδουν το νόημα που οι ίδιοι οι χρήστες επιδιώκουν. Παραδείγματος χάριν μια ιστοσελίδα μπορεί να ταξινομηθεί κάτω από την ετικέτα “βιολογική καλλιέργεια” και ένα βίντεο κάτω από την ετικέτα “διαχείριση παρασίτων με θηρευτές”. Οι ταξινομήσεις αυτές όταν τροφοδοτούνται από τις συνεισφορές πολλαπλών διαφορετικών χρηστών μπορούν να δημιουργήσουν ηλεκτρονικές κοινότητες ενδιαφέροντος που στη συνέχεια μπορούν να κατευθύνουν τόσο τα κριτήρια αναζήτησης σχετικής πληροφορίας όσο και να εξειδικεύσουν τις πρακτικές της ομάδας ή του οργανισμού που τις υιοθετεί.



Εικόνα 27: Διεπαφή των Hubs

Στην τρέχουσα έκδοση της εφαρμογής μας και για τις ανάγκες της περίπτωσης χρήσης της βιολογικής καλλιέργειας υποστηρίζεται μια αρχική κατάταξη αντικειμένων σε κατηγορίες και υποκατηγορίες που παρουσιάζονται στην Εικόνα 27. Από την εικόνα γίνεται προφανές αφενός το γεγονός ότι ένα αντικείμενο μπορεί να ταξινομηθεί σε πολλές

κατηγορίες και αφετέρου το συλλογικό μνημονικό (υπό μορφή ενός "hub cloud") που προκύπτει από τις επιλογές άλλων χρηστών.

Πέρα της ενσωματωμένης στο σύστημα ταξινόμησης, οι χρήστες μπορούν να δημιουργήσουν τις δικές τους ετικέτες και έτσι να μετατρέψουν μια ταξινόμηση από γενικευμένου τύπου (όπως αυτή στην Εικόνα 27) σε εστιασμένη και θεματικά προσαρμοσμένη. Ένα ενδεικτικό παράδειγμα δημιουργίας ετικετών ορισμένες από το χρήστη εμφανίζεται στην Εικόνα 28). Ο διάλογος στο κέντρο της σελίδας επιτρέπει στον χρήστη είτε να κατατάξει την τρέχουσα επιλογή σε μια ήδη υπάρχουσα θεματική ετικέτα είτε να δημιουργήσει νέα ετικέτα (ατομική ή δημόσια). Με τον τρόπο αυτό καθίσταται δυνατή αφενός η δημιουργία δεσμών με άλλους συμπράττοντες στο ίδιο γνωστικό αντικείμενο και αφετέρου η ανάπτυξη κοινοτικού κεφαλαίου που εμπλουτίζεται από τις επιλογές των χρηστών.

Εικόνα 28: Εμπλουτισμός ταξινόμησης από θεματικές ετικέτες ειδικού ενδιαφέροντος

5.3. Υλοποίηση & Σενάρια Χρήσης

Παρακάτω παραθέτονται τρία ενδεικτικά σενάρια χρήσης του συστήματος. Τα δύο πρώτα σενάρια χρήσης αφορούν την απλή αναζήτηση, ενώ το τρίτο αφορά την σύνθετη αναζήτηση. Για κάθε ένα από τα σενάρια χρήσης παρουσιάζεται το ερώτημα, η απόκριση του συστήματος και η μορφή των αποτελεσμάτων.

5.3.1. Σενάρια απλής αναζήτησης

Η Εικόνα 29 συνοψίζει το ερώτημα που ανακτά δεδομένα από ένα κοινωνικό ιστότοπο, αυτόν του Facebook. Η διαχείριση ενός τέτοιου ερωτήματος δρομολογείται με κατάλληλη κωδικοποίηση και την απόκριση που εμφανίζεται στην Εικόνα 30. Όσον αφορά

τα αποτελέσματα στην τρέχουσα έκδοση της βάσης δεδομένων, αυτά εμφανίζονται στην Εικόνα 31.

Εικόνα 29: Υποβολή Ερωτήματος "Select Pages from Facebook"

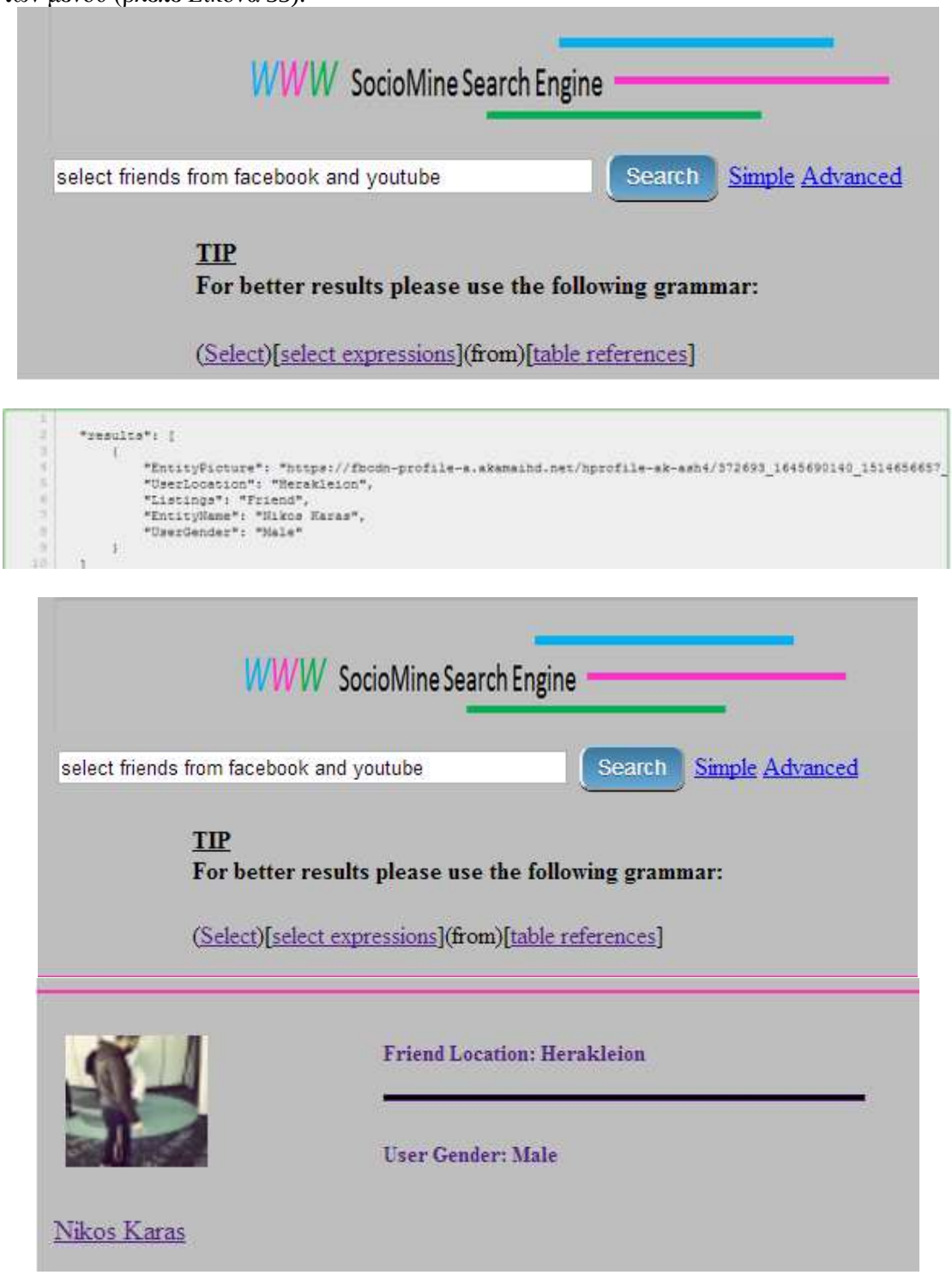
Εικόνα 30: Απόκριση Server στο ερώτημα "Select Pages from Facebook"

Εικόνα 31: Παρουσίαση Αποτελεσμάτων του ερωτήματος "Select Pages from Facebook"

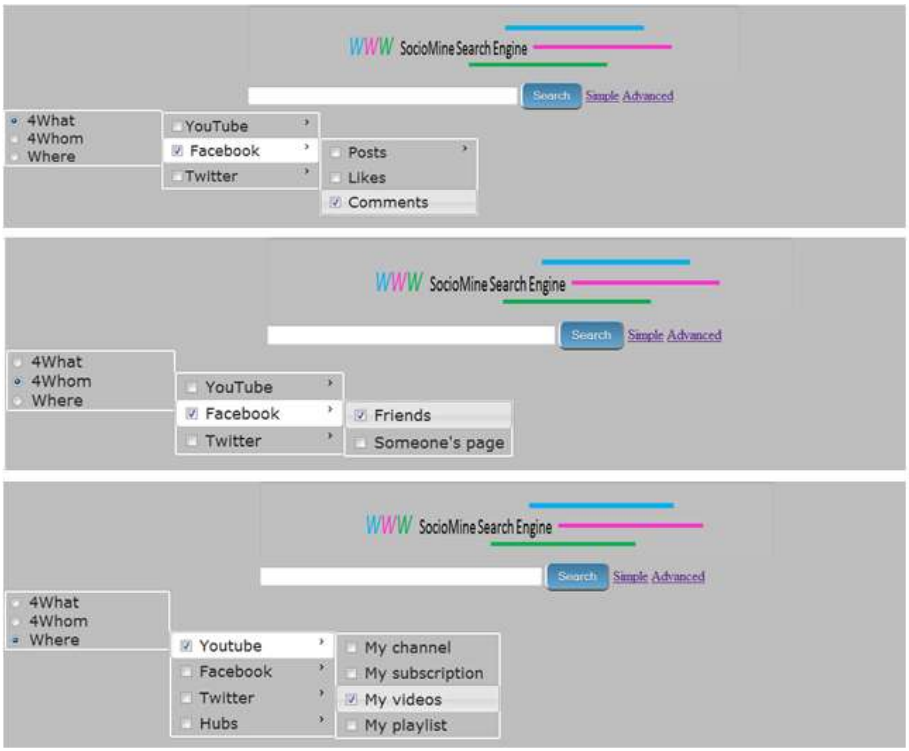
5.3.2. Σενάριο σύνθετης ή προχωρημένης αναζήτησης

Το δεύτερο παράδειγμα απλής αναζήτησης αφορά την ανάκτηση δεδομένων από περισσότερους του ενός κοινωνικούς ιστοτόπους. Ένα ενδεικτικό παράδειγμα της κατηγορίας

αυτής παρουσιάζεται στην Εικόνα 32. Η περίπτωση της σύνθετης αναζήτησης επιδιώκει να συλλέξει δεδομένα από περισσότερους του ενός κοινωνικούς ιστότοπους και να παρουσιάσει συνοπτικά τα αποτελέσματα με τρόπο ενιαίο. Ένα ενδεικτικό παράδειγμα ερωτήματος σύνθετης αναζήτησης είναι η ανάκτηση σχολίων (comments) που πραγματοποιήθηκαν από φίλους μου στο Facebook σχετικά με δικά μου video στο Youtube. Στην Εικόνα 32 αποτυπώνεται τόσο η λογική του συγκεκριμένου ερωτήματος όσο και τα ενδιάμεσα βήματα που καταλήγουν στην τελική απόκριση του συστήματος. Σε περίπτωση που ο χρήστης επιθυμεί να συλλέξει δεδομένα συγκεκριμένου τύπου τότε μπορεί να επιλέξει με την χρήση των μενού (βλέπε Εικόνα 33).



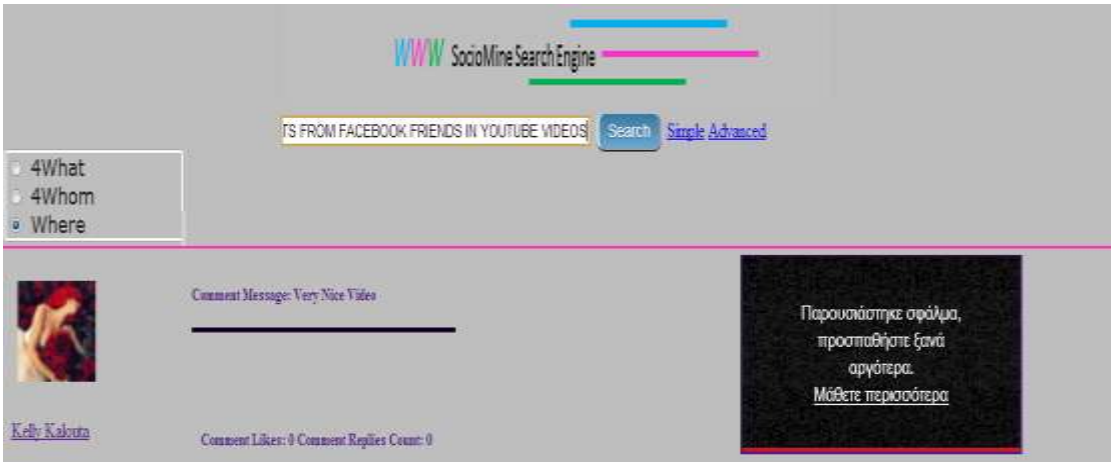
Εικόνα 32: (a)Υποβολή Ερωτήματος “Select Friends from Facebook & Youtube”, (b) Απόκριση του Server στο ερώτημα,(c) Παρουσίαση αποτελεσμάτων στο ερώτημα.



Εικόνα 33: Υποβολή ερωτήματος "Select comments from my Facebook friends in my Youtube videos"

Αντίστοιχα η απόκριση του εξυπηρετητή και τα αποτελέσματα εμφανίζονται στην Εικόνα 34 και Εικόνα 35. Υπενθυμίζεται ότι τα αποτελέσματα προκύπτουν από τα δεδομένα του στιγμιότυπου της βάσης δεδομένων που αξιοποιήθηκε για την παρούσα εργασία.

Εικόνα 34: Απόκριση Server στο ερώτημα "Select comments from my Facebook friends in my Youtube videos"



Εικόνα 35: Παρουσίαση Αποτελεσμάτων του ερωτήματος "Select comments from my Facebook friends in my Youtube videos"

5.3.3. Σχολιασμός

Οι περιπτώσεις χρήσης που μελετήθηκαν και παρουσιάστηκαν παραπάνω καταδεικνύουν τρείς τουλάχιστον περιοχές όπου οι συμβατικές μηχανές αναζήτησης υστερούν. Η πρώτη περιοχή αφορά τη δυνατότητα αξιοποίησης του συλλογικού μνημονικού και συσσωρευμένης εμπειρίας μιας κοινότητας χρηστών προκειμένου η ανάκτηση δεδομένων να καταστεί αμιγώς συνεργατική διαδικασία. Στην πλειοψηφία τους, οι σημερινές μηχανές αναζήτησης υιοθετούν την αρχή ότι η αναζήτηση δεδομένων είναι ατομική λειτουργία που διεκπεραιώνεται από ένα μεμονωμένο χρήστη ο οποίος γνωρίζει πλήρως το πεδίο διερεύνησης και μπορεί να καθορίσει με ακρίβεια τα κριτήρια ανάκτησης δεδομένων. Στην πραγματικότητα όμως κάτι τέτοιο δεν ισχύει παρά σε ειδικές περιπτώσεις και κατηγορίες χρηστών. Οι χρήστες συχνά αναζητούν διαφορετικά δεδομένα ανάλογα με τον ρόλο τους σε ένα οργανισμό, το πεδίο ενδιαφέροντος τους ή την εξειδίκευση που διαθέτουν. Σε πολλές περιπτώσεις μάλιστα η απλή καταγραφή χαρακτήρων ή προσδιοριστικών μεταβάλλουν σημαντικά τα αποτελέσματα της αναζήτησης (όσον αφορά το πλήθος αλλά και τη σειρά / κατάταξη τους). Το να μπορούν λοιπόν οι χρήστες να αξιοποιήσουν το συλλογικό μνημονικό που απορρέει από την εμπειρία άλλων χρηστών σε ίδιους ρόλους, με παρόμοια ενδιαφέροντα ή την ίδια επιχείρηση είναι σημαντικό για την κατάλληλη διαμόρφωση κριτηρίων αναζήτησης και την εστιασμένη ανάκτηση δεδομένων. Η πιλοτική έκδοση της εφαρμογής που αναπτύχθηκε στην παρούσα εργασία υποστηρίζει τέτοιου είδους αναζητήσεις μέσω του μηχανισμού των hubs. Τέτοιου είδους μηχανισμοί ουσιαστικά επιτρέπουν σε μια Διαδικτυακή ομάδα χρηστών (με κοινά ενδιαφέροντα, παρόμοιους ή/και συμπληρωματικούς ρόλους) να εξειδικεύσουν ένα πεδίο ενδιαφέροντος κατά τρόπο που αυτοί κρίνουν κατάλληλο και επομένως να αναπτύξουν κοινοτικό κεφάλαιο μέσω μιας άτυπης διαδικασίας κατανεμημένης νόησης. Περεταίρω, αυτό το κοινοτικό κεφάλαιο και η έμφυτη διαδικασία συνεχούς εμπλουτισμού τους μπορεί να τροφοδοτήσει επιμέρους δράσεις των εταίρων αλλά και να λειτουργήσει ως μηχανισμός ευγενούς ανταγωνισμού και άμιλλας μεταξύ των μελών της ομάδας.

Η δεύτερη περιοχή που χρήζει ιδιαίτερης αναφοράς σχετίζεται με την αξιοποίηση ψηφιακών ιχνών από διαφορετικούς κοινωνικούς οικισμούς με τρόπο που να εξυπηρετούνται πρωτίστως τα ενδιαφέροντα μιας ομάδας χρηστών, μιας εικονικής σύμπραξής ή ενός φορέα. Σε αυτό τον τομέα οι σημερινές μηχανές αναζήτησης παρουσιάζουν σχετική υστέρηση αφού στην πλειοψηφία τους αξιοποιούν περιορισμένο εύρος ψηφιακών ιχνών που στην καλύτερη των περιπτώσεων αποδίδουν κάποιας μορφής ιστορικού του μεμονωμένου χρήστη. Η εφαρμογή μας αποτελεί ένα βήμα προς την κατεύθυνση αξιοποίησης ευρύτερου φάσματος ψηφιακών ιχνών ενδιαφέροντος που είτε φιλοξενούνται από διαφορετικούς κοινοτικούς οικισμούς είτε υποδεικνύονται μέσω των hubs.

Τέλος, και ως απόρροια των παραπάνω αναδύεται έμμεσα η ανάγκη κατάλληλων διεπαφών χρήστη-υπολογιστή που αναδεικνύουν τα λειτουργικά χαρακτηριστικά εργαλείων με αμιγώς συνεργατικό χαρακτήρα. Προς αυτή την κατεύθυνση, η παρούσα εργασία αφενός υιοθέτησε μια δομημένη γραμματική διατύπωσης ερωτημάτων / κριτηρίων και αφετέρου προέταξε την καθοδήγηση του χρήστη έτσι ώστε να είναι δυνατή η δημιουργία ενός ερωτήματος μέσω της απόκρισης σε βασικά ερωτήματα που αφορούν το τί, ποιός και πού αναζητείται.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Η πτυχιακή εργασία που παρουσιάστηκε συνοπτικά στα προηγούμενα κεφάλαια επέτρεψε στους συγγραφείς αφενός να μελετήσουν σύνθετα θέματα που σχετίζονται με την αναζήτηση στο Διαδίκτυο γενικά και σε κοινοτικά δίκτυα ειδικότερα, και αφετέρου να εξοικειωθούν και να διεισδύσουν σε τεχνολογίες και εργαλεία της τρέχουσας γενιάς. Συνολικά, η εμπειρία που αποκτήθηκε ήταν ιδιαιτέρως χρήσιμη και προσέφερε ισχυρά εφόδια στους συγγραφείς για την μετέπειτα σταδιοδρομία τους. Ειδική αναφορά πρέπει να γίνει στην προσπάθεια που καταβλήθηκε για την πιλοτική ανάπτυξη εφαρμογών που εδράζονται σε ένα ευρύ φάσμα γνωστικών αντικειμένων όπως είναι η ανάπτυξη βάσεων δεδομένων, ο προγραμματισμός Διαδικτύου και οι διεπαφές χρήστη-υπολογιστή. Παρότι η εφαρμογή που τελικά αναπτύχθηκε χρήζει βελτιώσεων και επεκτάσεων είναι σημαντικό να τονιστεί ότι μας επέτρεψε να αποκτήσουμε σαφέστερη εικόνα για την ανάπτυξη εργαλείων και εφαρμογών λογισμικού.

Ειδικότερα, η εφαρμογή που υλοποιήθηκε στα πλαίσια της πτυχιακής αυτής εργασίας στόχευσε στο να προσφέρει στον τελικό χρήστη ένα εμπλουτισμένο αποτέλεσμα σε σχέση με το μέχρι τώρα προσφερόμενο από τις καθιερωμένες μηχανές αναζήτησης. Ουσιαστικά του δίνει τη δυνατότητα να αναζητήσει πληροφορία από διαφορετικά κοινοτικά δίκτυα, χρησιμοποιώντας μια θεμελιώδη γραμματική. Η όλη διαδικασία (προσωποποίηση) επιτυγχάνεται σε τρία βήματα:

- Για ποιόν χρήστη θέλω να κάνω αναζήτηση,
- για ποιο στοιχείο των κοινωνικών δικτύων (σχόλια, αναρτήσεις, εικόνες, κ.ο.κ.),
- και τέλος από ποιό κοινωνικό δίκτυο (facebook, twitter, youtube).

Η εμπειρία μας κατέδειξε πλήθος περιοχών που θα μπορούσαν να μελετηθούν για την βελτίωση του συστήματος. Καταρχήν, μια ενδεδειγμένη μελλοντική επέκταση του παραπάνω συστήματος θα μπορούσε να είναι προς την κατεύθυνση δημιουργίας ενός δικτυακού API καθώς επίσης και μιας πλατφόρμας κατανεμημένης πρόσβασης στο σύστημα. Αυτό θα επέτρεπε τη διασύνδεση με άλλες εφαρμογές πέραν αυτών που μελετήθηκαν πιλοτικά στο πλαίσιο της παρούσας εργασίας. Μια δεύτερη περιοχή ενδιαφέροντος αφορά τις διαδικασίες και τους μηχανισμούς ανάπτυξης και αξιοποίησης των κοινοτήτων ενδιαφέροντος που αναδύονται μέσω των hubs. Τέλος, θα ήταν επίσης χρήσιμο και σημαντικό να μελετηθούν περισσότερα κοινωνικά δίκτυα με τρόπο τέτοιο που να γίνεται εφικτή η ενιαία διαχείριση ψηφιακών ιχνών προερχόμενων από διαφορετικούς εικονικούς οικισμούς.

ΒΙΒΛΙΟΓΡΑΦΊΑ

[1] Wikipedia, "Wikipedia,the free encyclopedia," [Online]. Available: http://en.wikipedia.org/wiki/Web_search_engine. [Accessed 5 Ιούλιος 2013].

[2] Frantzis_Thrasyvoulos, «http://nemertes.lis.upatras.grr/jspui,» [Ηλεκτρονικό]. Available: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/2677/1/Diplwmatikh_Frantzis_Thrasyvoulos.pdf. [Πρόσβαση 5 Ιούλιος 2013].

[3] Wikipedia,the free encyclopedia, «Web Search Engine,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Web_search_engine. [Πρόσβαση 5 Ιούλιος 2013].

[4] Wikipedia,the free encyclopedia, «Filter Bubble,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Filter_bubble. [Πρόσβαση 5 Ιούλιος 2013].

[5] Wikipedia,the free encyclopedia, «Collaboration Search Engine,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Collaborative_search_engine. [Πρόσβαση 5 Ιούλιος 2013].

[6] G. Golovchinsky και J. Pickers, «Collaborative Exploratory Search,» σε *HCIR 2007 Workshop*, 2007.

[7] J. Pickens, G. Golovchinsky, C. Shah, P. Qvarfordt και M. Back, «Collaborative Exploratory Search,» σε *31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.

[8] M. R. Morris και E. Horvitz, «SearchTogether: An Interface for Collaborative Web Search,» σε *20th Annual ACM symposium on User Interface Software and Technology*, 2007.

[9] Wikipedia,the free encyclopedia, «Web Crawler,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Web_crawler. [Πρόσβαση 7 Ιούλιος 2013].

[10] C. Olston και M. Najork, «Web Crawling,» σε *Foundation and Trends in Information Retrieval*, 2010.

[11] Χ. Σταύρος, «Βελτιστοποίηση απόδοσης μηχανισμού διάσχισης του Διαδικτύου,» Θεσσαλονίκη, 2012.

[12] G. Pant, P. Srinivasan και F. Menczer, «Crawling The Web,» σε *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, Springer, 2004, pp. 153-178.

[13] C. Junghoo και H. Garcia-Molina, «Synchronizing a database to improve freshness,» σε *2000 ACM SIGMOD International Conference on Management of Data*, Dallas, 2000.

[14] Α. Μπάτζιος, «Εξόρυξη και Διαχείριση Σημασιολογικής Πληροφορίας στον Παγκόσμιο Ιστό,» Θεσσαλονίκη, 2009.

[15] Ι. Γαροφαλάκης και Β. Στεφανής, «Σχεδίαση και Υλοποίηση συστήματος αξιολόγησης της δομής και του περιεχομένου ιστοτόπων για κινητές συσκευές,» Πάτρα, Ιανουάριος

] 2008.

[16] F. Menczer, «ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery,» Morgan Kaufmann, Indiana, 1997.

[17] F. Menczer και R. K. Belew, «Adaptive Information Agents in Distributed Textual Enviroments,» ACM Press, Indiana, 1998.

[18] S. Brin και L. Page, «The Anatomy of a Large-Scale Hypertextual Web Search Engine,» Stanford University, Stanford, after 2000.

[19] Wikipedia,the free encyclopedia, «PageRank,» [Ηλεκτρονικό]. Available: www.en.wikipedia.org/wiki/PageRank. [Πρόσβαση 25 Οκτωβρίου 2012].

[20] G. Almpanidis, C. Kotropoulos και I. Pitas, «Combining text and link analysis for focused crawling--An application for vertical search engines,» *Information Systems,* τόμ. 32, pp. 886-908, 2007.

[21] J. M. Kleinberg, «Authority Sources in a Hyperlinked Enviroment,» *Journal of the ACM,* τόμ. 46, αρ. 5, pp. 604-632, 1999.

[22] Stanford University, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.

[23] Wikipedia,the free encyclopedia, «HITS algorithm,» [Ηλεκτρονικό]. Available: www.en.wikipedia.org/wiki/HITS_algorithm. [Πρόσβαση 25 Οκτωβρίου 2012].

[24] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson και J. Kleinberg, «Mining the Web's Link Structure,» *Computer,* τόμ. 32, αρ. 8, pp. 60-67, 1999.

[25] D. Cohn και H. Chang, «Learning to Probabilistically Identify Authoritative Documents,» σε *17th International Conference on Machine Learning*, San Francisco, 2000.

[26] Wikipedia,the free encyclopedia, «Google Panda,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Google_Panda. [Πρόσβαση 7 Ιούλιος 2013].

[27] Wikipedia,the free encyclopedia, «Google Penguin,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Google_Penguin. [Πρόσβαση 7 Ιούλιος 2013].

[28] D. Sullivan, «Search Engine Land,» 26 April 2012. [Ηλεκτρονικό]. Available: http://searchengineland.com/the-penguin-update-googles-webspam-algorithm-gets-official-name-119623. [Πρόσβαση 1 June 2013].

[29] Google, «Google Webmaster Tools,» Google, 10 June 2013. [Ηλεκτρονικό]. Available: https://support.google.com/webmasters/answer/35769?hl=en#3. [Πρόσβαση 1 July 2013].

[30] M. Cutts, «Another step to reward high-quality sites,» Google, 24 April 2012. [Ηλεκτρονικό]. Available: http://insidesearch.blogspot.gr/2012/04/another-step-to-

reward-high-quality.html. [Πρόσβαση 1 July 2013].

[31] D. K. Sharma και A. K. Sharma, «A Comparative Analysis of Web Page Ranking Algorithms,» *International Journal on Computer Science and Engineering,* τόμ. 2, αρ. 6, pp. 2670-2676, 2010.

[32] W. Xing και A. Ghorbani, «Weighted PageRank Algorithm,» σε *2nd Annual Conference on Communication Networks & Services Research*, 2004.

[33] R. Baeza-Yates και E. Davis, «Web Page Ranking Using Link Attributes,» σε *13th International Word Wide Web Conference on Alternate Track Papers & Posters*, 2004.

[34] K. Futjimura, I. Takafumi και S. Masayuki, *The EigenRumor Algorithm for Ranking Blogs,* Chiba: Second Anual Workshop on the Web Bloging EcoSystem, 2005.

[35] N. Yazdani και Z. Bidoki, «DistanceRank: An intelligent ranking algorithm for web pages,» *Information Processing and Management,* 2007.

[36] H. e. a. Jiang, «TIMERANK: A method of Improving Ranking Scores by Visited Time,» σε *7th International Conference on Machine Learning and Cybernetics*, Kunming, 2008.

[37] C. Clarke και G. Cormack, «Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System,» University Of Waterloo, 1995.

[38] D. A. Grossman και O. Frieder, Information Retrieval: Algorithms and Heuristics, Springer, 2004.

[39] C. T. a. S. Dwarkadas, «Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval,» σε *1st Symposium On Networked Systems Design and Implementation*, San Francisco, California,USA, 2004.

[40] Wikipedia, the online encyclopedia, «Keyword stuffing,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Keyword_stuffing. [Πρόσβαση 1 July 2013].

[41] Google, «Keyword Stuffing,» Google, 27 May 2013. [Ηλεκτρονικό]. Available: https://support.google.com/webmasters/answer/66358?hl=en. [Πρόσβαση 1 July 2013].

[42] D. I. Moldovan και R. Mihalcea, «Using WordNet and lexical operators to improve Internet searches,» *IEEE Internet Computing,* τόμ. 4, αρ. 1, pp. 34-43, 2000.

[43] D. Buscaldi, P. Rosso και E. S. Arnal, «A wordnet-based query expansion method for geographical information retrieval,» σε *Working Notes for the CLEF Workshop*, Vienna, Austria, 2005.

[44] P. M. Kruse, A. Naujocks, D. Roesner και M. Kunze, «Clever search: A wordnet based wrapper for internet search engines,» σε *Proceedings of the 2nd GermaNet Workshop*, 2005.

[45] R. Guha, R. McCool και E. Miller, «Semantic search,» σε *Proceedings of the 12th international conference on World Wide Web*, 2003.

[46] C. Rocha, D. Schwabe και M. P. de Aragão, «A hybrid approach for searching in the

] semantic web,» σε *Proceedings of the 13th international conference on World Wide Web*, 2004.

[47 J. Heflin και J. Hendler, «Searching the Web with SHOE,» σε *Artificial Intelligence for*
] *Web Search. Papers from the AAAI Workshop*, Menlo Park, CA, 2000.

[48 A. Maedche, S. Staab, N. Stojanovic, R. Studer και Y. Sure, «SEAL - A Framework for
] Developing Semantic Web Portals,» σε *Proceedings of the 18th British National Conference on Databases*, 2001.

[49 E. Mäkelä, E. Hyvönen και T. Sidoroff, «View-based user interfaces for information
] retrieval on the semantic web,» σε *Proceedings of the ISWC-2005 Workshop End User Semantic Web Interaction*, Galway, Ireland, 2005.

[50 E. Mäkelä, E. Hyvönen, S. Saarela και K. Viljanen, «OntoViews - A Tool for Creating
] Semantic Web Portals,» σε *Proceedings of hte Third Inernational Semantic Web Conference*, Hirosima, Japan, 2004.

[51 D. Reynolds, P. Shabajee και S. Cayzer, «Semantic Information Portals,» σε *Proceedings*
] *of the 13th International World Wide Web Conference on Alternate track papers & posters*, Manhattan, NY, USA, 2004.

[52 E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila και S.
] Kettula, «Museumfinland - finnish museums on the semantic web,» *Web Semantics: Science, Services and Agents on the World Wide Web*, τόμ. 3, αρ. 2-3, pp. 75-242, 2005.

[53 E. Hyvönen και E. Mäkelä, «Semantic autocompletion,» σε *Proceedings of the 1st Asia*
] *Semantic Web Conference*, Beijing, China, 2005.

[54 E. Mäkelä, K. Viljanen, P. Lindgren, M. Laukkanen και E. Hyvönen, «Semantic yellow
] page service discovery: The veturi portal,» σε *4th International Semantic Web Conference*, Galway, Ireland, 2005.

[55 N. Athanasis, V. Christoforides και D. Kotzinos, «Generating on the fly queries for the
] semantic web: The ics-forth graphical rql interface (grql),» σε *Third International Semantic Web Conference*, Hiroshima, Japan, 2004.

[56 R. Fikes, P. Hayes και I. Horrocks, «OWL-QL: A language for deductive query
] answering on the Semantic Web,» *Web Semantics: Science, Services and Agents on the World Wide Web*, τόμ. 2, αρ. 1, pp. 19-29, 2004.

[57 Y. Du, Y. Shi και X. Zhao, «Using Spam Farm to Boost PageRank,» ACM, 2007.
]

[58 Wikipedia,the free encyclopedia, «Page Hijacking,» [Ηλεκτρονικό]. Available:
] www.en.wikipedia.org/wiki/302_Google_jacking. [Πρόσβαση 25 Οκτωβρίου 2012].

[59 Wikipedia, the free encyclopedia, «Link Farm,» [Ηλεκτρονικό]. Available:
] en.wikipedia.org/wiki/Link_Farm. [Πρόσβαση 25 Οκτωβρίου 2012].

[60 Universiteit Leiden, «Internet History- Search Engines,» [Ηλεκτρονικό]. Available:

] www.internethistory.leidenuniv.nl/index.php3?c=7. [Πρόσβαση 28 October 2012].

[61 C. Castillo, «Effective Web Crawling,» 2004.
]

[62 M. Koster, «Robots in the Web: threat or treat?,» 2007. [Ηλεκτρονικό]. Available:
] http://www.robotstxt.org/threat-or-treat.html. [Πρόσβαση 30 October 2012].

[63 M.Koster, «About /robots.txt,» [Ηλεκτρονικό]. Available:
] http://www.robotstxt.org/robotstxt.html. [Πρόσβαση 31 October 2012].

[64 Wikipedia, the free encyclopedia, «Cloacking,» [Ηλεκτρονικό]. Available:
] http://en.wikipedia.org/wiki/Cloaking. [Πρόσβαση 1 July 2013].

[65 Google, «Cloacking,» Google, 27 May 2013. [Ηλεκτρονικό]. Available:
] https://support.google.com/webmasters/answer/66355?hl=en. [Πρόσβαση 1 July 2013].