ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Ψηφιακών Συστημάτων

Κατεύθυνση «Δικτυοκεντρικά Συστήματα»

# Η ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ GOOGLE

## Περιγραφή λειτουργίας, επιθέσεις και τρόποι αντιμετώπισης

**Λαμπρόπουλος Μιχάλης**


Επιβλέπων Καθηγητής: Λαμπρινουδάκης Κωνσταντίνος

# ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική έχει ως αντικειμενό της την έρευνα πάνω στις επιθέσεις που μπορεί να δεχθεί η διαδικτυακή μηχανή αναζήτησης Google, καθώς και την παρουσίαση μεθόδων και τεχνικών με βάση τις οποίες αντιμετωπίζει τις επιθέσεις αυτές. Το φάσμα των απειλών που μπορεί να δεχθεί η εν λόγω μηχανή αναζήτησης είναι μεγάλο και μπορεί να την ζημιώσει σε πολλούς τομείς, βλάπτοντας την αξιοπιστία της, την λειτουργία της και τα οικονομικά της κέρδη και ωφέλη. Βέβαια, πολλές από τις επιθέσεις αυτές δεν αφορούν μόνο την Google κάθε αυτή αλλά μπορούν να δημιουργήσουν προβλήματα και σε πολλές άλλες μηχανές αναζήτησης.

Περιληπτικά, η παρούσα εργασία είναι δομημένη στα παρακάτω κεφάλαια:

**Κεφάλαιο 1**

Επεξηγείται συνοπτικά η αρχιτεκτονική και η λειτουργία των μηχανών αναζήτησης, δίνοντας βαρύτητα στην λειτουργία των web crawlers.

**Κεφάλαιο 2**

Γίνεται αναφορά στην λειτουργία της μηχανής αναζήτησης Google, όσον αφορά το λογισμικό της, το υλικό της καθώς και στον αλγόριθμο PageRank, ο οποίος την έχει καθιερώσει ως κυρίαρχη μηχανή αναζήτησης.

**Κεφάλαιο 3**

Εξηγούνται συνοπτικά οι κυριότερες περιπτώσεις επιθέσεων, δίνοντας βάση στο web spam, καθώς και τις επιθέσεις στα οικονομικά προϊόντα της Google.

**Κεφάλαιο 4**

Αναφέρονται μέθοδοι και τεχνικές που χρησιμοποιεί η Google, για την αντιμετώπιση των επιθέσεων που περιγράφονται στο Κεφάλαιο 3, με πιο χαρακτηριστική την χρήση του αλγορίθμου TrustRank

## Περιεχόμενα

# ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

# ΚΕΦΑΛΑΙΟ 1
# ΑΝΑΣΚΟΠΗΣΗ ΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ

## 1.1 Εισαγωγή

Στο κεφάλαιο αυτό θα αναλύσουμε συνοπτικά την λειτουργία και τα βασικότερα χαρακτηριστικά των μηχανών αναζήτησης. Θα δοθεί περισσότερη σημασία στον τρόπο με τον οποίο λειτουργούν οι μηχανές αναζήτησης την σημερινή εποχή, ενώ θα επικεντρωθούμε στα χαρακτηριστικά τα οποία έχει η μηχανή αναζήτησης Google.

## 1.2 Ιστορική Αναδρομή

Οι πρώτες μηχανές αναζήτησης ήταν στην ουσία μια λίστα από συνδέσμους (links) σε διάφορους web servers. Η πρώτη από αυτές ήταν τοποθετημένη στο CERN και παραμένει ίδια, από το 1992[1], για ιστορικούς και μόνο λόγους. Η πρώτη πραγματική μηχανή αναζήτησης, δημιουργήθηκε από τον Oscar Nierstrasz, στις 2 Σεπτεμβρίου 1993 και είχε το όνομα W3Catalog[2]. Η συγκεκριμένη μηχανή αναζήτησης ήταν γραμμένη στην γλώσσα προγραμματισμού Perl και η μόνη δυνατότητα που είχε ήταν η αναζήτηση λέξεων-κλειδιών σε μια συλλογή από HTML αρχεία.

Στα μέσα της δεκαετίας του ’90, πολλές εταιρείες που άρχισαν να δραστηριοποιούνται στον χώρο του Διαδικτύου, συνειδητοποίησαν ότι η αγορά των μηχανών αναζήτησης ήταν πολλά υποσχόμενη. Έτσι συνεπώς, ενώ αρχικά οι διαδικτυακές πύλες (portals) της εποχής εκείνης χρησιμοποιούσαν τους διαδικτυακούς καταλόγους (web directories – την ιεραρχικά δομημένη και ταξινομημένη παρουσίαση πολλών διαδικτυακών τόπων ίδιου περιεχομένου) επένδυσαν στην έρευνα πάνω στον τομέα των διαδικτυακών μηχανών αναζήτησης. Έτσι, το 1996, οι 5 πιο διαδεδομένες[3] μηχανές αναζήτησης ήταν οι Yahoo!, Magellan, Lycos, Infoseek και Excite.

Το 2000, η πιο δημοφιλής πλέον μηχανή αναζήτησης, η Google, έκανε την εμφάνισή της. Η συγκεκριμένη εταιρεία αναζήτησης μπορούσε να επιτύχει καλύτερα

[1] http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html
[2] http://groups.google.com/group/comp.infosystems.www/browse_thread/thread/2176526a36dc8bd3/2718fd17812937ac?hl=en&lnk=gst&q=Oscar+Nierstrasz#2718fd17812937ac
[3] http://articles.latimes.com/1996-04-01/business/fi-53780_1_netscape-home

αποτελέσματα σε πολλές αναζητήσεις, χρησιμοποιώντας μια καινοτόμα για την εποχή τεχνική, η οποία ονομάζεται PageRank (βαθμολογία ιστοσελίδων). Όπως θα δούμε και παρακάτω, το PageRank είναι στην ουσία ένας επαναληπτικός αλγόριθμος, ο οποίος βαθμολογεί μια ιστοσελίδα/διαδικτυακό τόπο με μια βάση από κριτήρια. Ένα από τα πιο σημαντικά κριτήρια αυτά, είναι το πόσοι σύνδεσμοι οδηγούν στην εν λόγω ιστοσελίδα από άλλες ιστοσελίδες). Εώς και το 2009, η Google καταλαμβάνει την πρώτη θέση στην αγορά των μηχανών αναζήτησης στις περισσότερες χώρες του κόσμου[4], διατηρώντας ένα σημαντικά μεγάλο προβάδισμα σε σχέση με τους ανταγωνιστές της.

## 1.3 Η λειτουργία μιας διαδικτυακής μηχανής αναζήτησης

Στην ενότητα αυτή θα περιγράψουμε την βασική λειτουργία μιας σύγχρονης μηχανής αναζήτησης. Μια μηχανή αναζήτησης, ακολουθεί κατά βάση τα ακόλουθα βήματα:

- Web Crawling
- Ευρετηριοποίηση (Indexing)
- Αναζήτηση με βάση λέξεις κλειδιά (Web Search Query)

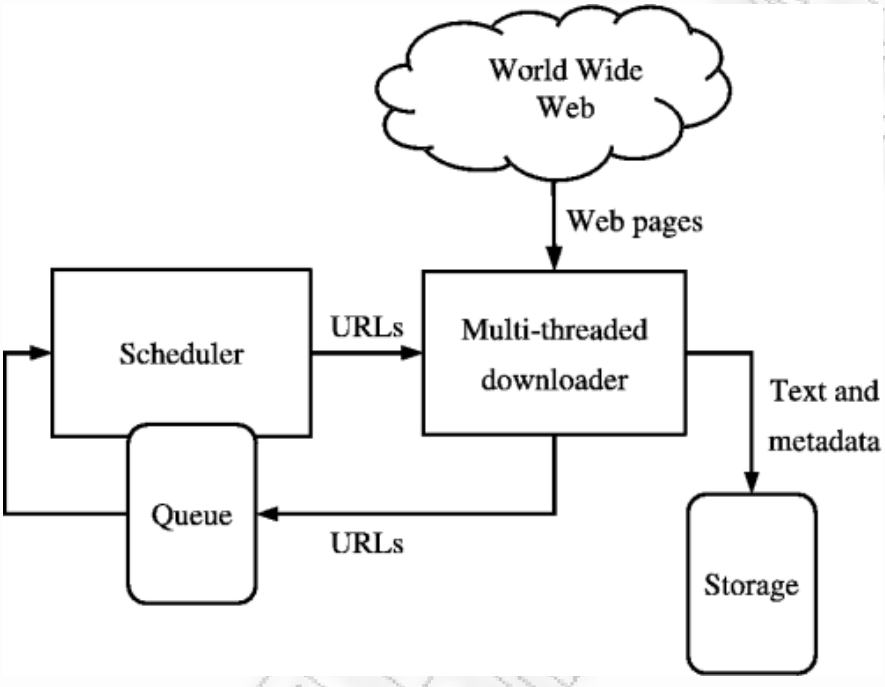
Θα αναλύσουμε συνοπτικά τα συστατικά στοιχεία των μηχανών αναζήτησης παρακάτω.

### 1.3.1 Web Crawling

Ο Web Crawler (διεθνής όρος) είναι ένα λογισμικό, το οποίο περιηγείται το Διαδίκτυο, με έναν αυτόματο, μεθοδικό τρόπο κι έχει ως σκοπό την αποθήκευση ή/και ανακάλυψη νέου περιεχομένου, έτσι ώστε να αξιοποιηθεί μελλοντικά από μια μηχανή αναζήτησης. Σε πολλές περιπτώσεις, εκτός από την κύρια αυτή λειτουργία του, ένας web crawler ελέγχει για την εγκυρότητα των συνδέσμων μέσα σε μια ιστοσελίδα κι ελέγχει για την σωστή συγγραφή (σύμφωνα με τα πρότυπα της W3C) του κώδικα HTML αυτής. Σε γενικές γραμμές, ένας Web crawler αρχίζει την ακολουθία των ενεργειών που εκτελεί, μέσω μιας αρχικής λίστας από διευθύνσεις διαδικτυακών τόπων και στην συνέχεια επεκτείνεται σε περαιτέρω διαδικτυακούς

[4] http://www.searchengineoptimizationcompany.ca/seoblog/World%20Wide%20Web/search-engines/worldwide-search-engine-market-share-part-1/05012009

τόπους, ακολουθώντας τους συνδέσμους που υπάρχουν στις αρχικές σελίδες. Ο Web crawler επαναλαμβάνει τις ενέργειες αυτές, εως ότου ικανοποιηθεί κάποια συνθήκη ή όταν έχει εξετάσει έναν μεγάλο αριθμό σελίδων. Η αρχιτεκτονική ενός τυπικού Web crawler παρουσιάζεται στην Εικόνα 1.1.




**Εικόνα 1.1:** η τυπική αρχιτεκτονική ενός Web crawler[5]

Η σημερινή μορφή του Διαδικτύου κάνει την λειτουργία των Web crawlers αρκετά δύσκολη, για τους παρακάτω λόγους:

- Το σημερινό Διαδίκτυο περιλαμβάνει ένα τεράστιο όγκο δεδομένων.
- Η δομή του Διαδικτύου και το περιεχόμενο πολλών ιστοσελίδων αλλάζει καθημερινώς.
- Πολλές σελίδες είναι δυναμικές κι όχι στατικές. Συνεπώς η εξέταση ενός στιγμιότυπού τους από τον Web crawler δεν είναι η μοναδική.

Ο συνδυασμός των παραπάνω χαρακτηριστικών δημιουργεί έναν πολύ μεγάλο αριθμό από URLs, τα οποία θα πρέπει να εξεταστούν. Με βάση το γεγονός αυτό, ένας Web Crawler μπορεί να εξετάσει έναν περιορισμένο σχετικά αριθμό διαδικτυακών τόπων και συνεπώς θα πρέπει με την εφαρμογή κάποιας «πολιτικής» (Web Crawler policy) να επικεντρωθεί σε κάποιους διαδικτυακούς τόπους, σε μεγαλύτερο βαθμό από ότι κάποιους άλλους. Η έρευνα πάνω στις διάφορες πολιτικές για την βελτιστοποίηση της

[5] http://www.codeproject.com/KB/IP/Crawler/WebCrawlerArchitecture.png

λειτουργίας των Web Crawlers, έχει αποτελέσει αντικείμενο πολλών ερευνητικών ομάδων τα τελευταία χρόνια [1].

Σε γενικές γραμμές, ένας Web Crawler εφαρμόζει τις ακόλουθες πολιτικές:

- μία **πολιτική επιλογής (selection policy)**, που ορίζει ποιες ιστοσελίδες θα πρέπει να εξετάσει.
- μία **πολιτική επανεπίσκεψης (re-visit policy)**, που ορίζει ποιες ιστοσελίδες θα πρέπει να επανεξετάσει για ενδεχόμενες αλλαγές στο περιεχόμενό τους.
- μία **πολιτική αποφυγής επιπρόσθετου φόρτου (politeness policy)**, που ορίζει το ότι οι επισκεπτόμενοι διαδικτυακοί τόποι δεν θα πρέπει να επιβαρύνονται από υπερβολικό αριθμό επισκέψεων του Web Crawler.
- μία **πολιτική παραλληλισμού (parallelization policy)**, που ορίζει τον συντονισμό των διάφορων Web Crawlers, οι οποία δουλεύουν με κατανεμημένο τρόπο.

Θα περιγράψουμε συνοπτικά τις πολιτικές αυτές παρακάτω.

**Πολιτική Επιλογής (Selection Policy)**

Δεδομένου του μεγέθους του σημερινού Διαδικτύου, ακόμα και οι μεγάλες διαδικτυακές μηχανές αναζήτησης, καλύπτουν μόνο ένα μέρος των δεδομένων τα οποία παρατίθενται. Μία μελέτη των Lawrence και Giles[2], έδειξε ότι περί το έτος 1999, καμία μηχανή της εποχής δεν είχε αποθηκεύσει πάνω από το 16% του Διαδικτύου. Η πιο ενδιαφέρουσα πολιτική επιλογής είναι αυτή του επικεντρωμένου Web Crawling (Focused Web Crawling). Σύμφωνα με την τεχνική αυτή, ένας Web Crawler επισκέπτεται τους συνδέσμους μιας ιστοσελίδας, όπου υπάρχει μεγάλη πιθανότητα τα περιεχόμενό τους να είναι σχετικό με το περιεχόμενο της τρέχουσας σελίδας. Η πληροφορία για το περιεχόμενο μιας ιστοσελίδας την οποία ο Web Crawler δεν έχει επισκεφθεί ακόμα, μπορεί να εξαχθεί από το κείμενο του συνδέσμου που οδηγεί στην ιστοσελίδα αυτή είτε από το URL της ιστοσελίδας. Βέβαια, δεν μπορούμε να θεωρήσουμε ότι τα στοιχεία αυτά δίνουν μια έγκυρη ένδειξη για το περιεχόμενο της ιστοσελίδας η οποία συνδέεται.

**Πολιτική Επανεπίσκεψης (Re-Visit Policy)**

Η δομή του σημερινού Διαδικτύου είναι δυναμική και το περιεχόμενό του αλλάζει συνεχώς. Για παράδειγμα, διαδικτυακοί τόποι περιοδικών ή εφημερίδων ενδέχεται να αλλάζουν 10 με 20 φορές την ημέρα, ίσως και παραπάνω. Η επίσκεψη ενός μέρους του Διαδικτύου από έναν Web Crawler, είναι μια διαδικασία που μπορεί να διαρκέσει

εβδομάδες, ακόμα και μήνες. Όταν η διαδικασία αυτή τελειώσει, πολλές από τις ιστοσελίδες τις οποίες επισκέφθηκε, είναι σίγουρο ότι θα έχουν αλλάξει. Πολλές έρευνες έχουν πραγματοποιηθεί στον τομέα των πολιτικών επανεπίσκεψης τα τελευταία χρόνια, με πιο αξιόλογη την έρευνα, είναι αυτή των Cho και Garcia-Molina[3]. Στην έρευνα αυτή, για κάθε URL της βάσης δεδομένων ενός Web Crawler, έστω $e_i$, ορίζονται δύο μετρικές:

1. η μετρική **freshness**, η οποία δηλώνει το πόσο πρόσφατο («φρέσκο») είναι το URL την χρονική στιγμή t και ορίζεται ως:

$$F(e_i; t) = \begin{cases} 1 & αν\ το\ e_i\ είναι\ ενημερωμένο\ την\ χρονική\ στιγμή\ t \\ 0 & σε\ διαφορετική\ περίπτωση \end{cases}$$

2. η μετρική της **ηλικίας (age)**, η οποία δηλώνει την ηλικία ενός URL την χρονική στιγμή t και ορίζεται ως:

$$A(e_i; t) = \begin{cases} 1 & αν\ το\ e_i\ είναι\ ενημερωμένο \\ & την\ χρονική\ στιγμή\ t \\ t - τελευταία\ χρονική\ ενημέρωση\ του\ e_i & αλλιώς \end{cases}$$

Αν υποθέσουμε ότι τα δεδομένα τα οποία είναι αποθηκευμένα στην βάση δεδομένων μεταβάλλονται με βάση κάποια κατανομή (οι Cho και Garcia-Molina υποθέτουν ότι μεταβάλλονται σύμφωνα με την κατανομή Poisson), μπορούμε να προσεγγίσουμε τον ρυθμό (ή συχνότητα) μεταβολής λ για κάθε ένα από τα URLs $e_i$ της βάσης δεδομένων. Με βάση λοιπόν τον ρυθμό μεταβολής λ, μπορούμε να ορίσουμε τις παρακάτω πολιτικές επανεπίσκεψης:

**Ομοιόμορφη πολιτική (Uniform change-frequency model):** σύμφωνα με την πολιτική αυτή, υποθέτουμε ότι όλα τα URLs της βάσης αλλάζουν με τον ίδιο ρυθμό. Η πολιτική αυτή είναι χρήσιμη σε περιπτώσεις όπου δεν γνωρίζουμε τον ρυθμό μεταβολής για κάποια από τα URLs της βάσης, οπότε θεωρούμε ότι όλα έχουν ρυθμό μεταβολής ίσο με τον μέσο όρο των ρυθμών μεταβολής των URLs, τα οποία γνωρίζουμε. Θα μπορούσε επίσης είναι χρήσιμη σε περιπτώσεις όπου οι ρυθμοί μεταβολής διαφέρουν ελάχιστα μεταξύ τους.

**Μη Ομοιόμορφη πολιτική (Non-Uniform change-frequency model):** σύμφωνα με την πολιτική αυτή, υποθέτουμε ότι κάθε URL $e_i$ έχει τον δικό του ρυθμό μεταβολής $\lambda_i$. Όπως είναι φυσικό, για να μπορέσει να εκτελεστεί η συγκεκριμένη πολιτική, θα πρέπει να είναι γνωστοί οι ρυθμοί μεταβολής για κάθε URL.

**Πολιτική Αποφυγής Επιπρόσθετου Φόρτου (Politeness Policy)**

Οι Web Crawlers έχουν την δυνατότητα να αποκτήσουν πρόσβαση στα δεδομένα μιας ιστοσελίδας, πολύ πιο γρήγορα και σε μεγαλύτερο βάθος, απ'ότι ένας ανθρώπινος χρήστης. Η δυνατότητά τους αυτή, μπορεί ενδεχομένως να δημιουργήσει προβλήματα στην απόδοση ενός διαδικτυακού τόπου. Για παράδειγμα, εάν ένας Web Crawler πραγματοποιεί πολλές αιτήσεις (HTTP requests) ανά δευτερόλεπτο ή/και κατεβάζει μεγάλα σε μέγεθος αρχεία ταυτόχρονα, ένας web server θα σπαταλά αρκετή υπολογιστική δύναμη αλλά και αρκετό εύρος ζώνης (bandwidth) ώστε να είναι σε θέση να ικανοποιήσει τις αιτήσεις αυτές. Μπορούμε να συνοψίσουμε το κόστος χρήσης των Web Crawlers παρακάτω:

- καταναλώνουν πολλούς πόρους του δικτύου από το οποίο εκτελούνται, ειδικά στην περίπτωση όπου πολλοί Web Crawlers εκτελούνται παράλληλα από το ίδιο δίκτυο.
- δημιουργούν προβλήματα στην λειτουργία των διαδικτυακών τόπων στους οποίους στοχεύουν, ειδικά όταν η συχνότητα των αιτήσεων είναι μεγάλη.
- Web Crawlers με λάθη στην υλοποίησή τους, μπορεί να οδηγήσουν πολλούς web servers σε τερματισμό της λειτουργίας τους (π.χ. λόγω προγραμματιστικού λάθους ένας Web Crawler μπορεί να κάνει συνεχόμενες αιτήσεις μέσω μιας ατέρμονης επανάληψης).
- Web Crawlers, οι οποίοι χρησιμοποιούνται από κοινούς χρήστες του Διαδικτύου κι όχι από μηχανές αναζήτησης, μπορεί να δημιουργήσουν πολλά προβλήματα.

Για τον περιορισμό των παραπάνω προβλημάτων, έχουν προταθεί πολλές λύσεις κατά καιρούς, με πιο σημαντικές και πρακτικές τις εξής:

- η ύπαρξη ενός άνω ορίου του αριθμού επισκέψεων ενός Web Crawler σε έναν διαδικτυακό τόπο ανά 24 ώρες.
- οι αιτήσεις που θα κάνει ένας Web Crawler θα πρέπει να γίνονται με μία χρονική καθυστέρηση μεταξύ τους (συνήθως της τάξης των δευτερολέπτων), έτσι ώστε ο διαδικτυακός τόπος να μην υπερφορτώνεται.
- η γενική αποδοχή του Robots Exclusion Protocol, το οποίο δίνει την δυνατότητα στον διαχειριστή ενός διαδικτυακού τόπου, να δώσει πληροφορίες για το ποιες ιστοσελίδες και πότε ένας Web Crawler μπορεί να επισκεφθεί.

**Πολιτική Παραλληλισμού (Parallelization Policy)**

Ένας παράλληλος Web Crawler είναι ένας Web Crawler που εκτελεί πολλαπλές διεργασίες του παράλληλα (ή κατανεμημένα). Ο στόχος είναι φυσικά η μεγιστοποίηση του αριθμού των διαδικτυακών τόπων που επισκέπτονται και το μεγαλύτερο εύρος αποτελεσμάτων. Φυσικά η παράλληλη λειτουργία ενός Web Crawler θα πρέπει να ικανοποιεί την εξής βασική απαίτηση: να μην επιτρέπει την επίσκεψη του ίδιου διαδικτυακού τόπου από δύο ή περισσότερους Web Crawlers. Ειδικά όσον αφορά το κατέβασμα μεγάλων σε μέγεθος αρχείων, η απαίτηση αυτή είναι επιτακτική. Για την αποφυγή της περίπτωσης αυτής, υπάρχει συνήθως κάποιος έλεγχος, έτσι ώστε όταν ένα νέο URL ανακαλύπτεται, ανατείθεται σε έναν και μόνο Web Crawler και όχι σε περισσότερους. Φυσικά ο έλεγχος αυτός δεν είναι πάντα εύκολος και σε μερικές περιπτώσεις ίσως αποβεί ιδιαίτερα χρονοβόρος.

### 1.3.2 Ευρετηριοποίηση (Indexing)

Η ευρετηριοποίηση (indexing), όσον αφορά της διαδικτυακές μηχανές αναζήτησης, είναι μια μέθοδος για την συλλογή, την κατηγοριοποίηση, την αποθήκευση δεδομένων, έτσι ώστε η αναζήτηση στα δεδομένα αυτά να γίνεται με γρήγορο και αποδοτικό τρόπο. Οι σύγχρονες μηχανές αναζήτησης, δεν επικεντρώνονται μόνο στην ευρετηριοποίηση αρχείων κειμένου, αλλά έχουν επεκταθεί και σε άλλους τύπους δεδομένων, όπως αρχείο video, μουσικής και εικόνων.

Ο κύριος σκοπός ενός ευρετηρίου είναι φυσικά η αύξηση της απόδοσης της αναζήτησης επί των αποθηκευμένων δεδομένων. Χωρίς την χρήση κάποιου ευρετηρίου, η μηχανή αναζήτησης θα έπρεπε να αναζητήσει πληροφορίες σε όλα τα τα αποθηκευμένα δεδομένα, γεγονός που θα είχε μεγάλο υπολογιστικό κόστος. Για παράδειγμα, η ευρετηριοποίηση ενός συνόλου δεδομένων αποτελούμενο από 20.000 έγγραφα, θα μπορούσε να επιστρέψει την απάντηση μιας επερώτησης σε milliseconds, ενώ η σειριακή αναζήτηση των αρχείων αυτών, θα μπορούσε να διαρκέσει ώρες. Φυσικά η αποθήκευση του ευρετηρίου καταλαμβάνει επιπρόσθετο χώρο, κάτι όμως που είναι μηδαμινό μπροστά στα οφέλη τα οποία έχει.

Οι παράγοντες τους οποίους μια μηχανή αναζήτησης θα πρέπει να λάβει υπ'όψη, όσον αφορά το θέμα της ευρετηριοποίησης, είναι οι παρακάτω:

- την συχνότητα με την οποία νέα δεδομένα εισάγονται στο ευρετήριο.

- την δομή με την οποία αποθηκεύονται τα δεδομένα στο ευρετήριο.
- το μέγιστο μέγεθος που μπορεί να έχει το ευρετήριο.
- τον μέσο χρόνο αναζήτησης των δεδομένων.
- το ποσοστό του λάθους που αφορούν τα αποτελέσματα μιας αναζήτησης.

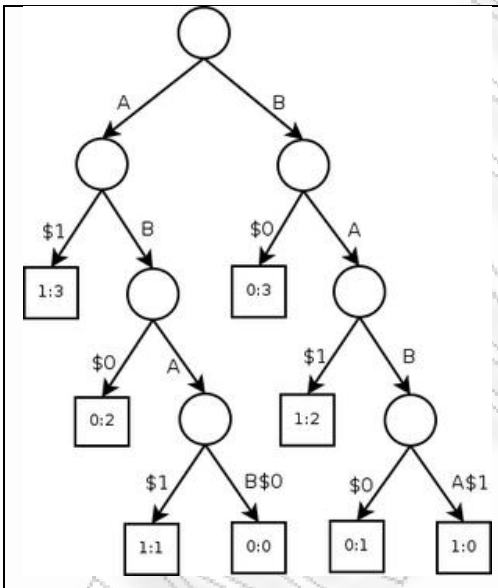
Επιγραμματικά, οι πιο συνηθισμένες δομές ευρετηρίων που χρησιμοποιούνται από τις μηχανές αναζήτησης, είναι οι παρακάτω:

**Δέντρα καταλήξεων (suffix trees):** αποθηκεύουν τις καταλήξεις των λέξεων, σε μορφή δέντρου (Εικόνα 1.2).

**Πίνακας όρων (Document-term matrix):** αποθηκεύει την εμφάνιση λέξεων σε έγγραφα, με την χρήση ενός 2-άστατου πίνακα (Εικόνα 1.3).

**Ορθό ευρετήριο (Forward index):** αποθηκεύει τις λέξεις για κάθε έγγραφο (η αναζήτηση γίνεται κατά έγγραφο – Εικόνα 1.4).

**Ανάστροφο ευρετήριο (Inverted index):** αποθηκεύει τα έγγραφα στα οποία βρίσκεται η κάθε λέξη (η αναζήτηση γίνεται κατά λέξη – Εικόνα 1.5).




**Εικόνα 1.2**

| | **το** | **τόπι** | **Από** |
|---|---|---|---|
| **A1** | 1 | 2 | 1 |
| **A2** | 0 | 2 | 2 |
| **A3** | 1 | 0 | 1 |

**Εικόνα 1.3**

| **Λέξεις** | **Αρχεία** |
|---|---|
| το | A1, A3 |
| Νίκος | A1, A4, A5 |
| από | A2, A5, A9 |

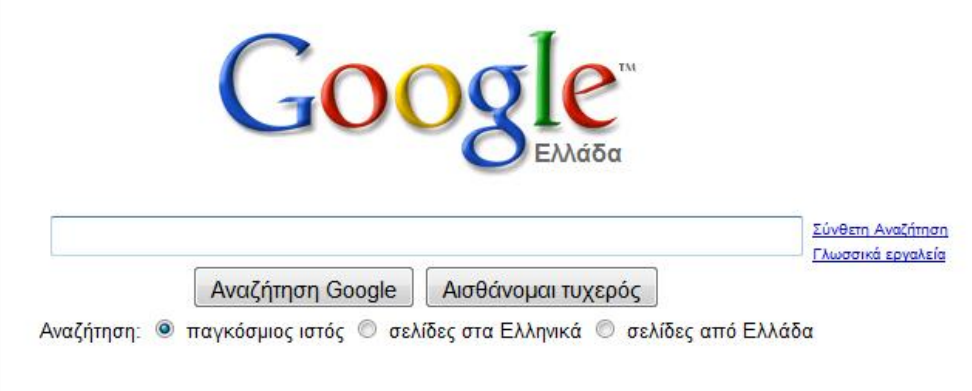
**Εικόνα 1.4**

| **Αρχεία** | **Λέξεις** |
|---|---|
| A1 | το, τόπι, από, σε, Μαρία |
| A2 | τόπι, από, εκεί |
| A3 | το, από, Νίκος |

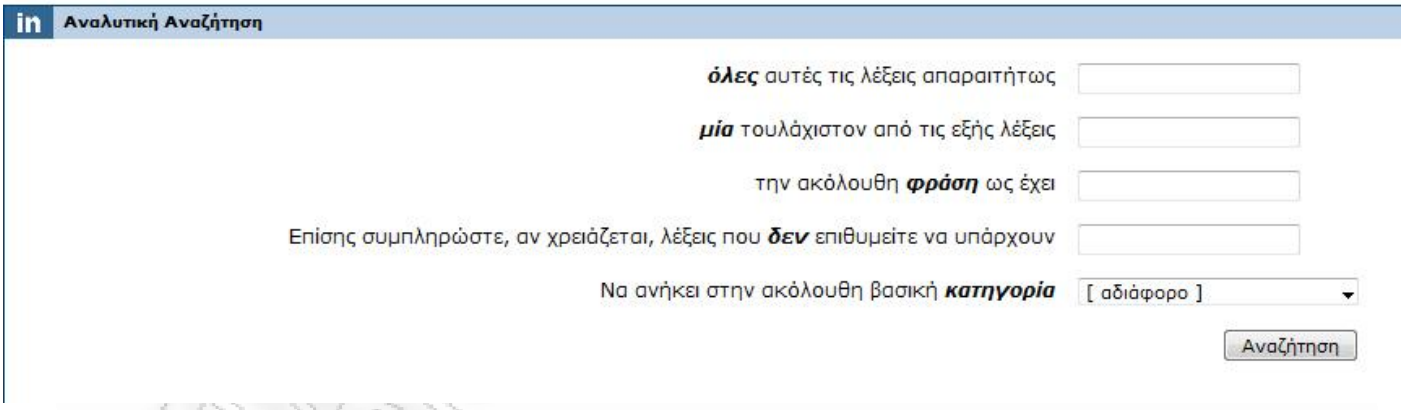
**Εικόνα 1.5**

### 1.3.3 Αναζήτηση με βάση λέξεις κλειδιά (Web Search Query)

Βασικό στοιχείο μιας μηχανής αναζήτησης (όχι μόνο διαδικτυακής) είναι η αλληλεπίδρασή της με τον χρήση. Η αλληλεπίδραση αυτή γίνεται συνήθως με μία φόρμα αναζήτησης, στην οποία ο χρήστης έχει την δυνατότητα εισαγωγής λέξεων/φράσεων, χωρισμένων μεταξύ τους με λογικούς τελεστές, τους οποίους η μηχανή αναζήτησης υποστηρίζει (Εικόνα 1.6). Φυσικά πολλές μηχανές αναζήτησης προσφέρουν διεπαφές με περισσότερο σύνθετες επιλογές, ειδικές για κάθε συγκεκριμένη μηχανή (Εικόνα 1.7).




**Εικόνα 1.6:** απλή αναζήτηση του ιστοτόπου Google




**Εικόνα 1.7:** αναλυτική αναζήτηση του ιστοτόπου in.gr

Τέτοιοι λογικοί τελεστές συνήθως είναι:

- **AND:** υλοποιεί το λογικό «ΚΑΙ».
- **OR:** υλοποιεί το λογική «Η'».
- **NOT:** υλοποιεί την λογική άρνηση.
- **FOLLOWED BY:** ένας όρος θα πρέπει να ακολουθείται από έναν δεύτερο

- **NEAR:** ένας από τους όρους θα πρέπει να βρίσκεται σε κοντική συντακτικά απόσταση από τους υπόλοιπους.

Σε γενικές γραμμές, υπάρχουν 3 κατηγορίες [4] επερωτήσεων που μπορούν να γίνουν:

**Πληροφοριακές επερωτήσεις (Informational queries):** οι επερωτήσεις αυτές αναζητούν ιστοσελίδες με πληροφορίες που αφορούν ένα συγκεκριμένο θέμα (π.χ. αυτοκίνητα). Τα αποτελέσματα φυσικά της μηχανής αναζήτησης συνήθως αφορούν έναν μεγάλο αριθμό σελίδων, οι οποίες αναφέρονται στο θέμα αυτό.

**Επερωτήσεις πλοήγησης (Navigational queries):** οι επερωτήσεις αυτές αφορούν την αναζήτηση πληροφοριών που αφορούν έναν και μόνο ιστότοπο. Για παράδειγμα, εάν ο χρήστης πληκτρολογήσει την λέξη-κλειδί «Lufthansa», αναμένει στα αποτελέσματα να εμφανιστεί το URL της ιστοσελίδας της εν λόγω εταιρείας κι όχι μια πληθώρα εναλλακτικών διαδικτυακών τόπων.

**Επερωτήσεις δοσοληψίας (Transactional queries):** οι επερωτήσεις αυτές αφορούν δοσοληψίες ενός χρήστη στο Διαδίκτυο [5] (π.χ. αγορά ενός εισιτηρίου, εύρεση ενός αρχείου για κατέβασμα, κτλ.). Για παράδειγμα, εάν ένας χρήστης δώσει ως φράση-κλειδί την «Beatles' Lyrics», θέλει θα επιθυμούσε στα αποτελέσματα της αναζήτησης αυτής να εμφανιστούν ιστοσελίδες που να έχουν ως περιεχόμενο τους στίχους των τραγουδιών των Beatles κι όχι ιστοσελίδες που περιέχουν πληροφορίες γενικώς για τους στίχους αυτούς.

Η βελτιστοποίηση των επερωτήσεων που πραγματοποιούνται, όσον αφορά την σημασιολογική τους ανάλυση κυρίως (semantic web search), αποτελεί αντικείμενο μιας ευρείας έρευνας, η οποία όμως ξεφεύγει από τα όρια της συγκεκριμένης μελέτης.

## 1.4 Μηχανές Μετά-Αναζήτησης (Meta-Search engines)

Οι μηχανές μετά-αναζήτησης (μετάφραση του όρου meta-search engines) αποτελούν μια ξεχωριστή κατηγορία των μηχανών αναζήτησης, οι οποίες δεδομένης μιας επερώτησης ενός χρήστη, προωθούν την επερώτηση αυτή σε άλλες μηχανές αναζήτησης ή/και σε άλλες βάσεις δεδομένων και στην συνέχεια ενώνουν τα δεδομένα τα οποία βρήκαν στο τελικό τους αποτέλεσμα. Το κύριο επιχείρημα της δημιουργίας μιας μηχανής μετά-αναζήτησης, είναι ότι το Διαδίκτυο αποτελεί έναν

τεράστιο σε όγκο χώρο αναζήτησης, ο οποίος δεν μπορεί να καλυφθεί από μία και μόνο μηχανή αναζήτησης.

Όπως είναι φυσικό, η συγκεκριμένη προσέγγιση έχει κι αρκετά μειονεκτήματα. Μια μηχανή μετά-αναζήτησης μπορεί να έχει πρόσβαση μόνο στα αποτελέσματα των μηχανών αναζήτησης κι όχι στην εσωτερική τους δομή (δηλαδή στο ευρετήριό τους και στα δεδομένα τα οποία έχουν αποθηκευμένα). Έτσι συνεπώς, η μηχανή μετά-αναζήτησης δεν είναι σε θέση να γνωρίζει ποια από τα δεδομένα στα οποία αποκτά πρόσβαση από διαφορετικές πηγές είναι περισσότερο σχετικά με την επερώτηση και ποια όχι. Ένα σημαντικό θέμα που προκύπτει από την λειτουργία των μηχανών μετά-αναζήτησης, είναι το ότι οι περισσότερες μηχανές αναζήτησης απαιτούν την έκδοση σχετικής άδειας ώστε να χρησιμοποιήσει κάποιος τρίτος τα αποτελέσματά τους. Ειδικά όσον αφορά την μηχανή αναζήτησης Google, η προστασία από την χωρίς άδεια χρήση των αποτελεσμάτων της κατοχυρώνεται νομικά[6]

[6] http://www.google.com/accounts/TOS

# ΚΕΦΑΛΑΙΟ 2
# Η ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ GOOGLE

## 2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα ασχοληθούμε με την λειτουργία της περισσότερο δημοφιλούς διαδικτυακής μηχανής αναζήτησης: της Google. Η Google δημιουργήθηκε το 1996, ως το αποτέλεσμα μιας έρευνας [6], των Larry Page και Sergey Brin, στα πλαίσια της διδακτορικής τους διατριβής, στο πανεπιστήμιου του Stanford της Καλιφόρνια. Ενώ οι περισσότερες μηχανές αναζήτησης ως την περίοδο εκείνη ταξινομούσαν τα αποτελέσματά τους με βάση το πόσες φορές εμφανίζεται ο όρος προς αναζήτηση σε μια ιστοσελίδα, οι δύο μαθηματικοί πρότειναν μία νέα μέθοδο, η οποία παράγει καλύτερα αποτελέσματα: την μέθοδο του **PageRank**. Σύμφωνα με την μέθοδο αυτή, στην βαθμολογία μιας ιστοσελίδας για την εμφάνισή της στα αποτελέσματα, σημαντικό ρόλο παίζει και η σημαντικότητά της, δηλαδή πόσοι σύνδεσμοι από άλλες ιστοσελίδες οδηγούν στην ιστοσελίδα αυτή. Ο διαδικτυακός τόπος, βρισκόταν αρχικά στο πανεπιστήμιο του Stanford και χρησιμοποιούσε το URL http://google.stanford.edu. Η διεύθυνση http://www.google.com/ κατοχυρώθηκε στις 15 Σεπτεμβρίου του 1997[7] και παραμένει ως σήμερα, η κύρια επιλογή για εκατομμύρια χρήστες του Διαδικτύου.

## 2.2 Η αρχιτεκτονική της Google

Η Google σε γενικές γραμμές, ακολουθεί την αρχιτεκτονική που παρουσιάστηκε στο Κεφάλαιο 1. Η Google εκτελείται σε ένα κατανεμημένο περιβάλλον, από χιλιάδες (μικρού κόστους) υπολογιστές και συνεπώς μπορεί να εκτελέσει γρήγορα παράλληλους υπολογισμούς. Με τον τρόπο αυτό, μπορεί να επεξεργαστεί παράλληλα έναν πολύ μεγάλο όγκο δεδομένων, αφού οι υπολογισμοί εκτελούνται την ίδια χρονική στιγμή. Το λογισμικό της Google αποτελείται από τρία κύρια μέρη:

- **Το Googlebot**, έναν web crawler που βρίσκει και αποθηκεύει ιστοσελίδες.

[7] http://whois.dnsstuff.com/tools/whois.ch?ip=google.com

- **Το λογισμικό ευρετηριοποίησης (google indexer)**, που ταξινομεί κάθε λέξη σε κάθε σελίδα και αποθηκεύει τα αποτελέσματα σε ένα ευρετήριο λέξεων σε μία τεράστια σε μέγεθος βάση δεδομένων.
- **Τον επεξεργαστή επερωτήσεων (query processor)**, που συγκρίνει το την επερώτηση του χρήστη, με τα αποθηκευμένα στο ευρετήριο δεδομένα και δίνει ως αποτέλεσμα τα έγγραφα εκείνα τα οποία κρίνει ότι είναι περισσότερο σχετικά, αφού προηγουμένως τα ταξινομήσει.

Θα δούμε συνοπτικά τα μέρη αυτά παρακάτω.

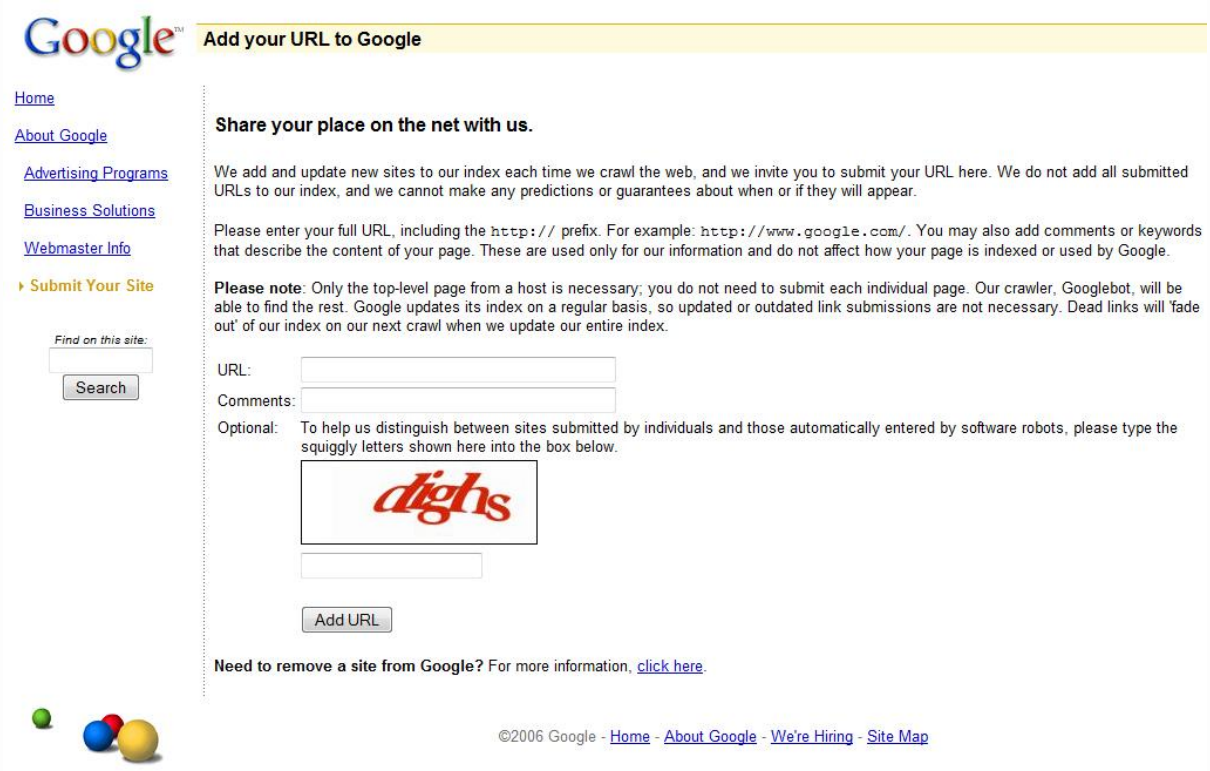
### 2.2.1 Googlebot

Το Googlebot είναι ο web crawler της Google. Χρησιμοποιείται για την εύρεση και την αποθήκευση ιστοσελίδων και αρχείων από τον παγκόσμιο ιστό, τα οποία στην συνέχεια παραδίδει στον Google indexer. Φυσικά, για να μην καταναλώνει πολλούς από τους πόρους μιας ιστοσελίδας, το Googlebot επισκέπτεται τις ιστοσελίδες με πολύ πιο αργό ρυθμό από ότι μπορεί στην πραγματικότητα να τις επισκεφθεί. Το Googlebot δίνει την δυνατότητα στον διαχειριστή μιας ιστοσελίδας, να απαγορεύσει την αποθήκευσή της στην Google, ενσωματώνοντας στον HTML κώδικά της, την εξής εντολή: <meta name="Googlebot" content="nofollow" />. Η Google δίνει επίσης την δυνατότητα, της εισαγωγής μιας ιστοσελίδας, μέσω μιας φόρμας, στην διεύθυνση http://www.google.com/addurl/?continue=/addurl (Εικόνα 2.1). Δυστυχώς, πολλοί spammers έχουν βρει τρόπους, έτσι ώστε να εισάγουν εκατομμύρια διευθύνσεις, αυτοματοποιημένα, στην παραπάνω διεύθυνση, με στόχο την διαφημιστική ή κάποιου άλλου είδους προπαγάνδα. Η Google ελέγχει όλα τα URLs τα οποία εισάγονται στην παραπάνω διεύθυνση και απομακρύνει όσα υποπτεύεται ότι αποτελούν spam.

Όταν το Googlebot επισκέπτεται μια ιστοσελίδα, βρίσκει όλους τους συνδέσμους που αναφέρονται στην ιστοσελίδα αυτή και τους προσθέτει σε μία ουρά αναμονής για μετέπειτα εξέταση. Φυσικά το Googlebot λαμβάνει υπ'όψη του όλα όσα αναφέρθηκαν στο Κεφάλαιο 1, περί λειτουργίας ενός web crawler.

Παρά το γεγονός ότι η λειτουργία του είναι σχετικά απλή, το Googlebot είναι προγραμματισμένο έτσι, ώστε να ξεπερνά πολλές προκλήσεις και δυσκολίες. Για παράδειγμα υπάρχει περίπτωση η ίδια σελίδα να εμφανίζεται πολλές φορές μέσα στην ουρά προτεραιότητας για μελλοντική εξέταση του Googlebot ή τα δεδομένα της να

έχουν ήδη ευρετηριοποιηθεί στο παρελθόν. Ένα πρόβλημα που επίσης δημιουργείται, είναι κάθε πότε το Googlebot θα πρέπει να επισκέπτεται ιστοσελίδες που βρίσκονται ήδη ευρετηριοποιημένες, έτσι ώστε να ελέγξει τυχόν αλλαγές. Ο χρόνος επανεξέτασης θα πρέπει να πρέπει να μην είναι πολύ συχνός, έτσι ώστε να μην ελεγχθούν ξανά ιστοσελίδες που δεν έχουν αλλαγές αλλά συνάμα κι όχι πολύ μεγάλος, αφού οι ιστοσελίδες θα πρέπει να αποθηκεύονται με τις απαιραίτητες ανανεώσεις (up-to-date results).




**Εικόνα 2.1:** http://www.google.com/addurl/?continue=/addurl

## 2.2.3 Λογισμικό ευρετηριοποίησης (Google Indexer)

Μετά την εξέταση μιας σελίδας, το Googlebot μεταβιβάζει το πλήρες κείμενο αυτής στο λογισμικό ευρετηριοποίησης της Google (Google Indexer). Ολόκληρη η σελίδα αποθηκεύεται σε μία βάση δεδομένων της Google (Doc Servers). Το ευρετήριο το οποίο δημιουργείται, για την αποδοτική αναζήτηση των σελίδων, είναι ταξινομημένο αλφαβητικά με βάση την κάθε λέξη. Για κάθε όρο αποθηκεύεται μία λίστα με όλα τα έγγραφα στα οποία η λέξη εμφανίζεται (forward index), καθώς και η θέση της μέσα

στο κείμενο. Η δομή αυτή δεδομένων, προσφέρει γρήγορη πρόσβαση στα αποθηκευμένα έγγραφα.

Για την βελτιστοποίηση της απόδοσης της αναζήτησης, η Google αγνοεί (δεν ευρετηριοποιεί) κοινές λέξεις (οι οποίες καλούνται stop words), όπως για παράδειγμα τις λέξεις "as", "the", "is", "in", "of", "how", "why", όπως επίσης και συγκεκριμένους μονοψήφιους αριθμούς και γράμματα. Οι λέξεις αυτές, είναι τόσο συνηθισμένες, που παρουσιάζονται σε σχεδόν κάθε κείμενο και δεν προσφέρουν καμία ουσιαστική παραπάνω πληροφορία στην αναζήτηση μιας συγκεκριμένης φράσης που τις περιέχει. Φυσικά εκτός των άλλων, ο Google Indexer αγνοεί και πολλά σημεία στίξης, καθώς και τους κενούς χαρακτήρες, ενώ επίσης μετατρέπει όλα τα γράμματα σε μικρά κι όχι σε κεφαλαία, έτσι ώστε να εντοπίσει την μία λέξη, ακόμα κι αν είναι γραμμένη διαφορετικά.

### 2.2.3 Επεξεργαστής Επερωτήσεων (Google's Query Processor)

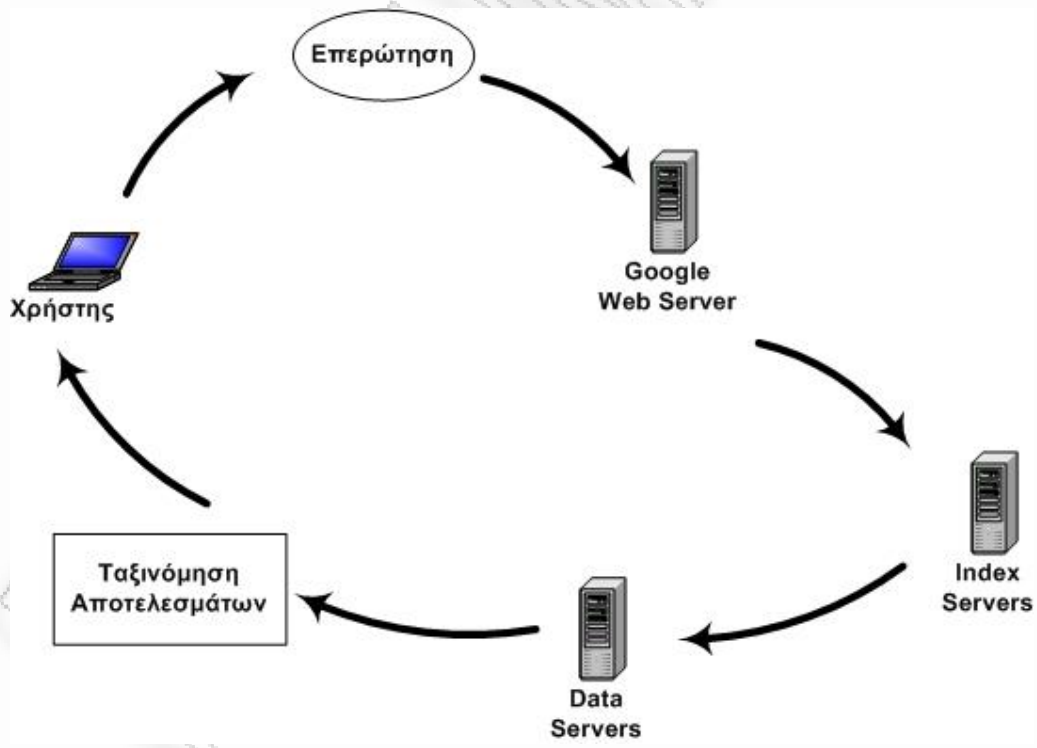
Ο επεξεργαστής των επερωτήσεων, αποτελείται από πολλά μέρη, συμπεριλαμβανομένων της γραφικής διεπαφής (με την μορφή μιας διαδικτυακής φόρμας – search box), ενός λογισμικού που αντιστοιχεί την επερώτηση σε σχετικά έγγραφα, καθώς και ενός λογισμικού που ταξινομεί τα αποτελέσματα.

Αρχικά, ο επεξεργαστής των επερωτήσεων προσπαθεί να «αντιληφθεί» την επερώτηση που έχει εισάγει ο χρήστης, η οποία ενδεχομένως να περιλαμβάνει πλην των λέξεων και ορισμένους τελεστές που έχουν ειδική σημασία για την μηχανή αναζήτησης. Στην συνέχεια, με την χρήση του Google Indexer, εντοπίζει όλες εκείνες τις ιστοσελίδες, στις οποίες οι λέξεις αυτές εμφανίζονται και αποκτά πρόσβαση στα δεδομένα τους. Το σημαντικότερο σημείο στην όλη διαδικασία, είναι η σειρά εμφάνισης των αποτελεσμάτων και είναι το στοιχείο αυτό που έχει κάνει την μηχανή αναζήτησης Google την δημοφιλέστερη παγκοσμίως. Ο επεξεργαστής των επερωτήσεων, ταξινομεί της σελίδας σε φθίνουσα σειρά ανάλογα με το PageRank τους. Έτσι, μια σελίδα με υψηλότερη βαθμολογία είναι περισσότερο σημαντική από μία με μικρότερη και συνεπώς θα εμφανιστεί πιο πάνω στα αποτελέσματα. Ο επεξεργαστής επερωτήσεων λαμβάνει υπ'όψην του παραπάνω από 100 παραμέτρους για τον υπολογισμό της βαθμολογίας μιας ιστοσελίδας. Ενδεικτικά, μερικές από τις παραμέτρους αυτές είναι η δημοτικότητα της ιστοσελίδας (δηλαδή σε πόσες άλλες

ιστοσελίδες εμφανίζεται ως σύνδεσμός), η θέση και το μέγεθος των όρων αναζήτησης μέσα στην σελίδα, το πόσο κοντά μεταξύ τους είναι οι όροι, κτλ.

Εκτός των άλλων, το λογισμικό επερωτήσεων εφαρμόζει και τεχνικές μηχανικής μάθησης (machine learning), έτσι ώστε να βελτιώσει την απόδοσή του, ύστερα από κάθε επερώτηση, ανακαλύπτοντας και «μαθαίνοντας» συσχετίσεις μεταξύ των αποθηκευμένων δεδομένων. Για παράδειγμα, η μηχανή αναζήτησης χρησιμοποιεί μια τέτοια τεχνική, στην περίπτωση όπου κάποιος ή κάποιοι από τους όρους, είναι γραμμένοι με ορθογραφικά λάθη. Αν κάτι τέτοιο ισχύει, η Google προτείνει εναλλακτικές επερωτήσεις, διορθώνοντας τις λάθος ορθογραφικά λέξεις (spelling-correcting system).

Η μηχανή αναζήτησης Google, δεν περιορίζεται μόνο στην αναζήτηση απλών λέξεων μέσα σε ιστοσελίδες. Εκτός των άλλων προσφέρει αναζήτηση εικόνων[8], αναζήτηση επιστημονικών εγγραφών[9], αναζήτηση ολόκληρων φράσεων, κτλ.
Μπορούμε να συνοψίσουμε την αρχιτεκτονική επεξεργασίας μιας επερώτησης, στην Εικόνα 2.2.




**Εικόνα 2.2:** η αρχιτεκτονική επεξεργασίας μιας επερώτησης

[8] http://images.google.com/
[9] http://scholar.google.com/

## 2.3 Το υλικό της μηχανής αναζήτησης Google

Το υλικό της μηχανής αναζήτησης Google παραμένει μυστικό στο ευρύ κοινό. Σύμφωνα με πολλές μαρτυρίες [7], η μηχανή αναζήτησης Google έχει προβεί σε εξαιρετικά μεγάλα μέτρα ασφαλείας, έτσι ώστε να μην αποκαλυφθούν χαρακτηριστικά της αρχιτεκτονικής της στο ευρύ κοινό. Οι εγκαταστάσεις όπου στεγάζεται το υλικό της, δεν είναι ανοιχτές για επίσκεψη.

Παρά το γεγονός αυτό, έχουν γίνει πολλές εκτιμήσεις για τις υπολογιστικές δυνατότητες του υλικού που χρησιμοποιεί η μηχανή. Σύμφωνα με εκτιμήσεις του 2005 [8], η μηχανή αναζήτησης χρησιμοποιούσε 200.000 υπολογιστές, ενώ σύμφωνα με εκτιμήσεις του 2006[10], ο αριθμός αυτός ανέρχεται στις 450.000.

Όσον αφορά τον Διαδικτυακό Διακοσμητή (Web Server) που χρησιμοιεί η μηχανή αναζήτησης, η Google έχει υλοποιήσει ένα δικό της λογισμικό, που ονομάζεται Google Web Server (GWS). Η Google διατηρεί σκόπιμα μυστικές τις προδιαγραφές και την υλοποίηση του συγκεκριμένου λογισμικού και η μόνη πληροφορία η οποία έχει δόσει είναι ότι εκτελείται σε λειτουργικό σύστημα Linux[11]. Υπάρχουν βέβαια ορισμένες ενδείξεις που υποδεικνύουν ότι ο GWS είναι μια τροποποιημένη έκδοση του δημοφιλούς Web Server Apache[12].

## 2.4 PageRank

### 2.4.1 Βασικές Έννοιες

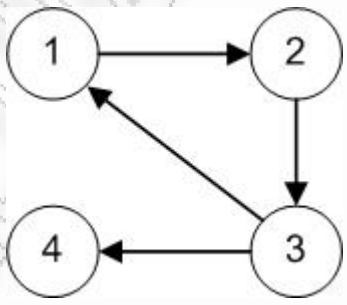
Στην συγκεκριμένη ενότητα, θα αναλύσουμε το σύστημα της βαθμολόγησης των ιστοσελίδων που πάνω στο οποίο βασίζεται η μηχανή αναζήτησης Google. Ο αλγόριθμος του PageRank προσδίδει μια βαθμολογία σε κάθε μία από τις δισεκκατομύρια ιστοσελίδες τις οποίες έχει αποθηκεύσει το Googlebot. Ο αλγόριθμος προσπαθεί να μοντελοποιήσει την συμπεριφορά ενός *ιδανικού χρήστη του Διαδικτύου* [6]. Σύμφωνα με την συμπεριφορά αυτή, ο χρήστης διαλέγει τυχαία έναν σύνδεσμο από μια ιστοσελίδα σε μία άλλη. Ο χρήστης συνεχίζει να επιλέγει συνδέσμους με τον ίδιο τρόπο (τυχαία) εως ότου επιλέξει κάποια ιστοσελίδα, επειδή το αποφάσισε κι όχι

[10] http://www.baselinemag.com/c/a/Infrastructure/How-Google-Works-1/
[11] http://news.cnet.com/8301-1001_3-10079685-92.html?tag=mncol;title
[12] http://googlesystem.blogspot.com/2007/09/googles-server-names.html

τυχαία. Η επιλογή αυτή δεν θα πρέπει να επηρεάζεται από τις προηγούμενες επιλογές, οι οποίες, θεωρητικά, έγιναν με τυχαίο τρόπο. Έτσι λοιπόν, η βαθμολογία μιας ιστοσελίδας με την χρήση του αλγορίθμου PageRank, αντιπροσωπεύει την πιθανότητα ο χρήστης να επέλεξε εθελημένα την ιστοσελίδα αυτή. Φυσικά, η συμπεριφορά αυτή δεν είναι εύκολο να μοντελοποιηθεί στα στενά όρια ενός αλγορίθμου που εκτελείται σε κάποιον υπολογιστή. Στις παρακάτω υπό-ενότητες θα δούμε σε αρκετά αφηρημένο επίπεδο την λογική του αλγορίθμου και θα αναλύσουμε συνοπτικά τα βασικά του χαρακτηριστικά.

### 2.4.2 Το Διαδίκτυο ως ένας κατευθυνόμενος γράφος

Για να μοντελοποιήσει την δραστηριότητα ενός τυχαίου ιδανικού χρήστη του Διαδικτύου, ο αλγόριθμος PageRank αναπαριστά τους συνδέσμους μεταξύ των ιστοσελίδων του Διαδικτύου σαν έναν *κατευθυνόμενο γράφο (directed graph)*. Οι ιστοσελίδες αναπαριστούν τους κόμβους του γράφου αυτού, ενώ οι σύνδεσμοι (web links) από μια ιστοσελίδα σε μία άλλη αναπαριστούν τις κατευθυνόμενες ακμές του. Παρά το γεγονός ότι ο κατευθυνόμενος γράφος του Διαδικτύου είναι υπερβολικά μεγάλος σε μέγεθος, ο αλγόριθμος του PageRank μπορεί να εφαρμοστεί σε οποιοδήποτε κατευθυνόμενο γράφο, οποιουδήποτε μεγέθους. Ένα παράδειγμα ενός τέτοιου γράφου με 4 κόμβους παρουσιάζεται στην Εικόνα 2.3.




**Εικόνα 2.3:** παράδειγμα ενός κατευθυνόμενου διαδικτυακού γράφου

Στο παραπάνω σχήμα, αναπαρίστανται 4 ιστοσελίδες και οι εξής σύνδεσμοι (web links): σύνδεσμός από την ιστοσελίδα 1 στην 2, σύνδεσμός από την ιστοσελίδα 2 στην 3 και σύνδεσμοι από την ιστοσελίδα 3, στις 1 και 4 αντίστοιχα. Όπως βλέπουμε, η ιστοσελίδα 4 δεν περιέχει κανέναν εξερχόμενο σύνδεσμο.

### 2.4.3 Πίνακας διαδικτυακών υπερσυνδέσμων

Η διαδικασία για την απόδοση βαθμολογίας σε κάθε έναν από τους $n$ στον αριθμό κόμβους (ιστοσελίδες) του κατευθυνόμενου γράφου, αρχίζει με την απεικόνιση του γράφου ως έναν πίνακα διάστασης $nxn$, ο οποίος καλείται πίνακας διαδικτυακών υπερσυνδέσμων (hyperlink matrix) και συμβολίζεται με $H$. Ας υποθέσουμε ότι μια ιστοσελίδα $i$ περιέχει $l_i \geq 1$ συνδέσμους σε άλλες ιστοσελίδες. Ας υποθέσουμε επίσης ότι η ιστοσελίδα $i$ περιέχει $k$ στον αριθμό συνδέσμους σε μια συγκεκριμένη ιστοσελίδα $j$. Αν κάτι τέτοιο ισχύει, τότε το στοιχείο που βρίσκεται στην γραμμή $i$ του πίνακα και στην στήλη $j$ παίρνει την τιμή $H_{ij} = \frac{k}{l_i}$. Σε περίπτωση που η ιστοσελίδα $i$ δεν περιέχει κανένα σύνδεσμο σε μια ιστοσελίδα $j$, τότε το αντίστοιχο στοιχείο $H_{ij}$ παίρνει την τιμή 0. Για παράδειγμα, ο πίνακας $H$ για τον κατευθυνόμενο γράφο της Εικόνας 2.3, είναι ο παρακάτω:

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

Ο κόμβος 4 είναι ένας κόμβος αδιεξόδου (dangling node) επειδή δεν περιέχει συνδέσμους σε άλλους ιστοσελίδες. Αυτό έχει ως αποτέλεσμα, όλες οι εγγραφές στην γραμμή 4 του παραπάνω παραδείγματος να είναι μηδέν. Αυτό σημαίνει πρακτικά ότι η πιθανότητα ένας τυχαίος χρήστης του Διαδικτύου θα κινηθεί από τον κόμβο 4 σε έναν οποιοδήποτε άλλον κόμβο, ακολουθώντας έναν σύνδεσμο, είναι μηδέν.

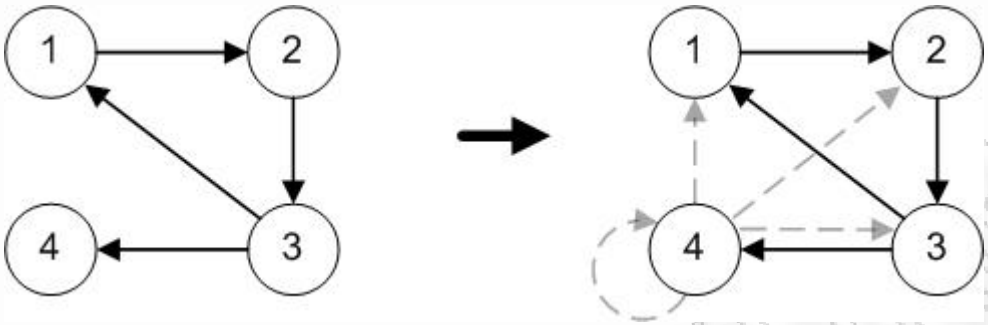
Η πλειοψηφία των ιστοσελίδων του Διαδικτύου αποτελούνται από τέτοιους κόμβους (στην υπόθεση αυτή συμπεριλαμβάνουμε και κόμβους που αποτελούν αρχεία, όπως εικόνες, αρχεία PDF, κτλ.), οπότε υπάρχουν πολλές μηδενικές γραμμές στον συνολικό πίνακα διαδικτυακών υπερσυνδέσμων του Διαδικτύου. Όταν ένας τυχαίος χρήστης βρεθεί σε μία τέτοια ιστοσελίδα, τότε για να συνεχίσει την περιήγησή του, θα πρέπει *ενεργά* να μεταβεί σε μία νέα, πληκτρολογώντας το URL της στον περιηγητή του. Επειδή ο πίνακας $H$, με την παραπάνω μορφή, δεν απεικονίζει την μετάβαση από έναν κόμβο αδιεξόδου σε έναν άλλον, η συμπεριφορά

ενός διαδικτυακού χρήστη δεν μπορεί να μοντελοποιηθεί πλήρως με τον παραπάνω πίνακα και μόνο.

### 2.4.4 Διόρθωση του προβλήματος των κόμβων αδιεξόδου

Για την μοντελοποίηση των επιλογών ενός τυχαίου χρήστη του Διαδικτύου, στην περίπτωση όπου βρεθεί σε έναν κόμβο αδιεξόδου, υπάρχουν αρκετές επιλογές. Η μηχανή αναζήτησης Google, δεν έχει κοινοποιήσει ακόμη ποια από τις επιλογές αυτές χρησιμοποιεί. Μία επιλογή είναι η ανάθεση σε κάθε στοιχείο j της γραμμής ενός κόμβου αδιεξόδου μίας πιθανότητας, έστω $w_j$, έτσι ώστε το άθροισμα των στοιχείων της γραμμής να ισούται με 1. Αυτό στην ουσία αναπαριστά το ότι η μετάβαση από τον κόμβο αδιεξόδου σε έναν οποιοδήποτε άλλον κόμβο είναι ισοπίθανη. Ο καινούριος πίνακας είναι ο $S = H + dw$, όπου $d$ είναι ένας πίνακας στήλη, που αντιστοιχεί σε έναν κόμβο αδιεξόδου (δηλαδή $d_i = 1$, εάν $l_i = 0$ ενώ $d_i = 0$ σε διαφορετική περίπτωση) και $w$ το διάνυσμα $[w_1\ w_2\ \ldots\ w_n]$, με $w_j \geq 0$, για κάθε $1 \leq j \leq n$ και $\sum_{j=1}^{n} w_j = 1$. Η πιο δημοφιλής επιλογή για τις πιθανότητες αυτές των μεταβάσεων, είναι η ομοιόμορφη κατανομή, όπου όλες οι μεταβάσεις είναι ισοπίθανες. Στην ουσία δηλαδή το διάνυσμα $w$ είναι το $\left[\frac{1}{n}\ \frac{1}{n}\ \ldots\ \frac{1}{n}\right]$. Για να γίνουν περισσότερο κατανοητά τα παραπάνω, θα εφαρμόσουμε την τεχνική αυτή στο παράδειγμα των προηγούμενων υπό-ενοτήτων.

Στο παραπάνω παράδειγμα, έχουμε n = 4 διαφορετικές ιστοσελίδες κι έτσι το διάνυσμα $w$ θα είναι το $\left[\frac{1}{4}\ \frac{1}{4}\ \ldots\ \frac{1}{4}\right]$. Με την αλλαγή της γραμμής που αφορά τον κόμβο αδιεξόδου 4, ο διαδικτυακός γράφος παίρνει την μορφή που απεικονίζεται στην Εικόνα 2.4. Παρατηρούμε ότι στον γράφο αυτό υπάρχει ανακύκλωση, όσον αφορά τον κόμβο αδιεξόδου (δηλαδή που ξεκινά και καταλήγει στον κόμβο 4). Η ακμή αυτή στην ουσία μοντελοποιεί την δυνατότητα του χρήστη, να εκτελέσει την λειτουργία ‘Refresh’ στον περιηγητή που χρησιμοποιεί, όσο βρίσκεται σε έναν κόμβο αδιεξόδου.




**Εικόνα 2.4:** ο νέος γράφος, μετά την διόρθωση του κόμβου αδιεξόδου 4

Ο νέος πίνακας $S = H + dw$, είναι:

$$S = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Παρά το γεγονός ότι η παραπάνω προσέγγιση διορθώνει ως έναν βαθμό το πρόβλημα του κόμβου αδιεξόδου, δεν το επιλύει πλήρως, αλλά φανερώνει μία ακόμα αδυναμία του διαδικτυακού γράφου. Παρά το γεγονός ότι δεν υπάρχει σύνδεσμος που να οδηγεί από την ιστοσελίδα 2, στην ιστοσελίδα 1, ένας τυχαίος χρήστης μπορεί να πραγματοποιήσει αυτή την μετάβαση, εισάγοντας το URL της ιστοσελίδας 1 στον περιηγητή του, όσο βρίσκεται στην ιστοσελίδα 2. Στην ουσία, το πρόβλημα το οποίο δημιουργείται λόγω του κόμβου αδιεξόδου, θα μπορούσε να δημιουργηθεί από έναν οποιοδήποτε κόμβο του γράφου. Φυσικά η πιθανότητα να κάνει κάτι τέτοιο, είναι μικρότερη από το να ακολουθήσει έναν από τους συνδέσμους της ιστοσελίδας, αλλά είναι υπαρκτή. Η μηχανή αναζήτησης Google χρησιμοποιεί επιπλέον προσεγγίσεις για την διόρθωση της παραπάνω ατέλειας, έτσι ώστε η συμπεριφορά ενός τυχαίου διαδικτυακού χρήστη να μοντελοποιηθεί όσο το δυνατόν καλύτερα.

### 2.4.5 Το Google Matrix

Για να μοντελοποιήσει την συνολική συμπεριφορά ενός τυχαίου χρήστη του Διαδικτύου (σύμφωνα και με τις παραπάνω παρατηρήσεις), η μηχανή αναζήτησης Google χρησιμοποιεί τον πίνακα (matrix) $G = aS + (1 - a)\mathbb{I}u$, όπου $0 \leq a < 1$ μία πραγματική τιμή, $\mathbb{I}$ το μοναδιαίο διάνυσμα στήλη (με όλα του τα στοιχεία δηλαδή 1),και $u$ το διάνυσμα που φέρει τις πιθανότητες μετάβασης από έναν κόμβο σε έναν άλλον για τον τυχαίο χρήστη. Το $u$ καλείται διάνυσμα εξατομίκευσης (personalization vector), ενώ το $a$ παράγοντας άμβλυνσης (damping factor). Ο παράγοντας αυτός μοντελοποιεί την τυχαία μετάβαση του χρήστη σε μία ιστοσελίδα διαφορετική από τις μεταβάσεις που υπαγορεύονται στον πίνακα S. Στην ουσία υποδηλώνει ότι η πιθανότητα μετάβασης από μία ιστοσελίδα σε μία άλλη, μέσω των υπαρκτών συνδέσμων είναι $a$, ενώ η μετάβαση σε κάποια άλλη σελίδα, μέσω πληκτρολόγησης ενός URL κι όχι μέσω συνδέσμων είναι $1 - a$. Η πλειοψηφία των πειραμάτων που πραγματοποιήθηκαν από τους Brin και Page [6] χρησιμοποίησαν για $a$ την τιμή 0.85 και για $u$ το διάνυσμα $[\frac{1}{n}\ \frac{1}{n}\ \ldots\ \frac{1}{n}]$. Οι πιο συνηθισμένες τιμές για τον παράγοντα $a$ κυμαίνονται μεταξύ 0.85 και 0.99 στις περισσότερες ερευνητικές δημοσιεύσεις.

Αναθέτοντας ως τιμή του διανύσματος $u$ το διάνυσμα που ακολουθεί την ομοιόμορφη κατανομή (δηλαδή όπου κάθε μετάβαση είναι ισοπίθανη με λόγο $\frac{1}{n}$), ένας τυχαίος χρήστης του Διαδικτύου θα μεταβεί σε κάποια σελίδα με ίση πιθανότητα. Η θεώρηση αυτή, κάνει την μέθοδο PageRank ιδιαίτερα ευάλωτη στην τεχνική του link spamming, όπως θα δούμε και παρακάτω, οπότε η Google πλέον δεν χρησιμοποιεί το ομοιόμορφο διάνυσμα. Με λίγα λόγια, αν κάποιος μπορούσε να αυξήσει τον αριθμό των διαφορετικών ιστοσελίδων που οδηγούν στην ιστοσελίδα του, τότε θα είχε και καλύτερες βασθμολογίες από την μέθοδο του PageRank. Το 2004, οι Gyongyi, Garcia-Molina και Pedersen πρότειναν τον αλγόριθμο του TrustRank [9] , ο οποίος δημιουργεί ένα διάνυσμα εξατομίκευσης, το οποίο επιλύει το συγκεκριμένο πρόβλημα. Η Google έχει αγοράσει τα δικαιώματα του συγκεκριμένου αλγόριθμου και ήδη τον έχει θέσει σε εφαρμογή.

### 2.4.6 Υπολογισμός του PageRank

Ο υπολογισμός της βαθμολογίας (PageRank) για κάθε ιστοσελίδα, γίνεται ως εξής:

Υπάρχει πάντοτε ένας πίνακας γραμμή $\pi$, για κάθε πίνακα $G$ για τον οποίο ισχύει ότι:

$$\pi G = \pi$$

Το $i$-οστό στοιχείο του πίνακα αυτού είναι και η βαθμολογία της $i$-οστής ιστοσελίδας.

Η απόδειξη ύπαρξης ενός τέτοιου πίνακα $\pi$ για κάθε πίνακα $G$, είναι καθαρά μαθηματική και η θεωρία που την πλαισιώνει υπερβαίνει τα όρια της συγκεκριμένης μελέτης[13]. Το στοιχείο το οποίο θα πρέπει να τονισθεί είναι ότι το άθροισμα των στοιχείων του διανύσματος $\pi$ είναι 1 [6] (είναι στην ουσία δηλαδή ένας πίνακας πιθανοτήτων). Αυτό που παρουσιάζει ιδιαίτερο ενδιαφέρον, είναι ο βαθμός που ο παράγοντας άμβλυνσης $a$ και το διάνυσμα εξατομίκευσης $u$ συμβάλλουν στον υπολογισμό της βαθμολογίας. Για να διαπιστώσουμε τον βαθμό αυτό, ας δούμε μερικά παραδείγματα με πραγματικές τιμές, για τον αρχικό πίνακα $S$ των παραπάνω υπό-ενοτήτων, τα οποία παρουσιάζονται στον Πίνακα 2.1 της επόμενης σελίδας.

Στον Πίνακα 2.1 παρουσιάζονται 4 διαφορετικοί πίνακες $G$ και οι αντίστοιχες βαθμολογίες των ιστοσελίδων. Τα παραδείγματα αποδεικνύουν ότι το διάνυσμα εξατομίκευσης επιδρά στον υπολογισμό του πίνακα $G$ σε μεγαλύτερο βαθμό, όταν ο παράγοντας άμβλυνσης είναι μικρός. Για παράδειγμα, με σταθερό $a$ = 0.85 και δύο διαφορετικά εξατομικευμένα διανύσματα (το διάνυσμα ομοιόμορφης κατανομής στο πρώτο παράδειγμα και ένα διαφορετικό διάνυσμα στο δεύτερο), λαμβάνουμε εντελώς διαφορετικές βαθμολογίες για τις 4 αρχικές ιστοσελίδες. Αντίθετα, αν ο παράγοντας άμβλυνσης είναι μεγάλος, τότε το διάνυσμα εξατομίκευσης επιδρά λιγότερο στον υπολογισμό των βαθμολογιών. Αυτό μπορεί να επαληθευθεί από τα παραδείγματα 3 και 4, όπου με με σταθερό $a$ = 0.95 και δύο διαφορετικά εξατομικευμένα διανύσματα λαμβάνουμε σχεδόν τις ίδιες βαθμολογίες. Συμπεραίνουμε λοιπόν, ο βαθμός επίδρασης του διανύσματος εξατομίκευσης είναι αντιστρόφως ανάλογως του

[13] Στην ουσία ο πίνακας $\pi$ αποτελεί το ιδιοδιάνυσμα του πίνακα $G$.

παράγοντα άμβλυνσης. Η ακριβής τιμή του παράγοντα άμβλυνσης που χρησιμοποιεί η Google, δεν έχει δοθεί στην δημοσιότητα.

| | **Παράγοντας άμβλυνσης ($a$)** | **Διάνυσμα εξατομίκευσης ($u$)** | **Google Matrix ($G$)** | **Πίνακας βαθμολογιών ($\pi$)** | **Ταξινόμηση των ιστοσελίδων (1 = μεγαλύτερη βαθμολογία)** |
|---|---|---|---|---|---|
| Παράδειγμα 1 | 0.85 | $\left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}\right)$ | $\begin{pmatrix} \frac{3}{80} & \frac{71}{80} & \frac{3}{80} & \frac{3}{80} \\ \frac{3}{80} & \frac{3}{80} & \frac{71}{80} & \frac{3}{80} \\ \frac{37}{80} & \frac{3}{80} & \frac{3}{80} & \frac{37}{80} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$ | $(0.21 \quad 0.26 \quad 0.31 \quad 0.21)$ | (3 2 1 3) |
| Παράδειγμα 2 | 0.85 | $(1 \quad 0 \quad 0 \quad 0)$ | $\begin{pmatrix} \frac{3}{20} & \frac{17}{20} & 0 & 0 \\ \frac{3}{20} & 0 & \frac{17}{20} & 0 \\ \frac{23}{40} & 0 & 0 & \frac{17}{40} \\ \frac{29}{80} & \frac{17}{80} & \frac{17}{80} & \frac{17}{80} \end{pmatrix}$ | $(0.30 \quad 0.28 \quad 0.27 \quad 0.15)$ | (1 2 3 4) |
| Παράδειγμα 3 | 0.95 | $\left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}\right)$ | $\begin{pmatrix} \frac{1}{80} & \frac{77}{80} & \frac{1}{80} & \frac{1}{80} \\ \frac{1}{80} & \frac{1}{80} & \frac{77}{80} & \frac{1}{80} \\ \frac{39}{80} & \frac{1}{80} & \frac{1}{80} & \frac{39}{80} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$ | $(0.21 \quad 0.26 \quad 0.31 \quad 0.21)$ | (3 2 1 3) |
| Παράδειγμα 4 | 0.95 | $(1 \quad 0 \quad 0 \quad 0)$ | $\begin{pmatrix} \frac{1}{20} & \frac{19}{20} & 0 & 0 \\ \frac{1}{20} & 0 & \frac{19}{20} & 0 \\ \frac{21}{40} & 0 & 0 & \frac{19}{40} \\ \frac{23}{80} & \frac{19}{80} & \frac{19}{80} & \frac{19}{80} \end{pmatrix}$ | $(0.24 \quad 0.27 \quad 0.30 \quad 0.19)$ | (3 2 1 4) |

**Πίνακας 2.1:** πειραματικοί υπολογισμοί του PageRank, συναρτήσει του παράγοντα άμβλυνσης $a$ και του διάνυσματος εξατομίκευσης $u$.

Για μικρούς σε διάσταση πίνακες, όπως αυτοί που παρουσιάζονται στον Πίνακα 2.1, ο υπολογισμός του PageRank ως η λύση της εξίσωσης $\pi G = \pi$ είναι εύκολος και μη απαιτητικός όσον αφορά το υπολογιστικό κόστος. Ο πίνακας $G$ για τις ιστοσελίδες που έχει αποθηκεύσει η μηχανή αναζήτησης Google, έχει περισσότερες από 25 δισεκατομμύρια γραμμές και στήλες, οπότε ο υπολογισμός των βαθμολογιών για κάθε ιστοσελίδα, ως λύση της παραπάνω εξίσωσης, θα απαιτούσε υπερβολικά μεγάλη υπολογιστική ισχύ και υπερβολικά πολύ χρόνο, κάτι το οποίο είναι μη ρεαλιστικό. Για τον υπολογισμό του διανύσματος με τις βαθμολογίες σε ρεαλιστικά πλαίσια χρόνου, η Google χρησιμοποιεί μεθόδους που προσεγγίζουν το διάνυσμα $\pi$, με σχετικά μεγάλη ακρίβεια.

Η πιο γνωστή από αυτές τις μεθόδους, είναι η επαναληπτική μέθοδος με την χρήση δυνάμεων. Σύμφωνα με την μέθοδο αυτή, δίνεται αρχικά στο διάνυσμα $\pi^{(0)}$ μια αρχική τιμή, π.χ. $\pi^{(0)} = u$. Στην συνέχεια, υπολογίζεται συνεχώς η παρακάτω σχέση:

$$\pi^{(k)} = \pi^{(k-1)} G$$

μέχρι να εκπληρωθεί κάποιο κριτήριο σύγκλισης (π.χ. μέχρι τα διανύσματα $\pi^{(k)}$ και $\pi^{(k+1)}$ να διαφέρουν το πολύ κατά ε = 0.01 ανά στοιχείο μεταξύ τους).
Πιο μαθηματικά, ο παραπάνω τύπος γράφεται ως:

$$\begin{aligned}
\pi^{(k)} &= \pi^{(k-1)} G \\
&= \pi^{(k-1)} [aS + (1-a)\mathbb{I}u] \\
&= \pi^{(k-1)} [a(H + dw) + (1-a)\mathbb{I}u] = \\
&= a\pi^{(k-1)} H + a(\pi^{(k-1)} d)w + (1-a)(\pi^{(k-1)}\mathbb{I})u
\end{aligned}$$

Επειδή ο πολλαπλασιασμός $\pi^{(k-1)}\mathbb{I}$ έχει ως αποτελέσμα τον αριθμό 1, αφού όπως αναφέρθηκε το άθροισμα των στοιχείων του διανύσματος $\pi$ είναι 1, η παραπάνω σχέση γράφεται ως:

$$\pi^{(k)} = a\pi^{(k-1)} H + a(\pi^{(k-1)} d)w + (1-a)u$$

Η παραπάνω έκφραση είναι στην ουσία ένα άθροισμα τριών πινάκων, εκ των οποίων οι δύο πρώτοι προκύπτουν από πολλαπλασιασμό πινάκων, ενώ ο τρίτος από πολλαπλασιαμό σταθεράς με διάνυσμα.

Στην ουσία, η μόνη χρονοβόρα υπολογιστικά μέθοδος είναι ο πολλαπλασιαμός των πινάκων $\pi^{(k-1)}H$. Σύμφωνα με μία έρευνα του 2004 [10], σχετικά με τα αρχεία που απαρτίζουν το Διαδίκτυο, βρέθηκε ότι ο μέσος όρος εξωτερικών συνδέσμων από μία ιστοσελίδα είναι κατά μέσο όρο 52. Αυτό πρακτικά σημαίνει, ότι μία μέση γραμμή του πίνακα του Διαδικτύου, θα περιλαμβάνει 52 από τα 25 δισεκατομμύρια περίπου μη μηδενικά στοιχεία. Συνεπώς ο πίνακας του Διαδικτύου είναι ένας αραιός πίνακας, όπου η πλειονότητα των στοιχείων του είναι μηδέν και άρα η Google μπορεί να εφαρμόσει κατάλληλους αλγόριθμους για αραιούς πίνακες, τόσο για την αποθήκευσή τους, όσο και για τον υπολογισμό του γινομένου τους με άλλους πίνακες.

## 2.5 Τα έσοδα της Google

Όπως αναφέρθηκε και παραπάνω, η Google αποτελεί την δημοφιλέστερη μηχανή διαδικτυακής αναζήτησης και φυσικά χρησιμοποιεί την ιδιότητά της αυτή ως μια διαρκή πηγή εσόδων. Στο πλήθος των υπηρεσιών που προσφέρει η εταιρεία, οι κυριότερες από αυτές είναι:

- Η υπηρεσία AdWords.
- Η υπηρεσία AdSense.

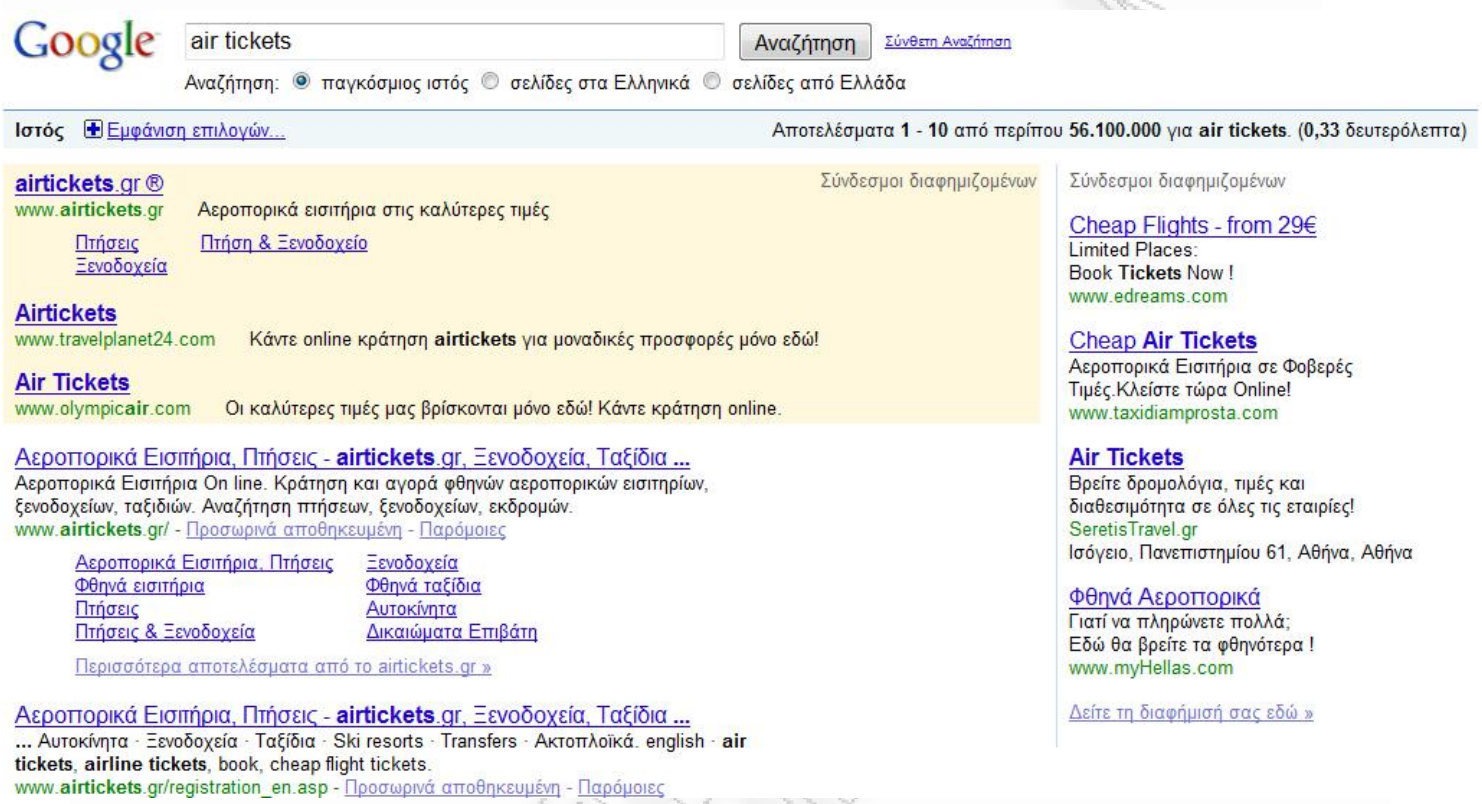
Θα αναλύσουμε τις δύο αυτές υπηρεσίες παρακάτω.

### 2.5.1 Η υπηρεσία AdWords

Η υπηρεσία AdWords αποτελεί την κύρια πηγή εσόδων της Google και ο συνολικός τζίρος γύρω από αυτό ήταν περίπου 23 δισεκατομμύρια δολλάρια το 2009 [14]. Η υπηρεσία AdWords αποτελεί μία pay-per-click (PPC) διαφημιστική υπηρεσία, σύμφωνα με την οποία, διαφημίσεις του κάθε πελάτη (με την μορφή υπερσυνδέσμων σε κάποια ιστοσελίδα του) εμφανίζονται στα αποτελέσματα συγκεκριμένων λέξεων κλειδιών της μηχανής Google. Όταν κάποιος χρήστης της μηχανής επιλέξει τον σύνδεσμο της διαφήμισης, τότε ο πελάτης οφείλει να πληρώσει την Google ένα

[14] http://investor.google.com/fin_data.html

συγκεκριμένο ποσό. Ένα παράδειγμα τέτοιων διαφημίσεων για τις λέξεις κλειδιά "air tickets" παρουσιάζεται στην Εικόνα 2.5.




**Εικόνα 2.5:** διαφημίσεις της υπηρεσίας AdWords

Οι ενδιαφερόμενοι, συμφωνούν αρχικά με την Google για τις λέξεις αναζήτησης όπου θα εμφανίζονται στα αποτελέσματα τα διαφημιστικά τους, καθώς και το ποσό όπου θα πληρώνουν για κάθε επιλογή ενός χρήστη. Η Google υλοποιεί έναν πολύπλοκο αλγόριθμο για την εμφάνιση των διαφημιστικών. Καταρχάς οι διαφημίσεις δεν είναι καθολικές, αλλά μπορεί να διαφέρουν από περιοχή σε περιοχή. Για παράδειγμα, αν κάποιος αναζητήσει αεροπορικά εισιτήρια στο ελληνικό διαδικτυακό τόπο της Google, θα εμφανιστούν κατά κύριο λόγο διαφημίσεις που αφορούν ελληνικές εταιρείες. Η ταξινόμηση των αποτελεσμάτων γίνεται βάση του ποσού που διαθέτει ο κάθε πελάτης αλλά και βάση της «βαθμολογίας ποιότητας» που έχει η κάθε ιστοσελίδα που διαφημίζεται. Η βαθμολογία ποιότητας είναι ένα μέγεθος της Google, όπου λαμβάνει κυρίως υπ'όψην και το ιστορικό της κάθε ιστοσελίδας, όσον αφορά προηγούμενες επιλογές χρηστών. Έτσι για παράδειγμα, μία ιστοσελίδα που διαφημίζεται και οι χρήστες την επιλέγουν έναντι των άλλων, θα εμφανίζεται ολοένα

και υψηλότερα στους συνδέσμους των διαφημιζόμενων. Επίσης, η βαθμολογία ποιότητας λαμβάνει υπ'όψην και την ποιότητα της ιστοσελίδας του διαφημιζόμενου, με βάση στοιχεία όπως το περιεχόμενό της, η ευκολία πλοήγησης, κτλ.

### 2.5.2 Η υπηρεσία AdSense

Η υπηρεσία AdSense αποτελεί την δεύτερη μεγαλύτερη πηγή εισόδων της Google. Σύμφωνα με την υπηρεσία αυτή, ένας διαδικτυακός τόπος μπορεί να συμμετάσχει στο συγκεκριμένο πρόγραμμα, προσθέτοντας σε αυτόν διαφημίσεις (οι οποίες περιλαμβάνουν κείμενο, εικόνα και ήχο). Οι διαφημίσεις αυτές προστίθενται κατ' επιλογήν του ιδιοκτήτη/διαχειριστή του διαδικτυακού τόπου, αλλά ελέγχονται και επιλέγονται από την ίδια την Google. Το όφελος από την συγκεκριμένη υπηρεσία, είναι ότι ο ιδιοκτήτης του διαδικτυακού τόπου πληρώνεται, κάθε φορά που κάποιος χρήστης επιλέξει κάποιον σύνδεσμο από αυτούς που παρουσιάζονται στις διαφημίσεις.

Οι διαφημίσεις οι οποίες εμφανίζονται, αφορούν συνήθως το περιεχόμενο του διαδικτυακού τόπου και ο σχεδιαστής του μπορεί να επιλέξει το σημείο όπου θα τις τοποθετήσει. Η υπηρεσία AdSense έχει γίνει ιδιαίτερα δημοφιλής, κυρίως επειδή οι διαφημίσεις οι οποίες παρουσιάζονται δεν έχουν την μορφή διαφημιστικών εικόνων (banners) και έτσι είναι περισσότερο καλαίσθητες. Έτσι, πλέον πάρα πολλοί διαδικτυακοί τόποι χρησιμοποιούν την υπηρεσία AdSense, η οποία είναι ιδιαίτερα βολική για μικρές σε μέγεθος επιχειρήσεις, οι οποίες δεν έχουν την δυνατότητα και τα οικονομικά μέσα, έτσι ώστε να δημιουργήσουν ένα ξεχωριστό τμήμα marketing, το οποίο θα δραστηριοποιείται στην εύρεση πελατών που ενδιαφέρονται να διαφημιστούν με τον τρόπο αυτό. Ιδιαίτερα ιστοσελίδες με πλούσιο και ενδιαφέρον περιεχόμενο, έχουν αρκετά κέρδη από την συγκεκριμένη υπηρεσία, όπως αυτό αναφέρεται και στην εν λόγω ιστοσελίδα της Google[15].

Η αύξηση των κερδών ενός διαδικτυακού τόπου με την χρήση της υπηρεσίας AdSense, μπορεί να γίνει με τους παρακάτω τρόπους:

- δημιουργούν καλαίσθητες σε εμφάνιση ιστοσελίδες, με πλούσιο και ενδιαφέρον περιεχόμενο, έτσι ώστε να προσελκύσουν το ενδιαφέρον διαφόρων εταιρειών που διαφημίζονται μέσω του AdSense.

[15] https://www.google.com/adsense/static/el/Success.html

- χρησιμοποιούν κείμενο στις ιστοσελίδες τους, το οποίο προσελκύει τους επισκέπτες να επιλέξουν τις διαφημίσεις του AdSense. Το κείμενο αυτό περιλαμβάνει φράσεις του είδους "Click on my AdSense Ads" ή "Sponsored Links" ή "Advertisements", κτλ. Η Google, αυξάνει το ποσό που θα λάβει ο ιδιοκτήτης μιας ιστοσελίδας, στην περίπτωση που εκτός από τις διαφημίσεις του AdSense εμφανίζονται και φράσεις όπως οι παραπάνω.

Όπως θα δούμε και στα παρακάτω κεφάλαια, η υπηρεσία AdSense αλλά και η υπηρεσία AdWords, είναι ιδιαίτερα ευάλωτες σε «επιθέσεις», με κυριότερες τις επιθέσεις όπου φαίνεται «εικονικά», ότι κάποιος χρήστης επέλεξε κάποιον από τους εικονικούς συνδέσμους. Αυτό θα μπορούσε για παράδειγμα να δημιουργήσει οικονομικά προβλήματα σε εταιρείες που διαφημίζονται μέσω του Google (για παράδειγμα έστω ότι κάποια ανταγωνίστρια εταιρεία επιλέγει συνεχώς τις διαφημίσεις μιας εταιρείας που εμφανίζονται μέσω της υπηρεσίας AdWords, με σκοπό να την ζημιώσει οικονομικά, αφού για κάθε επιλογή οφείλει να πληρώσει την Google) αλλά και στην ίδια εταιρεία Google (για παράδειγμα, μια εταιρεία που συμμετέχει στο AdSense, θα μπορούσε να επιλέγει συνεχώς τους συνδέσμους που εμφανίζονται στις διαφημίσεις της, εξαναγκάζοντας έτσι την Google να την πληρώνει συνεχώς). Φυσικά η Google έχει προνοήσει για τα προβλήματα αυτά κι έχει ήδη καταφύγει σε λύσεις, όπως θα δούμε και παρακάτω.

### 2.5.3 Λοιπές υπηρεσίες/προϊόντα

Εκτός από τις παραπάνω βασικές υπηρεσίες, οι οποίες αποτελούν και την κύρια πηγή των εισόδων της, η Google προσφέρει και μία πληθώρα άλλων υπηρεσιών, οι οποίες μεταξύ άλλων περιλαμβάνουν:

- την υπηρεσία ηλέκτρονικού ταχυδρομείου Gmail.
- την υπηρεσία χαρτών σε ψηφιακή μορφή Google Maps.
- τις διαφημιστικές υπηρεσίες Audio Ads, Click-to-Call, DoubleClick, κ.α.

# ΚΕΦΑΛΑΙΟ 3
# ΕΠΙΘΕΣΕΙΣ ΣΤΗΝ ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ GOOGLE

## 3.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο θα παρουσιαστούν αναλυτικά οι κυριότερες επιθέσεις που μπορούν να γίνουν στην μηχανή αναζήτησης Google. Η έννοια του όρου επίθεση σε μία μηχανή αναζήτησης, θα μπορούσε να πάρει πολλές έννοιες, κάποιες εκ των οποίων σχετίζονται με την λειτουργία της ως μηχανής αναζήτησης (π.χ. παραποίηση των αποτελεσμάτων) ενώ κάποιες άλλες με την λειτουργία της ως υπολογιστικό σύστημα (π.χ. καταστροφή δεδομένων της μηχανής αναζήτησης λόγω ιού). Στην συγκεκριμένη μελέτη θα αναφερθούμε κυρίως στην πρώτη κατηγορία επιθέσεων (αυτών που αφορούν την λειτουργία της μηχανής αναζήτησης) και πιο συγκεκριμένα στις παρακάτω:

- Η έννοια του Web Spam, τα είδη της και πως αυτά μπορούν να παραποιήσουν τα αποτελέσματα της μηχανής αναζήτησης.
- Google Bombing.
- Το φαινόμενο του Click Fraud και τα οικονομικά προβλήματα τα οποία μπορεί να προκαλέσει.
- Επιθέσεις σε άλλα προϊόντα της Google.
- Καταγεγραμμένες επιθέσεις εναντίον της Google.

Πολλά από τα παραπάνω είδη επιθέσεων, σχετίζονται με τον τρόπο λειτουργίας της μηχανής αναζήτησης Google και ιδιαίτερα με την μέθοδο PageRank. Παρά το γεγονός ότι πολλές από τις λεπτομέρειες υλοποίησης καθώς και τις παραμέτρους της μεθόδου δεν είναι γνωστές, οι επιτιθέμενοι, έχοντας γνώση της γενικής της λειτουργίας, μπορούν να παραποιήσουν (και μερικές φορές σε αρκετά μεγάλο βαθμό) τα αποτελέσματα της αναζήτησης ορισμένων λέξεων/φράσεων κλειδιών. Πολλές δε από τις πράξεις αυτές, είναι από την φύση τους παραπάνομες (και ιδιαίτερα αυτές οι οποίες αφορούν οικονομικά εγκλήματα). Στις παρακάτω ενότητες ακολουθεί μία αναλυτική περιγραφή των κυριότερων από αυτές τις επιθέσεις.

## 3.2 Ανάλυση του Web Spam

### 3.2.1 Ορισμός της έννοιας του Web Spam

Ο κύριος στόχος μιας μηχανής είναι αναζήτησης είναι η παράθεση διαδικτυακών τόπων, με περιεχόμενο σχετικό με μία ή περισσότερες λέξεις αναζήτησης. Οι διαδικτυακοί τόποι οι οποίοι εμφανίζονται, θα πρέπει να είναι ταξινομημένοι επίσης σε φθίνουσα σειρά, σχετικά με κάποιο κριτήριο, το οποίο συνήθως είναι ο βαθμός σχετικότητας που έχουν με το κλειδί αναζήτησης. Η έννοια της σχετικότητας μετράται διαφορετικά από κάθε μηχανή αναζήτησης. Η σχετικότητα συνήθως αφορά την ομοιότητα που παρουσιάζει το κλειδί αναζήτησης με το κείμενο της κάθε ιστοσελίδας.

Κάθε μηχανή αναζήτησης λοιπόν, ταξινομεί τα αποτελέσματα της με βάση ένα ή περισσότερα κριτήρια. Όπως είδαμε και στα παραπάνω κεφάλαια, η μηχανή αναζήτησης Google χρησιμοποιεί την τεχνική του PageRank για την απόδοση βαθμολογιών σε διαδικτυακούς τόπους, η οποία εκτός από την ομοιότητα του κειμένου της ιστοσελίδας με αυτό του κλειδιού αναζήτησης, λαμβάνει υπ'όψην και το πόσο «σημαντική» είναι η ιστοσελίδα (δηλαδή πόσες ιστοσελίδες αναφέρονται σε αυτή). Έτσι λοιπόν, ιστοσελίδες με μεγαλύτερες βαθμολογίες εμφανίζονται και υψηλότερα στα αποτελέσματα της αναζήτησης.

Χρησιμοποιούμε τον όρο **web spam** (ή αλλιώς ως **spamdexing**) για να αναφερθούμε σε ηθελημένες ενέργειες (από κάποιον άνθρωπο ή από κάποιο πρόγραμμα), οι οποίες έχουν ως στόχο να αυξήσουν (προς το καλύτερο) την βαθμολογία μιας ιστοσελίδας σε μία ή περισσότερες μηχανές αναζήτησης.

Ο παραπάνω ορισμός είναι στενά συνδεδεμένος με τις **μεθόδους βελτιστοποίησης μηχανών αναζήτησης (Search Engine Optimization – SEO)**. Μία μέθοδος βελτιστοποίησης μηχανών αναζήτησης, είναι μία συλλογή από τεχνικές, οι οποίες έχουν ως στόχο την βελτίωση την βαθμολογίας ενός διαδικτυακού τόπου. Πολλές από αυτές τις τεχνικές δεν έχουν απαραίτητα κακόβουλη έννοια και απλά χρησιμεύουν στην αναδιοργάνωση της ιστοσελίδας, έτσι ώστε να είναι καλύτερη αισθητικά αλλά και σε περιεχόμενο. Φυσικά κάποιες άλλες χρησιμοποιούν τεχνικές που όπως θα δούμε παρακάτω ενδεχομένως να μπερδέψουν μια μηχανή αναζήτησης και να την κάνουν να δίνει σε μια ιστοσελίδα βαθμολογία μεγαλύτερη της

πραγματικής. Υπάρχουν μάλιστα και εξειδικευμένες εταιρείες [16] [17], οι οποίες δραστηριοποιούνται στην βελτιστοποίηση της βαθμολογίας διαδικτυακών τόπων επί πληρωμή. Η τελευταία παρατήρηση λοιπόν, μας δείχνει ότι η έννοια του web spam είναι λίγο ως πολύ σχετική κι όχι απαραίτητα κακόβουλη και παράνομη, αφού κάποιος θα μπορούσε να θεωρήσει ως web spam και την απλή αναδιοργάνωση μιας ιστοσελίδες, με καλής ποιότητας περιεχόμενο. Συνήθως η έννοια του web spam χρησιμοποιείται όταν η εν λόγω μέθοδοι είναι εμφανές ότι έχουν ως μοναδικό σκοπό την αύξηση της δημοτικότητάς τους με όχι και τόσο ορθούς τρόπους.

Στις παρακάτω υπό-ενότητες, θα αναλύσουμε τις περισσότερο σημαντικές τεχνικές web spam και θα περιγράψουμε τα χαρακτηριστικά τους.

### 3.2.2 Εισαγωγή λέξεων κλειδιών στο σώμα της ιστοσελίδας

Η συγκεκριμένη μέθοδος **(keyword stuffing)** αποτελεί μία από τις πιο απλές και πιο δημοφιλείς τεχνικές web spam. Ο δημιουργός της ιστοσελίδας, εισάγει χειροκίνητα ή με την χρήση κάποιου προγράμματος, έναν μεγάλο όγκο από λέξεις κλειδιά σε όλη την ιστοσελίδα. Οι λέξεις κλειδιά μπορεί να είναι γενικού περιεχομένου ή να επικεντρώνονται σε ένα συγκεκριμένο θεματικό αντικείμενο. Για παράδειγμα, εάν μια ιστοσελίδα που αφορά ένα ηλεκτρονικό κατάστημα ανθοπωλείου επιθυμεί να χρησιμοποιήσει αυτή την τεχνική, τότε θα πρέπει να συμπεριλάβει στο περιεχόμενο της ιστοσελίδας λέξεις κλειδιά όπως «λουλούδια», «τριαντάφυλλα», «γαρύφαλα», κτλ. Ένα παράδειγμα μίας τέτοιας ιστοσελίδας παρουσιάζεται στην Εικόνα 3.1 της επόμενης σελίδας.

Φυσικά ένας πραγματικός χρήστης του Διαδικτύου μπορεί πολύ εύκολα να αντιληφθεί το γεγονός ότι μια ιστοσελίδα χρησιμοποιεί spam μεθόδους. Η διαπίστωση αυτή δεν είναι και τόσο εύκολο να γίνει όμως από μια μηχανή αναζήτησης, αφού πολλές φορές οι λέξεις κλειδιά έχουν τοποθετηθεί στην ιστοσελίδα με τρόπο τέτοιο, ώστε να μην γίνονται αντιληπτές, ακόμη και φαινομενικά.

[16] www.seoinc.com
[17] www.bruceclay.com


**Εικόνα 3.1:** παράδειγμα keyword stuffing web-spam

Όπως είναι φυσικό, ένας μεγάλος όγκος από λέξεις κλειδιά καθιστά την ιστοσελίδα ιδιαίτερα αντι-αισθητική και δύσχρηστη. Για τον λόγο αυτό, είναι ιδιαίτερα δημοφιλής η τεχνική του κρυφού κειμένου (hidden text). Σύμφωνα με την τεχνική αυτή, μέρη του κειμένου μιας ιστοσελίδας δεν είναι ορατά στον χρήστη, γιατί έχουν το ίδιο χρώμα με το φόντο της ιστοσελίδας. Ένα παράδειγμα παρουσιάζεται παρακάτω.

```
<body background=\white">
    <font color=\white">κρυφό κείμενο</font>
        .....
</body>
```

Για να μπορέσει κάποιος χρήστης να δει το κείμενο αυτό, θα πρέπει να δει τον HTML κώδικα της ιστοσελίδας. Βέβαια, η τεχνική αυτή δεν είναι απαραίτητα κακόβουλη, αφού θα μπορούσε για παράδειγμα να χρησιμοποιηθεί και σε ιστοσελίδες με γρίφους, παιχνίδια, κτλ. Η μηχανή αναζήτησης Google δεν λαμβάνει υπ'όψην το χρώμα του

κειμένου σε σχέση με το χρώμα του φόντου της ιστοσελίδας και συνεπώς πολλοί διαδικτυακοί τόποι που χρησιμοποιούν την τεχνική αυτή, ενδεχομένως να λαμβάνουν καλύτερες βαθμολογίες από την μέθοδο PageRank.

Εκτός από την παραπάνω τεχνική, δημοφιλής είναι και η τεχνική όπου εισάγει λέξεις κλειδιά στο κείμενο υπέρ-συνδέσμων. Η HTML χρησιμοποιεί την εντολή

```
<a href=”target_page.html”>Περιγραφή</a>
```

για την εισαγωγή υπέρ-συνδέσμων. Ο δημιουργός της ιστοσελίδας, για να μπορέσει να κρύψει ορισμένες spam διευθύνσεις, βάζει ως περιγραφή μία εικόνα, πολύ μικρών διαστάσεων (π.χ. ενός pixel με διάσταση 1x1), η οποία έχει το ίδιο χρώμα με το φόντο της ιστοσελίδας, έτσι ώστε να μην είναι ορατή από κάποιον φυλλομετρητή. Ένα παράδειγμα είναι το παρακάτω:

```
<a href=”target_page.html”><img src=”notVisibleImage.gif”></a>
```

### 3.2.3 Meta tag spam

Μία άλλη απλή και δημοφιλής τεχνική, η οποία όμως έχει αρχίσει να εγκαταλείπτεται τα τελευταία χρόνια, είναι η τεχνική του Meta tag spam, η οποία εισάγει λέξεις κλειδιά στο meta tags μέρος μιας HTML σελίδας, το οποίο έχει την παρακάτω σύνταξη:

```
<meta name=”keywords” content=”λέξη κλειδί 1, λέξη κλειδί 2, ...”>
```

Οι λέξεις κλειδιά χρησιμοποιούνται για να δώσουν μια περιγραφή για το περιεχόμενο της ιστοσελίδας,. Οι μηχανές αναζήτησης συνήθιζαν να δίνουν μεγάλη σημασία στα meta tags, αλλά όχι πλέον. Την σημερινή εποχή οι μηχανές αναζήτησης δίνουν συνήθως μεγαλύτερη σημασία στους όρους που εμφανίζονται στον τίτλο της ιστοσελίδας.

### 3.2.4 Πύλες Ιστοσελίδων (Doorway Pages)

Μία ακόμα συνηθισμένη τακτική, είναι αυτή των Πυλών Ιστοσελίδων. Μία πύλη είναι μια ιστοσελίδα, η οποία ανακατευθύνει στον επισκέπτη σε μία άλλη ιστοσελίδα. Η πύλη έχει στον περιεχόμενό της μία συλλογή από λέξεις κλειδιά, παρόμοια με αυτή που έχουν οι ιστοσελίδες που χρησιμοποιούν την τεχνική keyword stuffing, ευελπιστώντας ότι θα λάβει μεγάλη βαθμολογία από μια μηχανή αναζήτησης. Όταν κάποιος χρήστης μεταβεί στην ιστοσελίδα αυτή, τότε συνήθως μεταφέρεται αυτόματα στην πραγματική ιστοσελίδα, η οποία στην ουσία δεν έχει καμία σχέση με την πύλη της.. Το παραπάνω σενάριο φυσικά δεν είναι και το μόνο.

Πολλές ιστοσελίδες έχουν δύο εκδόσεις: μία έκδοση για τους web crawlers των μηχανών αναζήτησης και μία έκδοση για τους υπόλοιπους χρήστες (η συγκεκριμένη μέθοδος ονομάζεται cloaking). Η αναγνώριση για το εάν ο επισκέπτης της ιστοσελίδας είναι web crawler ή απλός χρήστης, μπορεί να γίνει με δύο τρόπους: είτε ελέγχοντας το πεδίο User-Agent κατά το στάδιο της αίτησης του πρωτοκόλλου HTTP ειτε ελέγχοντας την IP διεύθυνση του επισκέπτη. Όσον αφορά την πρώτη περίπτωση, η αίτηση ενός επισκέπτη, όσον αφορά το πρωτόκολλο HTTP, στην περίπτωση που πρόκειται για κάποιον web crawler, θα μπορούσε εν δυνάμει να είναι η παρακάτω:

```
GET / HTTP/1.1
Host: crawl-66-249-66-1.googlebot.com
User-Agent: Google Bot
```

Ο HTTP server για την συγκεκριμένη ιστοσελίδα, μπορεί να ρυθμιστεί με τρόπο τέτοιο ώστε όταν το πεδίο User-Agent έχει την ονομασία κάποιου γνωστού web crawler, να προωθεί την «ψεύτικη» σε αυτούς ιστοσελίδα, ενώ σε διαφορετική περίπτωση να προωθεί την «πραγματική».

Όπως είναι φυσικό, πολλοί web crawlers, έχοντας διαπιστώσει την παραπάνω τεχνική, ταυτοποιούν τον εαυτό τους χρησιμοποιώντας ονομάτα δημοφιλών φυλλομετρητών (π.χ. Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US)). Στην περίπτωση αυτή, ο HTTP server είναι ρυθισμένος έτσι, ώστε να ελέγχει την IP διεύθυνση του web crawler.

Βέβαια, η χρήση της παραπάνω μεθόδου δεν σημαίνει απαραίτητα ότι γίνεται με σκοπό την απόδοση μεγαλύτερων βαθμολογιών από τις διάφορες μηχανές αναζήτησης. Η μέθοδος αυτή χρησιμοποιείται από πολλές ιστοσελίδες, ώστε να δώσουν διαφορετικό περιεχόμενο σε χρήστες, ανάλογα με την γεωγραφική τους περιοχή. Για παράδειγμα, η ιστοσελίδα Amazon, έχει διαφορετική έκδοση για κάθε χωρά (ακόμα και η Google – www.google.com – ανακατευθύνει τον χρήστη ανάλογα με την χώρα επίσκεψης).

Επειδή η συγκεκριμένη μέθοδος (cloaking) παρουσιάζει ιδιαίτερη δυσκολία στον ακριβή ορισμό της, παρακάτω δίνονται μερικά παραδείγματα τα οποία εκλαμβάνονται μεθόδους επιθετικού cloaking (δηλαδή ως spam) από τις περισσότερες μηχανές αναζήτησης [14]:

- Η ιστοσελίδα η οποία στέλνεται στον web crawler περιέχει αρκετό σε περιεχόμενο κείμενο, ενώ η ιστοσελίδα η οποία στέλνεται στους φυλλομετρητές είναι σχεδόν κενή, περιέχοντας απλώς κώδικα JavaScript.
- Η ιστοσελίδα στέλνει κείμενο στον web crawler, ενώ στέλνει multimedia περιεχόμενο στους φυλλομετρητές (π.χ. macromedia Flash videos).
- Η ιστοσελίδα στέλνει περισσότερο κείμενο στον web crawler, απ'ότι στέλνει στον φυλλομετρητή. Η μέθοδος αυτή χρησιμοποιείται συνήθως όταν η ιστοσελίδα θέλει να στείλει περισσότερες λέξεις κλειδιά στον web crawler, οι οποίες όμως δεν θέλει να εμφανίζονται στον φυλλομετρητή του χρήστη.
- Όπως αναφέρθηκε και παραπάνω, μια ιστοσελίδα η οποία ανακατευθύνει τον επισκέπτη της σε καμία περίπτωση δεν θα πρέπει εξαρχής να θεωρηθεί ως ιστοσελίδα spam. Η μηχανή αναζήτησης θα πρέπει να ελέγξει αν οι διευθύνσεις στις οποίες γίνεται η ανακατεύθυνση σε web crawler και φυλλομετρητή ταυτίζονται και αν όχι, μόνον τότε η ιστοσελίδα θα πρέπει να ελεγχθεί περισσότερο.
- Οι σύγχρονες μηχανές αναζήτησης δίνουν μεγάλη σημασία για την απόδοση της τελικής βαθμολογίας, στις λέξεις που αναγράφονται στον τίτλο της ιστοσελίδας. Για τον λόγο αυτό πολλές ιστοσελίδες δίνουν διαφορετικό τίτλο ιστοσελίδας στους web crawlers και διαφορετικό στους φυλλομετρητές, κρατώντας το περιεχόμενο όμως της ιστοσελίδας ίδιο. Κάτι τέτοιο αντιμετωπίζεται πολύ αρνητικά από τις μηχανές αναζήτησης, οι οποίες επιβάλλουν χαμηλότερες βαθμολογίες στις εν λόγω ιστοσελίδες.

### 3.2.5 Link Spam

Όπως είδαμε και παραπάνω, η μέθοδος PageRank στηρίζεται συνοπτικά στο εξής: «μία ιστοσελίδα λαμβάνει καλύτερη βαθμολογία, όσο μεγαλύτερος είναι ο αριθμός των ιστοσελίδων που έχουν συνδέσμους στην ιστοσελίδα αυτή». Εκτός αυτού, η μέθοδος PageRank δίνει καλύτερες βαθμολογίες όταν οι ιστοσελίδες αυτές είναι περισσότερο σημαντικές από κάποιες άλλες. Για παράδειγμα, έχει διαφορετική σπουδαιότητα ένας σύνδεσμος σε μία ιστοσελίδα αν ο σύνδεσμος αυτός βρίσκεται σε ένα ελληνικό ιστότοπο, από ότι έχει αν βρίσκεται για παράδειγμα στο www.yahoo.com. Γνωρίζοντας το συγκεκριμένο γεγονός, πολλοί διαδικτυακοί τόποι, προσπαθούν να αυξήσουν την βαθμολογία τους αυξάνοντας με μια πληθώρα μεθόδων, τον αριθμό των εισερχόμενων προς αυτές συνδέσεων.

Οι συνηθέστερες μέθοδοι είναι οι εξής:

- Εικονικοί Ιστότοποι (Honey pots)
- Χρήση Διαδικτυακών Καταλόγων (Web Directories)
- Εισαγωγή συνδέσμων σε blogs, forums, κτλ.
- Link Farms

Παρακάτω θα αναφερθούμε συνοπτικά στις μεθόδους αυτές.

Οι εικονικοί ιστότοποι (Honey pots) είναι ένα σύνολο ιστοσελίδων, οι οποίες έχουν φαινομενικά ενδιαφέρον περιεχόμενο (π.χ. συλλογή με συγγράμματα εκπαιδευτικού περιεχομένου), αλλά περιέχουν κρυφούς συνδέσμους (με κάποια από τις μεθόδους που παρουσιάστηκαν παραπάνω) σε μία ή περισσότερες ιστοσελίδες. Αν το περιεχόμενό τους είναι αρκετά ενδιαφέρον, τότε υπάρχει μεγάλη πιθανότητα να εμφανιστούν ως σύνδεσμοι και σε άλλες ιστοσελίδες (στην ουσία δηλαδή να γίνουν περισσότερο «σημαντικές») με τελικό ωφελημένο τον αρχικό ιστότοπο που θέλουμε να βελτιώσουμε την βαθμολογία του.

Μία άλλη πολύ γνωστή μέθοδος είναι αυτή των Διαδικτυακών Καταλόγων. Ένας διαδικτυακός κατάλογος, είναι μία συλλογή από θέματα, όπου κάθε θέμα περιέχει συνδέσμους σε ιστοσελίδες με περιεχόμενο ίδιο με αυτό του εν λόγω θέματος. Η παρουσία μιας ιστοσελίδας σε κάθε θέμα, πολλών διαδικτυακών καταλόγων, θα βελτιώσει αναγκαστικά την βαθμολογία της.

Τα τελευταία χρόνια έχει γίνει ιδιαίτερα δημοφιλής, η μέθοδος εισαγωγής συνδέσμων σε blogs ή σε forums. Σύμφωνα με την μέθοδο αυτή, κάποιος μπορεί να

εισάγει συνδέσμους σε φαινομενικά αθώες απαντήσεις σε κάποια δημοσίευση ενός blog ή σε κάποιο θέμα ενός forum. Ειδικά σε μεγάλες διαδικτυακές κοινότητες, όπου η παρουσία ενός διαχειριστή ο οποίος θα ελέγχει όλες τις απαντήσεις είναι εξαιρετικά δύσκολη, η μέθοδος αυτή συναντάται πολύ συχνά. Εκτός από το πρόβλημα το οποίο δημιουργεί στις μηχανές αναζήτησης, η συγκεκριμένη μέθοδος δημιουργεί προβλήματα και στους χρήστες των blogs/forums, αφού η συνεχής παρουσία τέτοιων μηνυμάτων δυσχεραίνει την χρήση των εν λόγω διαδικτυακών τόπων. Ένα τέτοιο παράδειγμα παρουσιάζεται στην Εικόνα 3.2.


**Εικόνα 3.2:** παράδειγμα blog spam

Μία πολύ συνηθισμένη τέλος μέθοδος, είναι αυτή των Link Farms. Σύμφωνα με την μέθοδο αυτή, υπάρχει ένα σύνολο από ιστοσελίδες, όπου η κάθε η κάθε μία περιέχει έναν ή περισσότερους συνδέσμους προς όλες τις άλλες. Με τον τρόπο αυτό, όλες οι ιστοσελίδες επωφελούνται από την κοινή ανταλλαγή των υπερσυνδέσμων. Παρά το γεγονός ότι ορισμένες υπερσυνδέσεις ενδεχομένως να δημιουργήθηκαν από κάποιον χρήστη, στις περισσότερες των περιπτώσεων δημιουργούνται αυτόματα με την χρήση κάποιου προγράμματος. Η συγκεκριμένη μέθοδος, είναι πλέον πολύ προσιτή, αφού η αγορά μιας πληθώρας από διαδικτυακά ονόματα (domains) δεν απαιτεί μεγάλο κόστος.

### 3.2.6 Page Hijacking

Η συγκεκριμένη μέθοδος αποτελεί στην ουσία υπό-κατηγορία των Πυλών Ιστοσελίδων. Σύμφωνα με την μέθοδο αυτή, μία ιστοσελίδα δημιουργεί μία πανομοιότυπη έκδοση ενός δημοφιλούς διαδικτυακού τόπου, με την μόνη διαφορά ότι μεταφέρει τον χρήστη σε μία άλλη ιστοσελίδα, μετά από κάποιο σύντομο χρονικό διάστημα (π.χ. 1 δευτερόλεπτο). Φυσικά η ανακατεύθυνση αυτή εμφανίζεται μόνο στους χρήστες κι όχι στους web crawlers.

Αυτό θα έχει το εξής αποτέλεσμα: πολλές μηχανές αναζήτησεις, μετά την εκτέλεση της διαδικασίας του web crawling, ελέγχουν τις ιστοσελίδες που αποθήκευσαν, για τυχόν διπλότυπα. Αν υπάρχουν διπλότυπα, τότε συνήθως τα διαγράφουν (κρατώντας συνήθως αυτό που αποθηκεύτηκε σε αργότερο χρόνο). Αυτό όμως έχει ως αποτέλεσμα σε πολλές περιπτώσεις να μένει αποθηκευμένη η έκδοση της ψεύτικης κι όχι της αυθεντικής ιστοσελίδας, με αποτέλεσμα να εμφανίζεται αυτή στα αποτελέσματα. Για να γίνει περισσότερο κατανοητή η συγκεκριμένη μέθοδος, ας δούμε το παρακάτω παράδειγμα.

Έστω ότι μια ιστοσελίδα πουλά εξαρτήματα αυτοκινήτων και εμφανίζεται υψηλά στα αποτελέσματα της Google ως εξής:

| |
|---|
| Car Parts<br>Selling great car parts!<br>www.car-parts.com |

Έστω τώρα, κάποιος δημιουργεί μία πανομοιότυπη σε περιεχόμενο ιστοσελίδα, στην διεύθυνση www.fake-address.com. Υπάρχει ένα μεγάλο ενδεχόμενο, μετά από κάποιο χρονικό περιθώριο, το παραπάνω αποτέλεσμα να εμφανιστεί ως εξής:

| |
|---|
| Car Parts<br>Selling great car parts!<br>www.fake-address.com |

Όταν ένας χρήστης συνδεθεί με την συγκεκριμένη ιστοσελίδα, τότε αυτή θα τον ανακατευθύνει στην διεύθυνση www.car-parts.net, εταιρείας ανταγωνιστικής της

www.car-parts.com. Η συγκεκριμένη απάτη είναι αρκετά συχνή κι έχει δημιουργήσει πολλά προβλήματα στην μηχανή αναζήτησης Google.

### 3.2.7 Χρήση κλεμμένου υλικού (Article Spinning)

Πολλές ιστοσελίδες (κυρίως blogs), για να μπορέσουν να αυξήσουν την κατάταξή τους στην μηχανή αναζήτησης Google, χρησιμοποιούν την μέθοδο article spinning, η οποία στην ουσία αποτελεί μία έξυπνη παραλλαγή της μεθόδου keyword stuffing. Σύμφωνα με την μέθοδο αυτή, ιδιαίτερα ιστοσελίδες όπου έχουν άρθρα ως περιεχόμενό τους, παραθέτουν το ίδιο άρθρο πολλές φορές, παραφράσσοντας ορισμένες προτάσεις ή αλλάζοντας λέξεις με τα συνώνυμά τους. Για παράδειγμα, η αγγλική λέξη "picture", όταν εμφανίζεται σε κάποιο άρθρο, μπορεί να εμφανίζεται σε μία αντιγραφή του ως "photo" ή ως "image". Η αλλαγή των άρθρων δεν γίνεται απαραίτητα από κάποιον άνθρωπο και θα μπορούσε πολύ εύκολα να γίνει από κάποιο πρόγραμμα, το οποίο λαμβάνει τα συνώνυμα από κάποιο λεξικό. Η συγκεκριμένη μέθοδος αποτελεί έναν έξυπνο τρόπο για την τοποθέτηση λέξεων κλειδιών, με τρόπο τέτοιο όπου μια μηχανή αναζήτησης δύσκολα θα αντιληφθεί ως keyword stuffing. Φυσικά, πολλές παραφράσεις άρθρων δεν έχουν ιδιαίτερο νόημα αν διαβαστούν από κάποιον πραγματικό χρήστη, ιδιαίτερα αν για την δημιουργία τους χρησιμοποιήθηκε κάποιο πρόγραμμα. Για τον λόγο αυτό, πολλές από αυτές τις ιστοσελίδες χρησιμοποιούνται ως πύλες ιστοσελίδων, ώστε να ανακατευθύνουν τον χρήστη στο πραγματικό διαδικτυακό τόπο.

## 3.3 Βόμβες Google

Ο όρος «Βόμβες Google» ή αλλιώς Google Bombing αποτελεί μία από τις πιο δημοφιλείς τεχνικές, η οποία παραλλάσει τα αποτελέσματα της μηχανής αναζήτησης Google. Η μέθοδος αυτή περιγράφει μια πρακτική ηλεκτρονικού ακτιβισμού, η οποία αφορά εσκεμμένη απόπειρα να αλλοιωθεί η σειρά ταξινόμησης μιας συγκεκριμένης ιστοσελίδας στα αποτελέσματα που παράγονται από τη μηχανή αναζήτησης Google. Απώτερος σκοπός της μεθόδου, είναι η σύνδεση μιας ιστοσελίδας με λέξεις-κλειδιά που συνήθως έχουν προσβλητικό ή χιουμοριστικό περιεχόμενο.

Όπως είδαμε και στο κεφάλαιο 2, η μέθοδος PageRank δίνει μεγαλύτερες βαθμολογίες σε ιστοσελίδες, όπου υπάρχουν πολλοί σύνδεσμοι που οδηγούν σε αυτές. Η γενική σύνταξη της εντολής εισαγωγής ενός υπέρ-συνδέσμου της HTML είναι η παρακάτω:

```
<A href="www.site.com" title="A nice site">A nice site</A>
```

Ο τίτλος του κάθε συνδέσμου αλλά και η περιγραφή, χρησιμοποιείται από την μηχανή αναζήτησης Google για την απόδοση λέξεων-κλειδιών για κάθε ιστοσελίδα. Εάν για παράδειγμα ένας μεγάλος αριθμός συνδέσμων έχουν ως τίτλο ή/και ως περιγραφή για την ιστοσελίδα www.site.com το κείμενο "A nice site", τότε η μηχανή αναζήτησης μπορεί εύλογα να θεωρήσει ότι η παραπάνω φράση περιέχει λέξεις κλειδιά για την ιστοσελίδα αυτή.

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

Οι βόμβες Google οργανώνονται ανεπίσημα μεταξύ κατόχων ιστολογιών (blogs) ή άλλων ιστότοπων, με συμφωνία και εθελοντική τοποθέτηση τέτοιων συνδέσμων με το ίδιο κείμενο και προορισμό τον ίδιο ιστότοπο. Συνήθως πραγματοποιούνται είτε ως αστείο, είτε για 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, κάτι που χαρακτηρίζεται ως spam, και για την καταπολέμηση αυτού του φαινομένου, έχουν δημιουργηθεί διάφοροι τρόποι φιλτραρίσματος των καταχωρήσεων ή ακύρωσης των συνδέσμων.

Η Google, από την πλευρά της, προσπάθεί να καταπολεμήσει τέτοιου είδους προσπάθειες. Έτσι, τα αποτελέσματα της αναζήτησης διορθώνονται αμέσως μόλις εντοπιστεί κάποια προσπάθεια εξαπάτησης της μηχανής αναζήτησης.

Μερικά διάσημα παραδείγματα βομβών Google (κυρίως όσον αφορά τον ελληνικό χώρο), παρουσιάζονται παρακάτω:

- Ο συσχετισμός της ιστοσελίδας με την βιογραφία του προέδρου των ΗΠΑ Τζόρτζ Μπους, τον Ιούνιο του 2005 (Εικόνα 3.3).


**Εικόνα 3.3:** βόμβα Google με στόχο την ιστοσελίδα με την βιογραφία του προέδρου των ΗΠΑ Τζόρτζ Μπους. Η ιστοσελίδα συνδέεται με την φράση "miserable failure".

- Η λέξη κλειδί **ληστές**, οδηγούσε στον δικτυακό τόπο του Ο.Τ.Ε.. Έγινε σε ένδειξη διαμαρτυρίας για τις αυξήσεις στα τιμολόγια της πρόσβασης στο Διαδίκτυο μέσω τηλεφωνικής κλήσης (dial-up) που η εταιρία ανακοίνωσε τον Νοέμβριο του 2005.
- Η λέξη κλειδί **ατσαλάκωτος**, οδηγούσε στο site του δημάρχου Θεσσαλονίκης Βασίλη Παπαγεωργόπουλου. Έγινε μάλλον από πολιτικούς του αντιπάλους.

## 3.4 Click Fraud

Ο όρος Click Fraud αποτελεί ένα είδος μιας διαδικτυακής επίθεσης, σύμφωνα με την οποία ο επιτιθέμενος χρησιμοποιεί ένα πρόγραμμα, το οποίο επιλέγει εικονικά, πολλές φορές, διαφημίσεις της Google που προβάλλονται με την χρήση του AdSense. Η μέθοδος αυτή ενδέχεται να δημιουργήσει σοβαρά οικονομικά προβλήματα στον ιδιοκτήτη των εν λόγω διαφημίσεων.

Όπως είδαμε και σε παραπάνω ενότητα, η υπηρεσία AdSense είναι στην ουσία μία διαφημιστική μέθοδος, σύμφωνα με την οποία, κάποιος που θέλει να διαφημίσει τα προϊόντα του, επικοινωνεί με την Google και συμφωνεί στο κείμενο και στις εικόνες των διαφημίσεων. Στην συνέχεια, ένας ιδιοκτήτης ενός ιστότοπου, μπορεί να προσθέσει κατ'επιλογήν του τις διαφημίσεις αυτές, στον ιστότοπό του. Κάθε φορά που κάποιος επιλέγει μία διαφήμιση, τότε ο ιδιοκτήτης των διαφημίσεων οφείλει να καταβάλλει ένα ποσό στον ιδιοκτήτη του ιστοτόπου. Όσο περισσότερες είναι οι επιλογές σε μία συγκεκριμένη διαφήμιση, τόσο μεγαλύτερο είναι και το ποσό που θα πρέπει να καταβάλλει ο ιδιοκτήτης της.

Σύμφωνα με τα παραπάνω λοιπόν, η χρήση ενός αυτοματοποιημένου προγράμματος, το οποίο «επιλέγει» εικονικά, πάρα πολλές φορές διαφημίσεις ενός συγκεκριμένου ιδιοκτήτη, ενδέχεται να τον ζημιώσει οικονομικά σε μεγάλο βαθμό. Φυσικά η χρήση Η/Υ για την πρόκληση οικονομικού εγκλήματος, είναι παράνομη σε πολλές χώρες παγκοσμίως και φυσικά στις ΗΠΑ, όπου εδρεύει η εταιρεία Google.

Το πρόβλημα γίνεται ακόμα πιο έντονο, αφού σε πολλές περιπτώσεις ο ιδιοκτήτης των διαφημίσεων, είναι άμεσα ή έμμεσα είναι η ίδια η μηχανή αναζήτησης. Πολλές φορές η Google εισάγει με την χρήση του AdSense διαφημίσεις για τα προϊόντα της ή για προϊόντα για τα οποία είναι συνέταιρος. Έτσι λοιπόν, όταν κάποιος επιλέγει συνεχώς με την χρήση προγράμματος κάποιες από αυτές τις διαφημίσεις, τότε ζημιώνει οικονομικά την ίδια την Google, αφού θα πρέπει να καταβάλλει ένα οικονομικό ποσό στον ιδιοκτήτη του διαδικτυακού τόπου, όπου παρατίθενται οι διαφημίσεις της.

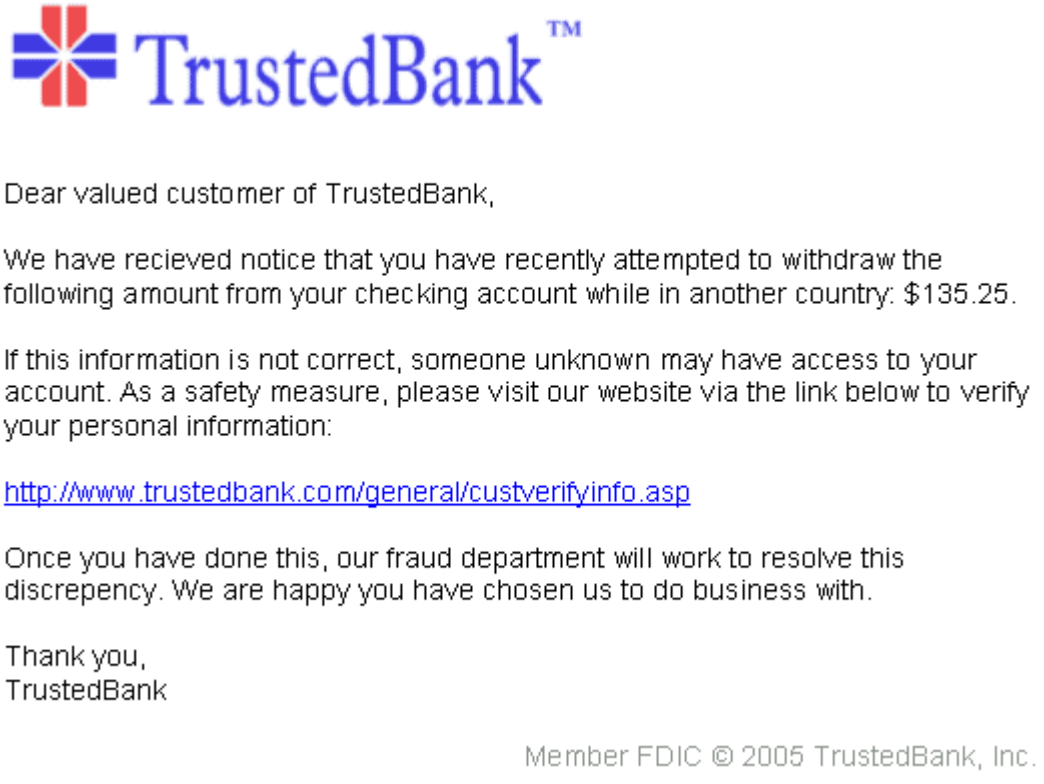
## 3.5 Η τεχνική του Phising

Μία από τις πιο συνηθισμένες μεθόδους εξαπάτησης μέσω διαδικτυακών ιστοσελίδων, η οποία δεν έχει να κάνει άμεσα με τις μηχανές αναζήτησης, είναι η μέθοδος Phising. Σύμφωνα με την μέθοδο αυτή, ο επιτιθέμενος προσπαθεί να υποκλέψει προσωπικές πληροφορίες, οι οποίες αφορούν συνήθως στοιχεία εγκυροποίησης (username/password) ή στοιχεία τραπεζικών λογαριασμών και πιστωτικών καρτών. Αυτό το κάνει, δημιουργώντας έναν διαδικτυακό τόπο, ο οποίος είναι παρόμοιος (ή μερικές φορές εντελώς ταυτόσημος) του πραγματικού, δίνοντας την ψευδαίσθηση στον χρήστη ότι εισάγει τα δεδομένα αυτά στον σωστό διαδικτυακό τόπο. Τα δεδομένα αυτά στην πραγματικότητα αποθηκεύονται από τον διαδικτυακό τόπο και στην συνέχεια χρησιμοποιούνται για την υπεξαίρεση χρηματικών ποσών ή άλλου είδους δολιοφθορές. Η επίσκεψη του χρήστη σε ένα τέτοιον διαδικτυακό τόπο γίνεται συνήθως είτε μέσω ψεύτικων ηλεκτρονικών μηνυμάτων (Εικόνα 3.4) είτε μέσω παραπλανητικών συνδέσμων σε ιστότοπους είτε μέσω λανθασμένων αποτελεσμάτων που εμφανίζονται σε αποτελέσματα σελίδων αναζήτησης.

Η παρουσία διαδικτυακών τόπων με τέτοιο περιεχόμενο στα αποτελέσματα μιας μηχανής αναζήτησης αποτελεί έναν κύριο παράγοντα που καθορίζει την αξιοπιστία της. Το πρόβλημα δεν είναι τόσο απλό να λυθεί, αφού η μηχανή αναζήτησης θα πρέπει να προειδοποιεί τον χρήστη (ή να μην περιλαμβάνει στα αποτελέσματά της) όχι μόνο για τις ιστοσελίδες που μιμούνται έναν διαδικτυακό τόπο αλλά και για τις ιστοσελίδες που περιέχουν συνδέσμους σε ιστοσελίδες που χρησιμοποιούν την εν λόγω μέθοδο. Μια ιστοσελίδα επίσης μπορεί να κάνει χρήση ενός συνδυασμού τεχνικών ώστε η μηχανή αναζήτησης να την συμπεριλάβει στα αποτελέσματά της (π.χ. μία ή περισσότερες από τις μεθόδους web spam που περιγράφηκαν παραπάνω), κάνοντας έτσι τον εντόπισμό της περισσότερο δύσκολο.

Όσον αφορά την μηχανή αναζήτησης Google, ισχυρίζεται ότι μπορεί να ανακαλύψει 9 στις 10 τέτοιες ιστοσελίδες, από αυτές που αποθηκεύουν οι web crawlers της {12]. Κάθε φορά που η Google ανακαλύπτει (ή υποπτεύεται) μία τέτοια ιστοσελίδα, εμφανίζει το παρακάτω μήνυμα:

> "Warning - phishing (web forgery) suspected. The site you are trying to visit has been identified as a forgery, intended to trick you into disclosing financial, personal or other sensitive information."




**Εικόνα 3.4:** παράδειγμα ψεύτικου ηλεκτρονικού μηνύματος που παραπέμπει σε διαδικτυακό τόπο με σκοπό την καταγραφή των στοιχείων ενός λογαριασμού [18]

## 3.6 Πρόσφατες επιθέσεις κατά της Google

Στην ενότητα αυτή θα ασχοληθούμε με πρόσφατες επιθέσεις κατά της εταιρείες Google, οι οποίες έλαβαν μεγάλη δημοσιότητα τα τελευταία χρόνια. Η περισσότερο γνωστή από αυτές τις επιθέσεις, είναι αυτή που συνέβη τον Δεκέμβριο του 2009 [11] και που αργότερα πήρε την ονομασία "Aurora". Η επίθεση κοινοποιήθηκε από την ίδια την Google, στις 12 Ιανουαρίου 2010, σε μία δημοσίευση στο επίσημο blog της [11]. Εκτός από την Google, οι επιτιθέμενοι στόχευσαν και άλλες εταιρείες που δραστηριοποιούνται στον τομέα της Πληροφορικής, όπως η Yahoo! και η Adobe.

Στην συγκεκριμένη δημοσίευση, η Google ισχυρίζεται ότι η συγκεκριμένη επίθεση είχε ως βάση της την Λαική Δημοκρατία της Κίνας. Η Google είχε αρχίσει να δραστηριοποιείται στην Κίνα, από τον Ιανουάριο του 2006 [11], με την προϋπόθεση ότι θα θέτει περιορισμούς και λογοκρισία στις ιστοσελίδες οι οποίες

[18] http://upload.wikimedia.org/wikipedia/commons/d/d0/PhishingTrustedBank.png

παρουσιάζονται στα αποτελέσματά της. Μετά την επίθεση, η Google ανακοίνωσε ότι θα άρει τους περιορισμούς αυτούς [11].

Κατά την επίθεση αυτή, η Google αναφέρει ότι κλάπηκαν μερικά στοιχεία τα οποία έχουν κατοχυρωθεί ως πνευματική της ιδιοκτησία νομικά. Οι επιτιθέμενοι έδειξαν ιδιαίτερο ενδιαφέρον για τους λογαριασμούς ηλεκτρονικού ταχυδρομείου (Gmail accounts) κάποιων κινέζων οι οποίοι είναι ακτιβιστές ανθρωπίνων δικαιωμάτων και διώκονται ποινικά στην Κίνα. Το μόνο που κατάφεραν είναι οι επιτιθέμενοι, είναι να αποκτήσουν πρόσβαση στους τίτλους των ηλεκτρονικών μηνυμάτων, καθώς και στις ημερομηνίες δημιουργίας τους, αλλά όχι στο περιεχόμενό τους. Επιπλέον η Google ανακάλυψε ότι ένας μεγάλος αριθμός από λογαριασμούς ηλεκτρονικού ταχυδρομείου που ανήκουν σε ακτιβιστές ανθρωπίνων δικαιωμάτων, προσπελάσσεται παράνομα από τρίτους χρήστες, οι οποίοι είναι συνδεδεμένοι από παροχείς Διαδικτύου (ISPs) της Κίνας. Τα στοιχεία ταυτοποίησης των εν λόγω λογαριασμών κλάπηκαν με κάποιον τρόπο από τους πραγματικούς τους ιδιοκτήτες και η Google δεν είχε καμία ευθύνη στην εν λόγω κλοπή.

# ΚΕΦΑΛΑΙΟ 4
# ΑΝΤΙΜΕΤΩΠΙΣΗ ΤΩΝ ΕΠΙΘΕΣΕΩΝ

## 4.1 Εισαγωγή

Οι επιθέσεις που αναφέρθηκαν στο προηγούμενο κεφάλαιο, αποτελούν ένα μεγάλο πρόβλημα για την Google (αλλά και για κάθε μηχανή αναζήτησης), αφού την ζημιώνουν οικονομικά και μειώνουν την αξιοπιστία και την δημοτικότητά της. Η αντιμετώπιση της κάθε μιας από τις υπάρχοντες επιθέσεις αποτελεί από μόνη της ένα ξεχωριστό πρόβλημα και δεν είναι πάντοτε εύκολο να αντιμετωπιστεί πλήρως. Στο συγκεκριμένο κεφάλαιο παρουσιάζονται οι κυριότερες μέθοδοι για την αντιμετώπιση των περισσότερο κοινών επιθέσεων. Συνοπτικά, οι μέθοδοι που παρουσιάζονται αφορούν τις παρακάτω επιθέσεις:

**Αντιμετώπιση του web spam**

Για την αντιμετώπιση του web spam χρησιμοποιείται μια πληθώρα τεχνικών, με περισσότερο διαδεδομένη τον αλγόριθμο του TrustRank, του οποίου χρήση κάνει και η μηχανή αναζήτησης Google [9]. Εκτός από τον αλγόριθμο αυτόν, γίνεται και χρήση και άλλων τεχνικών, με σημαντικότερες την εξέταση του ποσοστού των συνδέσμων σε σχέση με το συνολικό κείμενο της ιστοσελίδας, την ανάλυση με αλγορίθμους μηχανικής μάθησης, κτλ.

**Αντιμετώπιση του click fraud**

Για την αντιμετώπιση της απάτης του click fraud, η Google χρησιμοποιεί μια πληθώρα απο τεχνικές, οι οποίες χωρίζονται σε 3 διαδοχικά στάδια. Οι τεχνικές αυτές περιγράφονται αναλυτικά παρακάτω.

**Αντιμετώπιση του phising**

Όπως αναφέρθηκε και παραπάνω, η εν λόγω απάτη ενδέχεται να δημιουργήσει πολλά προβλήματα στην αξιοπιστία της Google και για τον λόγω αυτό θα πρέπει να προσεχθεί ιδιαίτερα. Η Google χρησιμοποιεί έναν συνδυασμό από τεχνικές για την αντιμετώπιση της εν λόγω απάτης, οι οποίες επίσης παρουσιάζουν θεαματικά αποτελέσματα [17].

## 4.2 Ο αλγόριθμος αξιολόγησης TrustRank

Όπως αναφέρθηκε και στο κεφάλαιο 2, η μηχανή αναζήτησης Google χρησιμοποιεί τον αλγόριθμο αξιολόγησης TrustRank, για την αντιμετώπιση των ιστοτόπων που χρησιμοποιούν spam μεθόδους. Η μέθοδος TrustRank παρουσιάστηκε αρχικά το 2004, από τους Gyongyi, Garcia-Molina και Pedersen [9]. Ο αλγόριθμος TrustRank λειτουργεί ως συμπληρωματικός αυτού του PageRank, δίνοντας επιπρόσθετες πληροφορίες για το κατά πόσο ένας ιστότοπος είναι έγκυρος και δεν χρησιμοποιεί μεθόδους spam. Τα τελευταία μάλιστα χρόνια, η έννοια του TrustRank έχει γίνει ευρέως γνωστή και πολλοί ιστότοποι δίνουν περισσότερη σημασία στο να ανεβάσουν την βαθμολογία που τους δίνει η μέθοδος TrustRank, απ'ότι η μέθοδος PageRank [13].

Η μέθοδος TrustRank σε γενικές γραμμές προσπαθεί να προσομοιώσει την λειτουργία ενός πραγματικού «ελεγκτή» ιστοσελίδων, ο οποίος ελέγχει όλα τα αρχεία τα οποία έχει αποθηκεύσει ο web crawler μιας μηχανής αναζήτησης. Όπως είναι φυσικό, ένας άνθρωπος μπορεί με μια απλή εξέταση μιας ιστοσελίδας, τόσο οπτικά όσο και σε επίπεδο HTML κώδικα, να διαπιστώσει αν χρησιμοποιεί μία ή περισσότερες spam μεθόδους. Για παράδειγμα, αν κάποιος πραγματικός επισκέπτης μιας ιστοσελίδας δει μια ακαταννόητη παράθεση διάφορων λέξεων ή έναν πολύ μεγάλο αριθμό συνδέσμων να οδηγούν στον ίδιο ιστότοπο, εύκολα διαπιστώνει ότι πρόκειται για ιστοσελίδα spam. Φυσικά μερικές φορές η διαπίστωση αυτή δεν είναι και τόσο προφανής, ειδικά όταν πρόκειται για μια πληθώρα από ιστοσελίδες, οι οποίες ανταλλάζουν συνδέσμους μεταξύ τους (Link Farms). Φυσικά σε μερικές περιπτώσεις, η κατηγοριοποίηση μιας ιστοσελίδας στην κατηγορία των spam ιστοσελίδων, ίσως αποτελεί υποκειμενικό ζήτημα του ανθρώπινου επισκέπτη και ο διαχωρισμός ίσως να μην είναι και τόσο εύκολος.

Η εξεέταση της κάθε ιστοσελίδας, από ανθρώπινο παράγοντα και μόνο. φυσικά δεν είναι αποδοτική, αφού ο αριθμός των αρχείων που αποθηκεύονται καθημερινά από τους web crawlers είναι τεράστιος σε μέγεθος. Σε διαφορά όμως με τον αλγόριθμο PageRank, ο αλγόριθμος TrustRank δεν είναι εντελώς αυτοματοποιημένος, αλλά λειτουργεί σε συνεργασία με κάποιον πραγματικό ελεγκτή, ο οποίος έχει αρχικά εξετάσει ένα μικρό υποσύνολο των ιστοσελίδων που

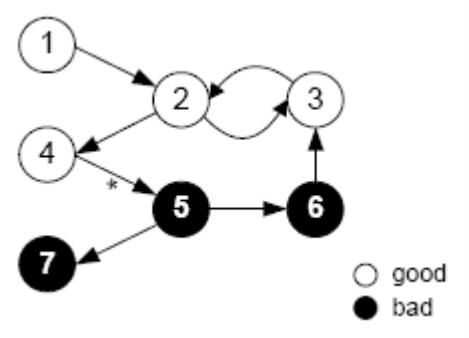
αποθήκευσε ο web crawler. Θα εξετάσουμε την λειτουργία του αλγορίθμου TrustRank συνοπτικά, στις παρακάτω υπό-ενότητες.

### 4.2.1 Το μοντέλο που χρησιμοποιεί ο αλγόριθμος TrustRank

Όπως και ο αλγόριθμος PageRank, έτσι και ο αλγόριθμος TrustRank, θεωρεί το Διαδίκτυο ως έναν κατευθυνόμενο γράφο, έστω $V$, όπου κόμβοι του γράφου αυτού είναι οι ιστοσελίδες του Διαδικτύου, ενώ οι ακμές του γράφου είναι οι σύνδεσμοι μεταξύ τους. Για να μοντελοποιήσουμε το γεγονός ότι κάποιος επισκέπτης αντιλαμβάνεται ότι μια ιστοσελίδα χρησιμοποιεί spam μεθόδους ή όχι, χρησιμοποιούμε την παρακάτω δυαδική συνάρτηση O, για κάθε ιστοσελίδα $p \in V$:

$$O(p) = \begin{cases} 0 & \text{αν η ιστοσελίδα } p \text{ χρησιμοποιεί spam μεθόδους} \\ 1 & \text{σε διαφορετική περίπτωση} \end{cases}$$

Στην εικόνα 4.1 παρουσιάζεται ένα παράδειγμα ενός μικρού τέτοιου διαδικτυακού γράφου με 7 κόμβους, όπου οι κανονικές ιστοσελίδες απεικονίζονται με λευκό χρώμα, ενώ οι ιστοσελίδες που χρησιμοποιούν spam μεθόδους με μαύρο. Για τον συγκεκριμένο γράφο ισχύει ότι O(1) = O(2) = O(3) = O(4) = 1 και O(5) = O(6) = O(7) = 0.




**Εικόνα 4.1**[19]: παράδειγμα διαδικτυακού γράφου με κανονικές (λευκές) ιστοσελίδες και ιστοσελίδες που χρησιμοποιούν κάποια μέθοδο spam (μαύρες)

[19] Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen, "Combating web spam with trustrank", proceedings of the Thirtieth international conference on Very large data bases - Volume 30, σελ. 578, Figure 2

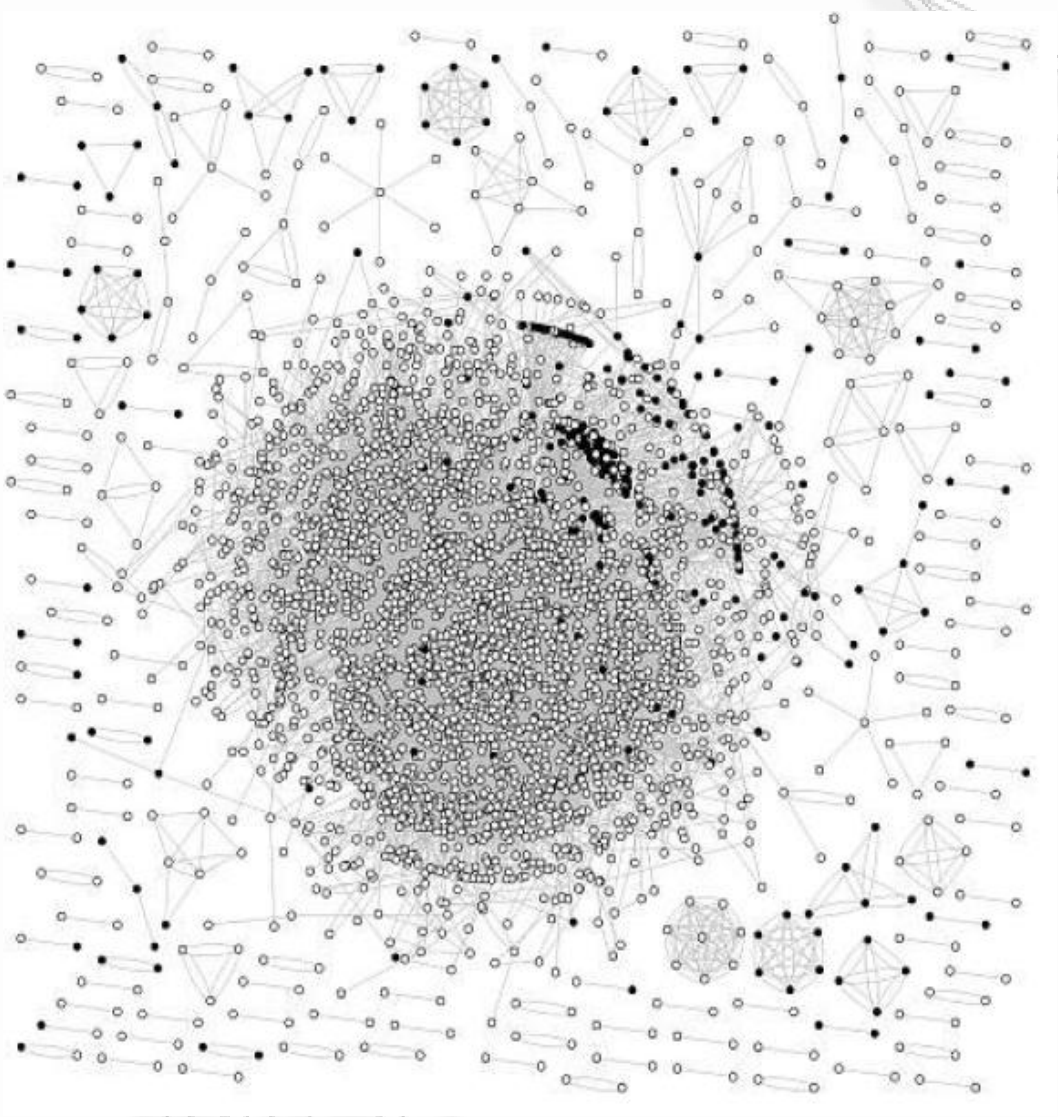
Ακόμη και αν η εκτίμηση της παραπάνω συνάρτησης ήταν εντελώς έγκυρη και αυτοματοποιημένη, η κλήση της για όλες τις ιστοσελίδες που έχει αποθηκεύσει ο web crawler της μηχανής αναζήτησης δεν είναι και τόσο αποδοτική (ένας web crawler μπορεί να αποθηκεύει εκατομμύρια ιστοσελίδες καθημερινά). Για την αντιμετώπιση του παραπάνω προβλήματος, ο αλγόριθμος TrustRank χρησιμοποιεί κάποιες παραδοχές, με στόχο να μειώσει το απαιτούμενο υπολογιστικό κόστος.

Θεωρούμε λοιπόν αρχικά, ότι ένα μικρό υποσύνολο των συνολικών ιστοσελίδων που έχουν αποθηκευθεί βαθμολογούνται με βάση την συνάρτηση Ο, είτε αυτοματοποιημένα είτε από κάποιον άνθρωπο. Η βασική παραδοχή η οποία κάνει ο αλγόριθμος TrustRank είναι η εξής: μία ιστοσελίδα που δεν χρησιμοποιεί κάποια spam μέθοδο έχει λίγες πιθανότητες να περιέχει έναν ή περισσότερους συνδέσμους σε ιστοσελίδες που χρησιμοποιούν spam μεθόδους. Φυσικά αυτό δεν ισχύει πάντα, αφού για παράδειγμα ο ιδιοκτήτης μιας τέτοια ιστοσελίδας ίσως πέσει θύμα απάτης και η ιστοσελίδα περιέχει εν τέλει συνδέσμους σε spam ιστοτόπους. Η συγκεκριμένη παραδοχή/ιδιότητα καλείται *«κλειστότητα κατά προσέγγιση του συνόλου των ιστοσελίδων που δεν χρησιμοποιούν spam μεθόδους» (approximate isolation of the good set)*. Ο κανόνας αυτός ισχύει πειραματικά, όπως παρουσιάζεται και στην Εικόνα 4.2 της επόμενης σελίδας, όπου υπάρχει μεγάλοι συγκέντρωση μεταξύ γειτονικών οι μαύρων κόμβων, οι οποίοι αποτελούν spam ιστοσελίδες Στον γράφο που παρουσιάζεται στην Εικόνα 4.1, παρατηρούμε ότι μία τέτοια περίπτωση είναι ο σύνδεσμος από την ιστοσελίδα 4 στην ιστοσελίδα 5 και για τον λόγο αυτό στην συγκεκριμένη ακμή απεικονίζεται ένα αστεράκι. Ο κανόνας ισχύει για όλες τις άλλες περιπτώσεις (συνδέσμους) του συγκεκριμένου γράφου.

Τα πράγματα περιπλέκονται ακόμη περισσότερο, όταν εκτός από την περίπτωση που περιγράφηκε παραπάνω, ενδεχομένως μια spam ιστοσελίδα περιλαμβάνει συνδέσμους σε κανονικές ιστοσελίδες. Όπως περιγράφηκε και στο κεφάλαιο 3, μία τέτοια μέθοδος είναι αρκετά συνηθισμένη, αφού η ιστοσελίδα αποσκοπεί στην εξαπάτηση μιας μηχανής αναζήτησης, έτσι ώστε η τελευταία να την θεωρήσει ως κανονική.

Για την αντιμετώπιση των παραπάνω προβλημάτων, είναι προφανές ότι θα πρέπει να γίνει χρήση κάποιας μεθόδου διαφορετικής από την τυφλή χρήση της συνάρτησης Ο. Για να το επιτύχει αυτό, ο αλγόριθμος TrustRank χρησιμοποιεί μία συνάρτηση πιθανότητας, η οποία καλείται *«συνάρτηση εμπιστοσύνης» (trust function)* Τ και η οποία παίρνει τιμές στο διάστημα [0,1] και δίνει την πιθανότητα μια

ιστοσελίδα $p \in V$ να είναι κανονική (T(p) = 1) ή να χρησιμοποιεί μεθόδους spam (T(p) = 0).



**Εικόνα 4.2**[20]: μέρος του πραγματικού Διαδικτυακού γράφου, όπου οι μαύροι κόμβοι αποτελούν spam ιστοσελίδες.

### 4.2.2 Ιδιότητες της συνάρτησης εμπιστοσύνης

Μαθηματικά, η συνάρτηση εμπιστοσύνης T ορίζεται ως εξής:

$$T(p) = Pr[O(p) = 1]$$

[20] Luca Becchetti, Web Spam Detection: link-based and content-based techniques, σελ. 3, Figure 1, διαθέσιμο στο: http://www.chato.cl/papers/becchetti_2008_link_spam_techniques.pdf

Για να εξηγήσουμε περισσότερο την χρήση της, θα δώσουμε ένα παράδειγμα. Ας θεωρήσουμε έναν διαδικτυακό γράφο V με 100 κόμβους. Έστω ότι η συνάρτηση επιστοσύνης T δίνει τιμή 0.7 για κάθε έναν από τους 100 κόμβους. Αυτό πρακτικά σημαίνει ότι κάθε κόμβος έχει πιθανότητα 70% να είναι κανονικός και 30% να χρησιμοποιεί κάποια μέθοδο spam. Αν από τους 70 από τους 100 αυτούς κόμβους ισχύει ότι O(p) = 1 και για τους υπόλοιπους 30 ισχύει ότι O(p) = 0, τότε η συνάρτηση εμπιστοσύνης T έχει ποσοστό επιτυχίας 100%.

Σε πραγματικές συνθήκες, είναι πολύ δύσκολο να κατασκευάσουμε μία συνάρτηση εμπιστοσύνης με τόσο μεγάλο ποσοστό επιτυχίας. Παρά το γεγονός αυτό, αφού η συνάρτηση T μας δίνει την πιθανότητα μία ιστοσελίδα να είναι κανονική, μπορούμε να την χρησιμοποιήσουμε για να διατάξουμε τις ιστοσελίδες του διαδικτυακού γράφου, ανάλογα με την πιθανότητα αυτή. Έτσι για παράδειγμα για δύο ιστοσελίδες p και q, αν ισχύει ότι T(p) > T(q), τότε η ιστοσελίδα p έχει μεγαλύτερες πιθανότητες να είναι κανονική από την q και συνεπώς η μηχανή αναζήτησης μπορεί να την εμφανίσει σε καλύτερο σημείο στα αποτελέσματα απ'ότι θα εμφανίσει την q. Έτσι λοιπόν, μια απαραίτητη ιδιότητα της συνάρτησης εμπιστοσύνης, είναι η ακόλουθη:

$$T(p) < T(q) \Leftrightarrow Pr[O(p) = 1] < Pr[O(q) = 1]$$
$$T(p) = T(q) \Leftrightarrow Pr[O(p) = 1] = Pr[O(q) = 1]$$

Βέβαια τα αποτελέσματα της σύγκρισης δεν αποτελούν σημαντικό παράγοντα για την ταξινόμηση των ιστοσελίδων σε φθίνουσα σειρά ανάλογα με την πιθανότητά τους να μην χρησιμοποιούν κάποια μέθοδο spam. Ας φανταστούμε την περίπτωση ότι η καλύτερη ιστοσελίδα λαμβάνει τιμή $T(p_{best}) = 0.25$ και η χειρότερη $T(p_{worst}) = 0.05$. Παρά το γεγονός ότι μπορεί να εφαρμοστεί ταξινόμηση στα αποτελέσματα αυτά, όλες οι ιστοσελίδες εμφανίζουν μεγάλη πιθανότητα να χρησιμοποιούν κάποια spam μέθοδο. Για να αποφύγουμε προβλήματα αυτού του είδους, συνήθως ορίζεται μία πιθανότητα κατωφλίου δ (threshold probability), την οποία αν η τιμή της συνάρτησης εμπιστοσύνης μιας ιστοσελίδας ξεπερνά, μπορούμε να θεωρήσουμε ότι η συγκεκριμένη ιστοσελίδα είναι κανονική:

$$T(p) > \delta \Leftrightarrow O(p) = 1$$

### 4.2.3 Ο υπολογισμός της συνάρτησης εμπιστοσύνης

Στο σημείο αυτό θα πρέπει να δείξουμε το πως είναι δυνατός ο υπολογισμός μιας συνάρτησης εμπιστοσύνης Τ. Για να γίνει καταννοητός ο εν λόγω υπολογισμός, θα αρχίσουμε την περιγραφή με μερικές απλές προσεγγίσεις. Έστω λοιπόν ότι $L$ είναι το σύνολο με όλους τους κόμβους (ιστοσελίδες) του διαδικτυακού γράφου. Έστω ότι αρχικά επιλέγουμε ένα υποσύνολο $S$ του $L$ και κάποιος άνθρωπος βαθμολογεί τις ιστοσελίδες του $S$ με χρήση της συνάρτησης Ο. Συμβολίζουμε με $S^+$ το υποσύνολο του $S$ που περιέχει τις κανονικές ιστοσελίδες (δηλαδή αυτές για τις οποίες ισχύει O(p) = 1) και με με $S^-$ το υποσύνολο του $S$ που περιέχει ιστοσελίδες που χρησιμοποιούν κάποια spam τεχνική (δηλαδή αυτές για τις οποίες ισχύει O(p) = 1). Για τις υπόλοιπες σελίδες του συνόλου $L$ που δεν ελέγχθηκαν (δηλαδή για αυτές που ανήκουν στο σύνολο $L \setminus S$) τους δίνουμε την τιμή πιθανότητας 1/2 (50% δηλαδή να αποτελούν κανονική ιστοσελίδα και 50% όχι). Όλα τα παραπάνω, αποτυπώνονται μαθηματικά στην παρακάτω συνάρτηση εμπιστοσύνης $T_0$:

$$T_0(p) = \begin{cases} O(p) & \alpha\nu\, p \in S \\ \frac{1}{2} & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Αν για παράδειγμα στον γράφο της Εικόνας 4.1 θεωρήσουμε ότι S = {1,3,6}, τότε τα παρακάτω διανύσματα $o$ και $t_0$ αναπαριστούν τις τιμές της συνάρτησης Ο και $T_0$ αντίστοιχα για κάθε ιστοσελίδα:

$$o = [1, 1, 1, 1, 0, 0, 0]$$
$$t_0 = [1, \frac{1}{2}, 1, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}]$$

Όπως είναι φυσικό, οι τιμές της συνάρτησης $T_0$ για τις ιστοσελίδες που δεν ανήκουν στο σύνολο $S$ δεν μπορούν να παραμείνουν ως έχουν. Για τον υπολογισμό των βαθμολογιών των συγκεκριμένων ιστοσελίδων, θα χρησιμοποιήσουμε την ιδιότητα της κλειστότητας κατά προσέγγιση του συνόλου των ιστοσελίδων που δεν χρησιμοποιούν spam μεθόδους.

Αρχικά, όπως και για τον υπολογισμό της συνάρτησης εμπιστοσύνης $T_0$, επιλέγουμε ένα υποσύνολο $S$ του $L$ και κάποιος άνθρωπος εφαρμόζει σε αυτό την συνάρτηση Ο. Όσον αφορά τα υπόλοιπα στοιχεία του συνόλου $L$, θεωρώντας ότι οι κανονικές ιστοσελίδες περιέχουν συνδέσμους μόνο σε κανονικές ιστοσελίδες, τότε δίνουμε την τιμή 1 για όσες ιστοσελίδες υπάρχει μονοπάτι μήκους Μ (μέσω ακολουθίας Μ συνδέσμων) από την αρχική ιστοσελίδα προς αυτές. Στις ιστοσελίδες που απομένουν, τους προσδίδουμε την τιμή 1/2 (αβεβαιότητα – 50% να αποτελούν κανονική ιστοσελίδα και 50% όχι). Η παραπάνω συνάρτηση εμπιστοσύνης συμβολίζεται ως $T_M$ και ο μαθηματικός της ορισμός δίνεται παρακάτω:

$$T_M(p) = \begin{cases} O(p) & \text{αν } p \in S \\ 1 & \text{αν } p \notin S \text{ και } \exists\, q \in S^+ : q \rightsquigarrow_M p \\ \frac{1}{2} & \text{αλλιώς} \end{cases}$$

όπου ο συμβολισμός $q \rightsquigarrow_M p$ δηλώνει ότι υπάρχει ακολουθία Μ συνδέσμων που ξεκινούν από την ιστοσελίδα $q$ και καταλήγουν στην ιστοσελίδα $p$. Ένα τέτοιο μονοπάτι δεν θα πρέπει να περιλαμβάνει spam ιστοσελίδες.

Πιο συγκεκριμένα, για τον γράφο της Εικόνας 4.1 και θεωρώντας ότι S = {1,3,6}, παίρνουμε τις παρακάτω βαθμολογίες για τρεις διαφορετικές τιμές του Μ:

$$M = 1, \quad t_1 = [1, 1, 1, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}]$$

$$M = 2, \quad t_1 = [1, 1, 1, 1\ \frac{1}{2}, 0, \frac{1}{2}]$$

$$M = 3, \quad t_1 = [1, 1, 1, 1, 1, 0, \frac{1}{2}]$$

Ενώ θα περιμέναμε η συνάρτηση εμπιστοσύνης $T_M$ να δουλεύει καλύτερα από την συνάρτηση $T_0$, παρατηρούμε ότι για Μ = 3, η ιστοσελίδα 5 παίρνει βαθμολογία 1, αφού βρίσκεται σε απόσταση 3 βημάτων από την ιστοσελίδα 1, η οποία ανήκει στο σύνολο $S^+$. Στην περίπτωση όπου Μ = 4, το πρόβλημα αυξάνει ακόμη περισσότερο, αφού και η ιστοσελίδα 7 θα είχε 1 ως τιμή της συνάρτησης.

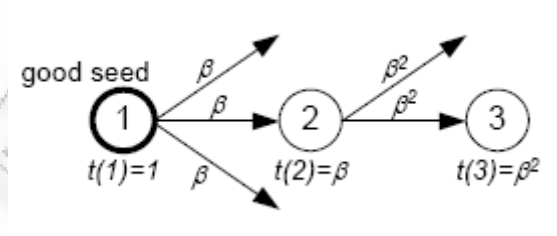
Η αιτία του προβλήματος είναι το γεγονός του ότι δεν μπορούμε να είμαστε σίγουροι ότι μία κανονική ιστοσελίδα θα οδηγεί και σε άλλες κανονικές ιστοσελίδες, σε μεγάλο βάθος του διαδικτυακού γράφου. Για να αντιμετωπίσουμε το συγκεκριμένο πρόβλημα, θα πρέπει να θεωρήσουμε ότι η πιθανότητα μία κανονική ιστοσελίδα να οδηγεί μέσω συνδέσμων σε άλλες κανονικές, μειώνεται όσο αυξάνεται το βάθος του μονοπατιού.

Για να αντιμετωπίσουμε το συγκεκριμένο πρόβλημα, μπορούμε να εφαρμόσουμε μία από τις παρακάτω δύο τεχνικές:

- Τεχνική της μείωσης της εμπιστοσύνης (trust dampening)
- Τεχνική της διάσπασης της εμπιστοσύνης (trust splitting)

**Τεχνική της μείωσης της εμπιστοσύνης (Trust Dampening)**

Σύμφωνα με την συγκεκριμένη μέθοδο, όπως και πριν, οι αρχικοί κανονικοί κόμβοι, παίρνουν την τιμή 1. Στην συνέχεια, κάθε κόμβος $p$, παίρνει ως τιμή της συνάρτησης την τιμή $\beta^i$, όπου $0 \leq \beta \leq 1$ μία σταθερά, η οποία ονομάζεται *παράγοντας μείωσης εμπιστοσύνης (damp factor)* και $i$ το μήκος του μονοπατιού από τον αρχικό κανονικό κόμβο στον κόμβο $p$. Η διαδικασία αυτή αναπαρίσταται σχηματικά στην Εικόνα 4.3.




**Εικόνα 4.3**[21]: παράδειγμα εφαρμογής της μεθόδου μείωσης της εμπιστοσύνης (trust dampening)

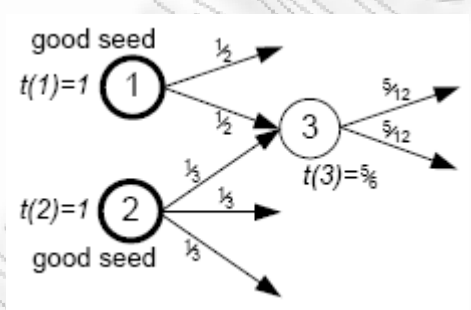
**Τεχνική της διάσπασης της εμπιστοσύνης (trust splitting)**

Η δεύτερη τεχνική η οποία χρησιμοποιείται, στηρίζεται στην εξής παρατήρηση/παραδοχή: η φροντίδα με την οποία ο διαχειριστής της ιστοσελίδας εισάγει τους συνδέσμους σε αυτή, είναι ανιστρόφως ανάλογη του αριθμού των συνδέσμων. Έτσι για παράδειγμα, αν μία ιστοσελίδα περιέχει λίγους συνδέσμους,

[21] Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen, "Combating web spam with trustrank", proceedings of the Thirtieth international conference on Very large data bases - Volume 30, σελ. 580, Figure 3

θεωρούμε ότι ο δημιουργός της έδωσε ιδιαίτερη προσοχή στους συνδέσμους αυτούς (συνεπώς έλεγξε ότι οι ιστοσελίδες δεν χρησιμοποιούν κάποια spam μέθοδο), ενώ αν περιέχει πολλούς συνδέσμους, θεωρούμε ότι η προσοχή που έδωσε ο δημιουργός της μειώθηκε.

Η παραπάνω παραδοχή μας οδηγεί στην εξής τεχνική: αν μία ιστοσελίδα $p$ έχει τιμή εμπιστοσύνης $T(p)$ και περιέχει $\omega(p)$ συνδέσμους σε άλλες ιστοσελίδες, τότε συνεισφέρει στον υπολογισμό της τιμής εμπιστοσύνης κάθε μιας από τις ιστοσελίδες αυτές την τιμή $\frac{T(p)}{\omega(p)}$. Η τιμή εμπιστοσύνης κάθε ιστοσελίδας υπολογίζεται ως άθροισμα των τιμών αυτών (δηλαδή το άθροισμα των τιμών που δίνουν οι ιστοσελίδες που περιέχουν σύνδεσμο στην εν λόγω ιστοσελίδα). Ένα παράδειγμα της μεθόδου παρουσιάζεται στην Εικόνα 4.4. Στην εν λόγω εικόνα, η τιμή εμπιστοσύνης του κόμβου 3, είναι το άθροισμα των τιμών που του προσδίδουν οι κόμβοι 1 και 2 ($t(3) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$).




**Εικόνα 4.4**[22]: παράδειγμα εφαρμογής της μεθόδου διάσπασης της εμπιστοσύνης (trust splitting)

[22] Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen, "Combating web spam with trustrank", proceedings of the Thirtieth international conference on Very large data bases - Volume 30, σελ. 580, Figure 4

### 4.2.4 Ο αλγόριθμος TrustRank

Με βάση όλα τα παραπάνω, μπορούμε πλέον να ορίσουμε τον αλγόριθμο TrustRank (μία έκδοση αυτού). Ο αλγόριθμος TrustRank παρουσιάζεται στον Πίνακα 4.1.

Αλγόριθμος TrustRank

είσοδος

T πίνακας γειτνίασης του διαδικτυακού γράφου

N αριθμός ιστοσελίδων

L αριθμός των ιστοσελίδων για τις οποίες θα υπολογιστεί αρχικά η συνάρτησης O

$\alpha_B$ παράγοντας μείωσης εμπιστοσύνης

$M_B$ αριθμός των βημάτων του αλγορίθμου

έξοδος

t* βαθμολογίες των ιστοσελίδων

αρχή

(1)

// Δημιούργησε το σύνολο s, το οποίο περιέχει τις αρχικές επιθυμητές ιστοσελίδες, πάνω στις οποίες

// θα εφαρμοστεί η συνάρτηση O

s = SelectSeed(…)

(2)

// Ταξινόμησε το σύνολο αυτό, ανάλογα με την προτίμηση

σ = Rank({1, . . . ,N}, s)

(3)

// Υπολόγισε τις τιμές τις συνάρτησης O για το σύνολο σ

d = $0_N$

for (i = 0; i < L; i++)

if (O(σ(i)) == 1) then

d(σ(i)) = 1

(4)

// Κανονικοποίησε τον πίνακα d

d = d/|d|

(5)

// Υπολογισμός του πίνακα βαθμολογιών t*

t* = d

for (i = 0; i < $M_B$; i++)

t* = $a_B \cdot T \cdot t^* + (1 - a_B) \cdot d$

return t*

τέλος

**Πίνακας 4.1**: ο αλγόριθμος TrustRank

## 4.3 Άλλοι μέθοδοι αντιμετώπισης του web spam

Αν και η μέθοδος TrustRank είναι αρκετά βασική και δημοφιλής, δεν αποτελεί τον μοναδικό τρόπο αντιμετώπισης του web spam. Η Google χρησιμοποιεί μία πληθώρα από άλλες μεθόδους, οι οποίες βοηθούν στην αποδοτικότερη ανακάλυψη spam ιστοσελίδων. Στην συγκεκριμένη ενότητα, παρουσιάζουμε τις περισσότερες βασικές από τις μεθόδους αυτές.

**Σύνταξή της ιστοσελίδας**

Πολλές από τις μηχανές αναζήτησης χρησιμοποιούν μεθόδους ανάλυσης του κειμένου που περιέχει μια ιστοσελίδα. Αν το κείμενο της ιστοσελίδας δεν είναι καλά δομημένο και έχει πολλά συντακτικά λάθη, τότε είναι πολύ πιθανό η συγκεκριμένη ιστοσελίδα να χρησιμοποιεί κάποια spam μέθοδο. Η ανάλυση ενός κειμένου, απαιτεί την χρήση ειδικών αλγορίθμων ανάλυσης φυσικής γλώσσας και διαφέρει σημαντικά από γλώσσα σε γλώσσα (για παράδειγμα σε μερικές γλώσσες η ανάλυση ενός κειμένου είναι περισσότερο δύσκολη απ'ότι σε άλλες). Εκτός από την σύνταξη, σημαντικό ρόλο παίζει και το πλήθος των λέξεων μιας ιστοσελίδας. Αν οι λέξεις είναι πολλές, τότε η ιστοσελίδα είναι υποψήφια για χρήση κάποιας spam μεθόδου και θα πρέπει να ελεγχθεί περαιτέρω.

Φυσικά, υπάρχει η περίπτωση του «έξυπνου» keyword stuffing, όπου ο δημιουργός της spam ιστοσελίδας εισάγει μία πληθώρα κλειδιών αναζήτησης με τρόπο τέτοιο, έτσι ώστε το κείμενο να βγάζει αρκετό νόημα και να είναι σωστό συντακτικά. Για την αντιμετώπιση της συγκεκριμένης τεχνικής, μία μηχανή αναζήτησης μπορεί να χρησιμοποιήσει κάποια μέθοδο μηχανικής μάθησης (π.χ. νευρωνικά δίκτυα). Η λογική είναι η ακόλουθη: ο αλγόριθμος μηχανικής μάθησης αρχικά εκπαιδεύεται με την χρήση κάποιου συνόλου εκπαίδευσης, το οποίο περιέχει spam ιστοσελίδες που χρησιμοποιούν την παραπάνω μέθοδο. Ο αλγόριθμος είναι φτιαγμένος με τέτοιον τρόπο, έτσι ώστε με την διαδικασία της εκπαίδευσης μπορεί να μεταβάλλει τα εσωτερικά του χαρακτηριστικά και να είναι σε θέση να αναγνωρίσει νέες, παρόμοιες με το σύνολο εκπαίδευσης, spam ιστοσελίδες. ΄

Το σύνολο εκπαίδευσης θα μπορούσε να διαφέρει από την απλή συλλογή κάποιων spam ιστοσελίδων και θα μπορούσε να χρησιμοποιεί μεθόδους εκτίμησης της συμπεριφοράς των χρηστών που επισκέπτονται μια ιστοσελίδα [18]. Σύμφωνα με την εν λόγω προσέγγιση, η μηχανή αναζήτησης προσπαθεί να προσομοιώσει τις

επιλογές ενός κανονικού χρήστη, όταν αυτός επισκεφθεί μία spam ιστοσελίδα. Στην πιο συνηθισμένη περίπτωση, ο χρήστης θα φύγει από την συγκεκριμένη ιστοσελίδα, αφού αντιληφθεί ότι πρόκειται για spam. Η προσομοίωση αυτή δεν είναι πάντοτε εύκολη και η συγκεκριμένη μέθοδος χρησιμοποιεί αρκετές παραδοχές για την εφαρμογή της.

**Ποσοστό λέξεων που αποτελούν συνδέσμους**

Μία άλλη, σχετικά προφανής κι αυτή μέθοδος, είναι η εκτίμηση του ποσοστού των λέξεων που χρησιμοποιούνται σε συνδέσμους, σε σχέση με το σύνολο των λέξεων που παρατίθενται στο κείμενο. Η τεχνική είναι χρήσιμη για την ανακάλυψη ιστοσελίδων που περιέχουν μία πληθώρα συνδέσμων που οδηγούν σε έναν και συγκεκριμένο ιστότοπο. Φυσικά η μέθοδος αυτή θα πρέπει να δίνει ως αποτέλεσμα τις ύποπτες για spam ιστοσελίδες, αφού κάποιες ιστοσελίδες μπορεί να περιέχουν μεγάλο αριθμό συνδέσμων και να είναι απόλυτα νόμιμες (π.χ. web directories).

**Ποσοστό του ορατού μέρους μιας ιστοσελίδας**

Πολλές μέθοδοι spam (εισαγωγή λέξεων με ίδιο χρώμα με το φόντο της ιστοσελίδας, εισαγωγή μη ορατών εικόνων που αποτελούν συνδέσμους σε ιστοσελίδες) εισάγουν σε μία ιστοσελίδα μη ορατά στοιχεία, τα οποία δεν γίνονται αντιληπτά σε έναν χρήστη μέσω του φυλλομετρητή του. Μία μέθοδος καταπολέμησης του spam είναι η εξέταση του ποσοστού των λέξεων της ιστοσελίδας που είναι ορατές (κάτι τέτοιο θα μπορούσε να υπολογιστεί ως το κλάσμα αριθμός μη ορατών λέξεων/συνολικός αριθμός). Αν το ποσοστό αυτό ξεπερνά ένα συγκεκριμένο όριο (π.χ. 20%) τότε μπορούμε να υποθέσουμε ότι η συγκεκριμένη ιστοσελίδα χρησιμοποιεί κάποια spam μέθοδο. Η εφαρμογή της συγκεκριμένης τεχνικής είναι περισσότερο δύσκολη, όταν αφορά την ανακάλυψη του ποσοστού των μη ορατών εικόνων της ιστοσελίδας, αφού θα απαιτούσε την χρήση ειδικών αλγορίθμων επεξεργασίας εικόνας.

**Διατήρηση ιστορικών στοιχείων**

Μία γνωστή επίσης μέθοδος είναι η διατήρηση, η εξέταση και η ανάλυση ιστορικών στοιχείων ιστοσελίδων. Τα ιστορικά αυτά στοιχεία αφορούν προηγούμενες αποθηκευμένες εκδόσεις ιστοσελίδων από κάποιον web crawler. Κατά τακτά χρονικά διαστήματα, οι νέες εκδόσεις των ιστοσελίδων που αποθηκεύει ο web crawler της μηχανής αναζήτησης, ελέγχονται με τις προηγούμενες. Αν παρατηρηθεί ιδιαίτερα μεγάλη αύξηση τους συνδέσμους τους οποίους περιέχει η ιστοσελίδα, τότε η μηχανή αναζήτησης θεωρεί την συγκεκριμένη ιστοσελίδα ύποπτη για την χρήση κάποιας μεθόδου spam.

**Αναζήτηση σε επιπέδο διευθύνσεων**

Πολλές φορές η μηχανή αναζήτησης ελέγχει ιστοσελίδες οι οποίες ανήκουν στο ίδιο όνομα Διαδικτύου (domain name), καθώς αποτελεί συχνό φαινόμενο κάποιος να δημιούργησε μία link farm χρησιμοποιώντας ένα συγκεκριμένο όνομα Διαδικτύου (π.χ. χρησιμοποιώντας 100 διευθύνσεις spam_address01.spam-example.com, …. spam_address100.spam-example.com οι οποίες να περιέχουν συνδέσμους προς την ιστοσελίδα www.spam-example.com).

## 4.4 Τρόποι αντιμετώπισης του Click Fraud

Όπως αναφέρθηκε και στο κεφάλαιο 3, η απάτη του click fraud μπορεί να δημιουργήσει σοβαρά οικονομικά προβλήματα, τόσο στην ίδια την Google, όσο και στους πελάτες αυτής. Η απόδειξη ότι γίνεται χρήση της συγκεκριμένης τεχνικής είναι ένα δύσκολο πρόβλημα, αφού κανείς δεν μπορεί να καθορίσει με ασφάλεια για το ποιος βρίσκεται πίσω από έναν υπολογιστή και ποιες είναι οι προθέσεις του. Έτσι, το καλύτερο δυνατό το οποίο μπορεί να κάνει η Google, είναι να εντοπίζει κάθε φορά τις επιλογές εκείνες των διαφημίσεων που *κατά πάσα πιθανότητα* αποτελούν προϊόν απάτης και να μην χρεώνει για τις επιλογές τον αντίστοιχο διαφημιζόμενο.

Όπως αναφέρεται και στο blog της Google, η ανακάλυψη του click fraud αποτελεί μία δύσκολη διαδικασία [15]. Οι περισσότερο συνηθισμένες τεχνικές για την αντιμετώπιση της εν λόγω απάτης παρουσιάζονται παρακάτω.

**Αυτοματοποιημένα φίλτρα άμεσης αναγνώρισης μη έγκυρων επιλογών**

Κάθε επιλογή μιας διαφήμισης εξετάζεται προσεκτικά από την Google, με την χρήση ειδικών φίλτρων, τα οποία αναλύουν μια πληθώρα παραμέτρων, όπως η IP διεύθυνση του χρήστη, την ώρα που έγινε επιλογή, τον αριθμό των επιλόγών που πραγματοποιήθηκαν από την εν λόγω IP διεύθυνση, κτλ. Η ανάλυση των παραπάνω δεδομένων δημιουργεί ένα «προφίλ» για την εν λόγω επιλογή και στην συνέχεια το προφίλ αυτό ελέγχεται με ήδη γνωστά πρότυπα, τα οποία αναφέρονται σε αποκλίνουσες συμπεριφορές. Αν το προφίλ παρουσιάζει μεγάλη ομοιότητα με κάποιο από τα πρότυπα αυτά, τότε η εν λόγω επιλογή δεν λαμβάνεται υπ'όψην και ο διαφημιζόμενος δεν χρεώνεται από αυτή. Επίσης η Google διατηρεί μία βάση δεδομένων, η οποία περιλαμβάνει IP διευθύνσεις, οι οποίες κατά πάσα πιθανότητα

χρησιμοποίησαν την εν λόγω απάτη. Έτσι, σε μελλοντικές επιλογές από τις εν λόγω διευθύνσεις, τα φίλτρα δεν εφαρμόζονται καν και η επιλογή απορρίπτεται εξ'αρχής. Ο έλεγχος στο στάδιο αυτό, πραγματοποιείται σε πραγματικό χρόνο, αμέσως μετά την ενεργοποίηση της επιλογής (click) από κάποιον χρήστη.

**Ανάλυση σε δεύτερο χρόνο (Offline Analysis)**

Η εφαρμογή των φίλτρων, όπως αναφέρθηκε, θα πρέπει να εκτελείται σε πραγματικό χρόνο, αμέσως μετά την επιλογή ενός χρήστη. Η χρονική εκτέλεση των φίλτρων θα πρέπει να είναι μικρή σχετικά σε χρόνο, έτσι ώστε να μην καθυστερείται η εύρυθμη λειτουργία των υπηρεσιών AdSense και AdWords. Για τον λόγο αυτό, η Google, αποθηκεύει τα αποτελέσματα του κάθε ελέγχου και τα εξετάζει με μεγαλύτερη λεπτομέρεια, εφαρμόζοντας ειδικούς αλγορίθμους αναγνώρισης προτύπων και εξόρυξης πληροφορίας [16].

Μία από τις περισσότερο χρησιμοποιούμενες μεθόδους, είναι αυτή της ανάλυσης των IP διευθύνσεων [16]. Για παράδειγμα, είναι ιδιαίτερα ύποπτο αν όλες οι επιλογές μιας διαφήμισης προέρχονται από την ίδια χώρα, πόσο μάλλον από την ίδια πόλη ή ακόμη περισσότερο από την ίδια IP διεύθυνση. Αυτό βέβαια δεν σημαίνει απαραίτητα ότι η συγκεκριμένες επιλογές αποτελούν κάποιο προϊόν απάτης, αλλά είναι σίγουρα κάτι που κινεί τις υποψίες και απαιτεί την διεξαγωγή περαιτέρω έρευνας.

Σε περίπτωση που ανακαλυφθεί κάποιο λάθος που πραγματοποιήθηκε στο πρώτο στάδιο (π.χ. χρέωση κάποιου διαφημιζόμενου λόγω μη ανακάλυψης της απάτης) τότε ο εν λόγω λογαριασμός του διαφημιζόμενου λαμβάνει μελλοντικά προτερήματα έναντι των άλλων (π.χ. μειώμενες χρεώσεις, δωρεάν τοποθέτηση διαφημίσεων σε ιστότοπους της Google, κτλ.).

**Ανάλυση από ειδικούς της Google (Google Team Analysis)**

Επειδή τα αυτοματοποιημένα εργαλεία δυστυχώς δεν εξασφαλίζουν πλήρη αξιοπιστία, η Google διατηρεί μία ομάδα από εξειδικευμένους τεχνικούς, οι οποίοι ερευνούν ειδικές περιπτώσεις (συνήθως μετά από αίτημα κάποιου διαφημιζόμενου είτε επειδή οι αλγόριθμοι των δύο παραπάνω επιπέδων δεν κατέληξαν σε κάποιο συμπέρασμα).

## 4.5 Τρόποι αντιμετώπισης της μεθόδου Phising

Όπως αναφέρθηκε και στο κεφάλαιο 3, η μέθοδος του Phising αποτελεί μία από τις περισσότερο δημοφιλείς μεθόδους απάτης που χρησιμοποιούνται και μπορεί να δημιουργήσει πολλά προβλήματα, όσον αφορά την φήμη και την αξιοπιστία της μηχανής αναζήτησης. Όπως και η ίδια η Google αναφέρει στο blog της [17], καθημερινώς εξετάζονται εκατομμύρια τέτοιες περιπτώσεις.

Η αρχική εξετάση η οποία γίνεται αφορά την διεύθυνση (URL) της ιστοσελίδας. Συνήθως, οι διευθύνσεις των ιστοσελίδων που χρησιμοποιούν κάποια τεχνική phising είναι μεγάλες σε μήκος και τις περισσότερες των περιπτώσεων περιλαμβάνουν φράσεις όπως "banking" και "login". Οι φράσεις αυτές καθιστούν τις ιστοσελίδες αυτές ύποπτες για την υποκλοπή στοιχείων των χρηστών τους.

Φυσικά η απλή εξετάση της διεύθυνσης δεν μπορεί να αποτελέσει από μόνη της κριτήριο αξιολόγησης μιας ιστοσελίδας, αφού οι εν λόγω λέξεις είναι ιδιαίτερα κοινές και χρησιμοποιούνται σε πολλές κανονικές ιστοσελίδες. Σε δεύτερο χρόνο λοιπόν, εξετάζεται προσεκτικά το περιεχόμενο της εν λόγω ιστοσελίδας. Αν η ιστοσελίδα περιέχει φόρμες εισαγωγής στοιχείων με λέξεις όπως "password" ή "PIN", τότε η ιστοσελίδα κρίνεται ύποπτη και θα πρέπει να εξεταστεί περαιτέρω. Μία άλλη επίσης μέθοδος είναι η γεωγραφική εξέταση του server όπου είναι αποθηκευμένη η ιστοσελίδα. Αν η ιστοσελίδα αφορά μία αμερικάνικη τράπεζα και ο server βρίσκεται τοποθετημένος σε κάποια άλλη χώρα, τότε είναι πολύ πιθανό ότι πρόκειται για απάτη. Παράλληλα, η Google ελέγχει και την βαθμολογία της ιστοσελίδας που δίνει ο αλγόριθμος PageRank. Αν η ιστοσελίδα δεν είναι δημοφιλής, τότε κατά πάσα πιθανότητα πρόκειται για μία ψεύτικη έκδοση της πραγματικής ιστοσελίδας. Τέλος, η Google ελέγχει και το διαδικτυακό όνομα της ιστοσελίδας, αφού έχει παρατηρηθεί ότι συνήθως τέτοιες ιστοσελίδες αποθηκεύονται σε ίδιους servers με αυτούς που αποθηκεύουν ιστοσελίδες spam.

Ο συνδυασμός των παραπάνω τεχνικών δημιουργεί ένα δίκτυ ασφαλείας, το οποίο παρουσιάζεται εξαιρετικές επιδόσεις. Σύμφωνα με την Google [17], οι παραπάνω μέθοδοι είναι σε θέση να αναγνωρίζουν 9 στις 10 ιστοσελίδες που χρησιμοποιούν τέτοιου είδους απάτη ενώ ταξινομούν λάθος (ως ιστοσελίδα phising) μόνο 1 στις 10.000 ιστοσελίδες.

# ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Edwards, J., McCurley, K. S., και Tomlin, J. A. (2001). "An adaptive model for optimizing performance of an incremental web crawler". In Proceedings of the Tenth Conference on World Wide Web (Hong Kong: Elsevier Science): 106–113. doi:10.1145/371920.371960.

[2] Lawrence, Steve; C. Lee Giles (1999-07-08). "Accessibility of information on the web". Nature 400 (6740): 107. doi:10.1038/21987.

[3] Cho, J. και Garcia-Molina, H. (2003). Effective page refresh policies for web crawlers. ACM Transactions on Database Systems, 28(4). Διαθέσιμο στο: http://oak.cs.ucla.edu/~cho/papers/cho-tods03.pdf , Ημερομηνία Καταγραφής: 27/01/2010

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze (2007), http://nlp.stanford.edu/IR-book/pdf/19web.pdf , Ημερομηνία Καταγραφής: 31/01/2010

[5] In-Ho Kang, Transactional Query Identification in Web Search, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.4753&rep=rep1&type=pdf , Ημερομηνία Καταγραφής: 31/01/2010

[6] Sergey Brin και Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Stanford University, 1996

[7] Randall Stross, Planet Google, New York: Free Press. σελ. 61, 2008

[8] Strassmann, Paul A., "A Model for the Systems Architecture of the Future.", 5 Δεκεμβρίου, 2005, Διαθέσιμο στο: http://www.strassmann.com/pubs/gmu/LectureV4.pdf, Ημερομηνία Καταγραφής: 24/02/2010

[9] Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen, "Combating web spam with trustrank", proceedings of the Thirtieth international conference on Very large data bases - Volume 30

[10] Allison Woodruff, Paul M. Aoki, Eric Brewer, Paul Gauthier, Lawrence A. Rowe, "An Investigation of Documents from the World Wide Web", Διαθέσιμο στο: http://www.paulaoki.com/papers/www5-color.pdf, Ημερομηνία Καταγραφής: 22/03/2010

[11] A new approach to China, Google Official Blog, http://googleblog.blogspot.com/2010/01/new-approach-to-china.html, Ημερομηνία Καταγραφής: 21/04/2010

[12] Phising Free, Google Official Blog, http://googleonlinesecurity.blogspot.com/2010/03/phishing-phree.html, Ημερομηνία Καταγραφής: 01/05/2010

[13] Darren W Chow, Google Pagerank Vs Google Trustrank, Διαθέσιμο στο: http://ezinearticles.com/?Google-Pagerank-Vs-Google-Trustrank&id=3168616, Ημερομηνία Καταγραφής: 06/05/2010

[14] Baoning Wu, Brian D. Davison, Cloaking and Redirection: A Preliminary Study, Διαθέσιμο στο: http://airweb.cse.lehigh.edu/2005/wu.pdf, Ημερομηνία Καταγραφής: 12/05/2010

[15] Using Data to help prevent fraud, Google Official Blog, http://googleblog.blogspot.com/2008/03/using-data-to-help-prevent-fraud.html, Ημερομηνία Καταγραφής: 13/05/2010

[16] Invalid Clicks – Google's Overall Numbers, Google Official Blog, http://adwords.blogspot.com/2007/02/invalid-clicks-googles-overall-numbers.html, Ημερομηνία Καταγραφής: 13/05/2010

[17] Phising Phree, Google Official Blog,, Διαθέσιμο στο: http://googleonlinesecurity.blogspot.com/2010/03/phishing-phree.html, Ημερομηνία Καταγραφής: 15/05/2010

[18] Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru, Identifying Web Spam with User Behavior Analysis, Διαθέσιμο στο: http://airweb.cse.lehigh.edu/2008/submissions/liu_2008_spam_user_behavior.pdf, Ημερομηνία Καταγραφής: 16/05/2010