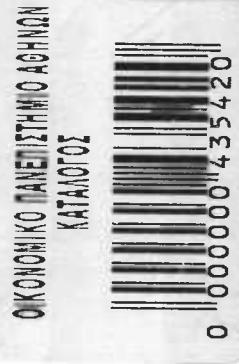




ΙΚΟΝΟΜΙ^Α
ΑΝΕΠΙΣΤΗΜ
ΑΘΗΝΩΝ
ΙΒΛΙΟΥ^Α
68656
004.678
ΔΡΑ

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc) στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

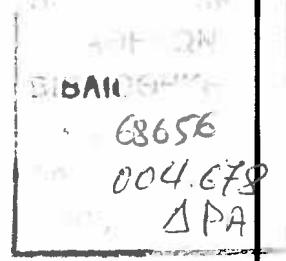
«Το Web ως γράφημα»

Δρατζίδης Κωνσταντίνος
M3990022

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2001



**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Το Web ως γράφημα»

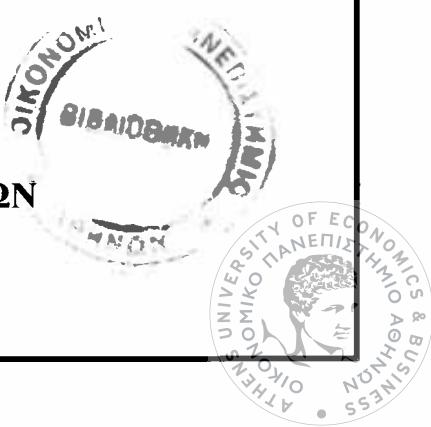
Δρατζίδης Κωνσταντίνος

M3990022

**Επιβλέπων Καθηγητής: Μάρθα Σιδέρη
Εξωτερικός Κριτής: Καθηγητής Γ. Πολύζος**

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2001



*Ευχαριστώ τους
Σεραφείμ Καραπάτη
Μάρθα Σιδέρη*



Πίνακας Περιεχομένων

Εισαγωγή.....	2
Συλλογή του ιστού	4
Τύποι Συλλεκτών ιστού	4
Συλλέκτες δικτυακού τόπου.....	4
Συλλέκτες γενικής χρήστης	4
Γενικές παρατηρήσεις	5
Περιγραφή AltaVista Search Engine	5
i. Συλλέκτης αρχείων και σελίδων	6
ii. Μετατροπέας Αρχείων	6
iii. Webindexer	6
Από τη συλλογή των δεδομένων στο γράφημα των συνδέσμων	7
AltaVista Search Engine και Connectivity Server	7
Υλοποίηση	7
Εσωτερική δομή του γραφήματος.....	8
Στατιστικά Στοιχεία	10
Γενικά στατιστικά στοιχεία γραφήματος	10
Κατανομές έσω-βαθμού και έξω-βαθμού	12
Αποτελέσματα.....	12
Πηγές πληροφορίας.....	14
Αλγόριθμος HITS.....	14
Μετρήσεις	16
Ισχυρές συνεκτικές συνιστώσες	19
Χαρακτηριστικά του αλγορίθμου	20
Βήματα	20
Κόστος υλοποίησης.....	20
Αποτελέσματα	21
Επίλογος – Συμπεράσματα.....	22
Βιβλιογραφία.....	24
Περίληψη	25
Executive Summary	27

Εισαγωγή

Η αναζήτηση πληροφορίας στον παγκόσμιο ιστό (World Wide Web) είναι ένα από τα πλέον ενδιαφέροντα και προκλητικά πεδία έρευνας. Η ανακάλυψη ενδιαφέρουσας πληροφορίας σχετικά με κάποιο θέμα ή ακόμα και η ακριβής στατιστική μελέτη των ιστοσελίδων είναι ζητήματα αρκετά δύσκολα λόγω της χαοτικής εξέλιξης του παγκόσμιου ιστού και της έλλειψης οργάνωσης αυτού. Καθημερινά δημιουργούνται περίπου ένα εκατομμύριο ιστοσελίδες για να προστεθούν στις εκατοντάδες εκατομμύρια ήδη υπάρχουσες. Ο τεράστιος αυτός όγκος πληροφορίας συνδέεται με περισσότερους από ένα δισεκατομμύριο συνδέσμους (hyperlinks). Το ερώτημα που γεννάται είναι πως μπορεί κάποιος να ανακτήσει σελίδες υψηλής ποιότητας που να σχετίζονται άμεσα με τις συγκεκριμένες ανάγκες του. Για το σκοπό αυτό οι χρήστες έχουν καταφύγει στις μηχανές αναζήτησης (search engines) [Altavista, Google, Infoseek, Yahoo]. Ο κλασικός τρόπος λειτουργίας μιας μηχανής αναζήτησης είναι η συντήρηση ενός ευρετηρίου (index) με τις ήδη γνωστές σελίδες για κάθε δυνατό ερώτημα (query) του χρήστη, η ανάκτηση πληροφορίας μέσω αυτού του ευρετηρίου και η διαβάθμιση των σελίδων μέσω ευρετικών κανόνων, τις περισσότερες φορές μη αποδοτικών. Προβλήματα όπως πως μια μηχανή αναζήτησης θα επιλέξει τις 20 «καλύτερες» σελίδες όταν το ερώτημα του χρήστη εμφανίζεται σε δεκάδες χιλιάδες σελίδες είναι εύλογα και δύσκολα να λυθούν αποτελεσματικά.

Πρόσφατα έχουν αναπτυχθεί νέες τεχνικές και αλγόριθμοι αναζήτησης πληροφορίας που βασίζονται στη δομή του Παγκοσμίου Ιστού παρά στο περιεχόμενο της κάθε ιστοσελίδας. Οι τεχνικές αυτές βασίζονται στην δομή γραφήματος που παρουσιάζει ο Παγκόσμιος Ιστός. Πράγματι, ο Παγκόσμιος Ιστός μπορεί να θεωρηθεί σαν ένα τεράστιο προσανατολισμένο γράφημα, όπου κάθε κόμβος του παριστάνει μια ιστοσελίδα και κάθε ακμή μεταξύ δύο κόμβων παριστάνει έναν σύνδεσμο από μια ιστοσελίδα σε μια άλλη. Το γράφημα αυτό είναι δυναμικό και εξελίσσεται καθημερινά καθώς νέες ιστοσελίδες δημιουργούνται, μεταβάλλονται ή παύουν να υπάρχουν [YBCFW00]. Ο τρόπος δημιουργίας και εξέλιξης του Παγκόσμιου Ιστού έχει σαν αποτέλεσμα το γράφημα αυτό, γνωστό ως *Γράφημα του Ιστού* (Web Graph),

να παρουσιάζει συγκεκριμένες δομές και ιδιότητες, οι οποίες μπορούν να αξιοποιηθούν από τα συστήματα ανάκτησης πληροφορίας. [Adamic99, KPR, Kleinberg00]. Για παράδειγμα, οι σύνδεσμοι μεταξύ των ιστοσελίδων παρέχουν σημαντική πληροφορία για το πόσο σχετίζονται οι ιστοσελίδες μεταξύ τους και για το πόσο σημαντική είναι μια ιστοσελίδα ή όχι.

Με βάση αυτή την υπόθεση, έχουν αναπτυχθεί αποδοτικές τεχνικές για την εύρεση «ποιοτικών» ιστοσελίδων του παγκόσμιου ιστού. Στην εργασία [KKRRT] δίνεται μια σύντομη περιγραφή της τεχνικής αυτής καθώς και τα μοντέλα και οι μετρικές που χρησιμοποιούνται. Η βασική μέθοδος που εφαρμόζεται στο γράφημα του Ιστού είναι η μέθοδος HITS του Kleinberg [Kleinberg97] Ο αλγόριθμος HITS είναι ένας αλγόριθμος αναζήτησης που σχεδιάστηκε για να εντοπίζει ιστοσελίδες υψηλής ποιότητας που σχετίζονται με το συγκεκριμένο θέμα (topic) που δίνει ως ερώτημα ένας χρήστης, χρησιμοποιώντας δειγματοληπτικά ένα τμήμα του γραφήματος.

Η εύρεση των συνεκτικών συνιστώσων εξερευνεί το γράφημα και ψάχνει για συνεκτικά κομμάτια του. Δίνει με αυτό τον τρόπο μια πιο μακροσκοπική εικόνα των δεδομένων από την απλή αναπαράσταση κόμβων και ακμών. Ενδιαφέρον παρουσιάζει η μορφή του γραφήματος των συνεκτικών συνιστώσων για το δηλαδή εάν τελικά υπάρχει κάποια δομή ή βγαίνει κάποιο συμπέρασμα.

Στόχος της εργασίας είναι η μελέτη και κατανόηση των τεχνικών αυτών, η υλοποίηση των αλγορίθμων με χρήση αποδοτικών δομών και η εφαρμογή τους στο Ελληνικό τμήμα του Παγκόσμιου ιστού (.gr).

Συλλογή του ιστού

Για να μπορέσουμε να μελετήσουμε τον ιστό ως γράφημα, θα πρέπει αρχικά να τον συλλέξουμε και έπειτα να κατασκευάσουμε το γράφημα των συνδέσμων. Στην κατασκευή του γραφήματος των συνδέσμων χρειάζεται μόνο η πληροφορία της διασύνδεσης και όχι και το περιεχόμενο των σελίδων. Αυτό μας δίνει μια αρχική απαίτηση για τον συλλέκτη. Στην κατασκευή του γραφήματος θεωρούμε ότι ένας κόμβος αντιπροσωπεύει μια σελίδα και μια ακμή ένα σύνδεσμο μεταξύ δύο σελίδων. Λόγω της υφής του προβλήματος, θα πρέπει να δημιουργήσουμε ένα προσανατολισμένο γράφημα. Αυτή είναι μια άλλη απαίτηση η οποία θα χρησιμοποιηθεί για την κατασκευή του γραφήματος από τη συλλογή.

Τύποι Συλλεκτών ιστού

Συλλέκτες δικτυακού τόπου

Σε αυτή τη κατηγορία ανήκουν οι συλλέκτες που είναι ικανοί να μας δώσουν πληροφορίες για ένα μόνο δικτυακό τόπο. Χρησιμοποιούνται συνήθως για ελέγχουν τη δομή ενός δικτυακού τόπου και να ανακαλύψουν σπασμένους συνδέσμους ή ορφανά αρχεία. Ο τρόπος λειτουργίας τους είναι απλός. Δίνεται ένα σημείο εκκίνησης της συλλογής και ο συλλέκτης, χρησιμοποιώντας μια στοίβα, διασχίζει όλο το δικτυακό τόπο. Η συνθήκη τερματισμού είναι το άδειασμα της στοίβας. Είναι, δε, συνήθως γραμμένοι σε scripting γλώσσα, όπως, για παράδειγμα, Perl και στη γλώσσα προγραμματισμού Java.

Συλλέκτες γενικής χρήσης

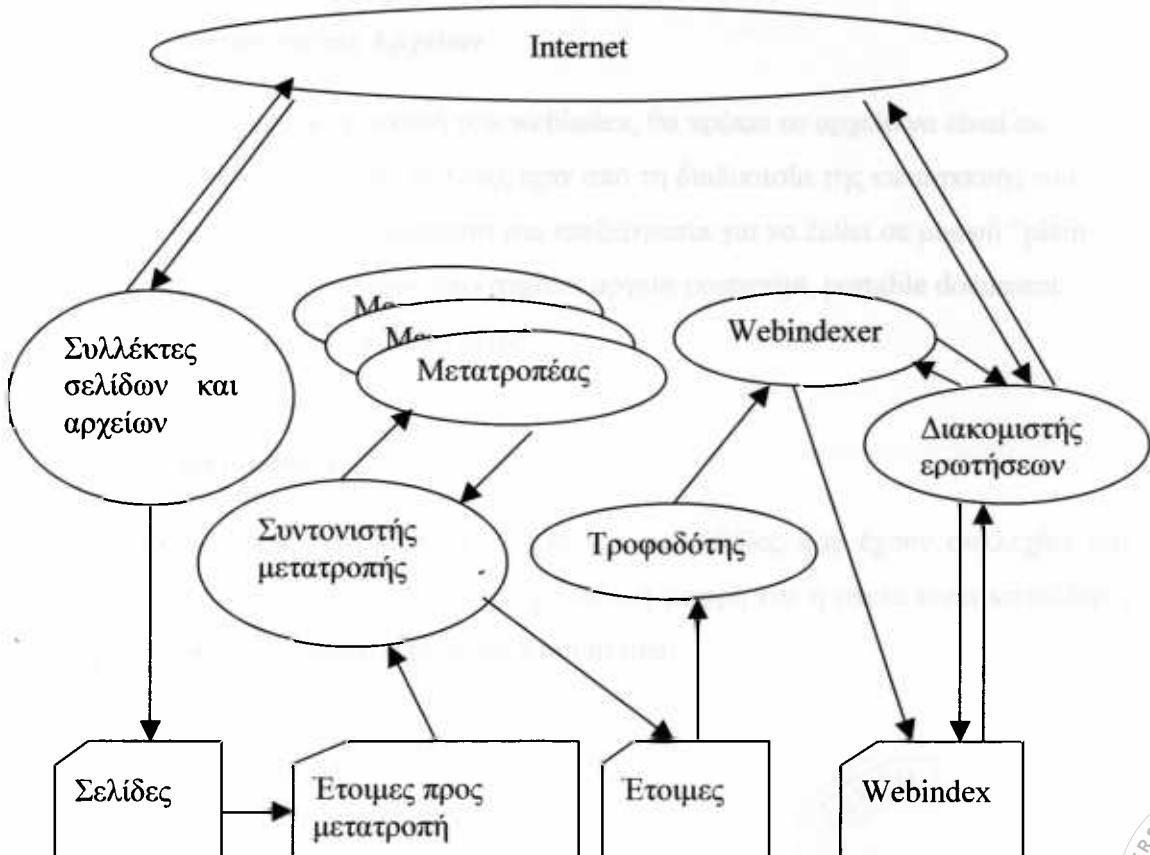
Σε αυτή τη κατηγορία ανήκουν οι συλλέκτες που είναι γενικής χρήσης. Με αυτούς είναι δυνατή αφενός η συλλογή του κειμένου των σελίδων και αφετέρου της συνδεσμολογίας ανάμεσα και σε διαφορετικούς δικτυακούς τόπους. Ο τρόπος υλοποίησης τους είναι με βάση τους agents και υπάρχει μια στοίβα η οποία είναι διαμοιρασμένη σε αυτούς. Η γλώσσα προγραμματισμού που χρησιμοποιείται συνήθως είναι η Java και η C++.

Γενικές παρατηρήσεις

Παρατηρούμε πως η Java είναι μια γλώσσα στην οποία έχει γραφτεί το μεγαλύτερο ποσοστό των συλλεκτών ιστού. Επίσης όλοι οι συλλέκτες λειτουργούν, κατά κύριο λόγο, με βάση τη στοίβα και στη πλειοψηφία τους χρησιμοποιούν πολυνηματική επεξεργασία. Όλοι οι τύποι συλλεκτών έχουν ειδική αρχιτεκτονική για την αποφυγή κύκλων και παγίδων κατά τη διάρκεια της σάρωσης.

Περιγραφή AltaVista Search Engine

Η AltaVista χρησιμοποιεί ως κεντρικό μηχανισμό συλλογής τον Mercator, ο οποίος έχει υλοποιηθεί σε Java. Ο Mercator δεν υπάρχει πουθενά ως αυτόνομο προϊόν, αλλά δίνεται μαζί με την AltaVista. Η AltaVista αποθηκεύει πληροφορία τόσο για το περιεχόμενο των σελίδων, όσο και για τη συνδεσμολογία μεταξύ τους. Στο παρακάτω σχήμα έχουμε το διάγραμμα λειτουργίας της AltaVista Search Engine. Αποτελείται από τους συλλέκτες αρχείων και σελίδων, από τους μετατροπείς αρχείων και από τον webindexer.



To Web ως γράφημα

i. Συλλέκτης αρχείων και σελίδων

Ο συλλέκτης αρχείων είναι υπεύθυνος για την εύρεση (find) και την ανάκτηση (retrieval) των αρχείων και σελίδων στο Internet. Στη διάλεκτο του Internet τα αντίστοιχα προγράμματα ονομάζονται “spiders”, “robots” ή “crawlers”. Για τη συλλογή αρχείων και σελίδων χρειάζονται δύο παράμετροι: μια λίστα αρχικών διευθύνσεων από τις οποίες θα αρχίσει η συλλογή και τις συνθήκες exclude και include, έτσι ώστε να γνωρίζει αν θα περιληφθεί μια σελίδα στη περαιτέρω συλλογή ή όχι [HM99].

Ένα άλλο χαρακτηριστικό ενός συλλέκτη είναι η τήρηση του “Robot Exclusion Standard”, σύμφωνα με το οποίο ένα “robot” πριν κάνει συλλογή αρχείων από ένα web site θα πρέπει να διαβάσει το ειδικό αρχείο που του παρέχει αυτό και μέσα στο οποίο αναγράφεται εάν έχει πρόσβαση και σε ποιες περιοχές του το εν λόγω “robot”. Εάν τηρηθεί αυτή η διαδικασία, τότε ο συλλέκτης συλλέγει μόνο τις σελίδες και τα αρχεία στα οποία έχει πρόσβαση. Η τήρηση των κανόνων του “Robot Exclusion Standard” είναι προαιρετική και μια τέτοια συμπεριφορά από τον συλλέκτη θεωρείται ως “polite”.

ii. Μετατροπέας Αρχείων

Για να γίνει η κατασκευή του webindex, θα πρέπει το αρχείο να είναι σε μορφή “plain text”. Για αυτό το λόγο, πριν από τη διαδικασία της κατασκευής του webindex, το κάθε αρχείο υφίσταται μια επεξεργασία για να έρθει σε μορφή “plain text”. Οι μετατροπέας αρχείων υποστηρίζει αρχεία postscript, portable document format (pdf) και αρχεία του MS Office.

iii. Webindexer

Ο webindexer αναλύει τα αρχεία και τις σελίδες που έχουν συλλεχθεί και κατασκευάζει το ευρετήριο το οποίο έχει ειδική μορφή και η οποία είναι κατάλληλη για ερωτήσεις που γίνονται από μέσω του Internet.



Από τη συλλογή των δεδομένων στο γράφημα των συνδέσμων

Μπορούμε να δούμε κάθε συλλογή V από συνδεδεμένες σελίδες ως ένα κατευθυνόμενο γράφημα $G = (V, E)$. Οι κόμβοι του γραφήματος αντιστοιχούν στις σελίδες και μια κατευθυνόμενη ακμή $(p, q) \in E$ σημαίνει ότι υπάρχει σύνδεσμος από τη σελίδα p στη σελίδα q . Ως έξω-βαθμό ενός κόμβου ορίζουμε το πλήθος των ακμών που δείχνει αυτός και ως έσω-βαθμό το πλήθος των ακμών που τον δείχνουν. Από ένα γράφημα G μπορούμε να απομονώσουμε τμήματά του, που ονομάζονται και υπογραφήματα. Έστω $W \subseteq V$ ένα υποσύνολο των σελίδων. Χρησιμοποιούμε τον συμβολισμό $G[W]$ για να δηλώσουμε το σύνολο που προκύπτει από το G έτσι ώστε κάθε κόμβος του συνόλου να ανήκει στο W και οι εν λόγω ακμές των κόμβων να ανήκουν και αυτές στο W .

AltaVista Search Engine και Connectivity Server

Με βάση αυτή τη πληροφορία από το webindex, ο Connectivity Server της Compaq, δημιουργεί το γράφημα των συνδέσμων το οποίο είναι προσανατολισμένο. Θα πρέπει να αναφέρουμε εδώ πως ο Connectivity Server δεν διατίθεται για εμπορικούς σκοπούς. Για να κατασκευάσει κάποιος το γράφημα από το ευρετήριο της AltaVista θα πρέπει να στείλει τα δεδομένα του στη Compaq.

Υλοποίηση

Για τη συλλογή των δεδομένων χρησιμοποιήθηκε το προϊόν της AltaVista, AltaVista Search Engine [AVSE]. Για κάθε νέα συλλογή δεδομένων απαιτούνται περίπου τρεις εβδομάδες μέχρι αυτή τη στιγμή υπάρχουν τρεις συλλογές έτοιμες και μια εν εξελίξει. Η πρώτη έγινε τον Μάρτιο, η δεύτερη τον Μάιο, η οποία αποδείχθηκε ότι για τεχνικούς λόγους ήταν ανακριβής, και η τρίτη τον Ιούλιο. Η τελευταία έχει αρχίσει από τα τέλη Οκτώβρη και είναι ακόμα σε εξέλιξη.

Για κάθε συλλογή πρέπει να καθοριστούν οι εξής παράμετροι :

- μια αρχική διεύθυνση από όπου θα αρχίσει το crawling
- ο τομέας (domain) ο οποίος θα εξερευνηθεί
- οι τύποι αρχείων που θα εισαχθούν στο webindex της AltaVista
- ο χρονοπρογραμματισμός του crawling.

Στην περίπτωση των τριών πρώτων συλλογών, αρχικές διευθύνσεις ήταν οι <http://www.in.gr>, <http://www.aueb.gr> και <http://www.ntua.gr>, ενώ σε αυτό που είναι σε εξέλιξη έχουν χρησιμοποιηθεί είκοσι αρχικές διευθύνσεις. Ο τομέας και στις τρεις περιπτώσεις ήταν ο ελληνικός (.gr) και οι τύποι αρχείων που χρησιμοποιούνται είναι μόνο web αρχεία (htm, html, asp, jhtml, xml, dhtml, cgi, php). Οι σελίδες που έγιναν αρχειοθετήθηκαν ήταν περίπου δύο εκατομμύρια σε όλες τις περιπτώσεις.

Μετά από τη συλλογή και την αρχειοθέτηση της AltaVista Search Engine, γίνεται μεταφορά των αποτελεσμάτων σε μορφή γραφήματος. Για αυτό το λόγο χρησιμοποιείται το AltaVista Search Engine Development Kit το οποίο αποτελείται από βιβλιοθήκες C και επιτρέπει πρόσβαση στο ευρετήριο (webindex) που έχει ήδη δημιουργηθεί. Η μεταφορά των αποτελεσμάτων γίνεται με ερωτήσεις (queries) για τα backlinks μιας σελίδας. Με τον όρο backlinks εννοούμε το σύνολο των σελίδων που δείχνουν την συγκεκριμένη σελίδα. Με αυτό τον τρόπο κατασκευάζουμε το γράφημα αντίστροφα, βασιζόμενοι στα backlinks και όχι απευθείας στους συνδέσμους (links) μιας σελίδας.

Μετά από την εξαγωγή των αποτελεσμάτων από το ευρετήριο της AltaVista γίνεται η κατασκευή δομών δεδομένων για γραφήματα. Σε αυτό το σημείο της διαδικασίας χρησιμοποιείται η LEDA, η οποία περιέχει τα κατάλληλα αντικείμενα για την κατασκευή των δικών μας δομών δεδομένων.

Εσωτερική δομή των γραφήματος

Όπως αναφέρθηκε κάθε σελίδα (κόμβος του γραφήματος) αντιπροσωπεύεται από ένα μοναδικό αριθμητικό αναγνωριστικό. Οι ακμές συνιστούν ζεύγη τέτοιων κόμβων. Το σύστημα κρατά για κάθε κόμβο τη λίστα των γειτόνων του, δηλαδή τους κόμβους στους οποίους καταλήγουν οι ακμές του (*outlinks*). Επίσης κρατά σε μια

άλλη λίστα τους κόμβους οι ακμές των οποίων καταλήγουν σε έναν άλλον (*inlinks*). Η δομή του φυλάσσονται όλα αυτά είναι των πινάκων κατακερματισμού (hash tables). Κλειδί αποτελεί το μοναδικό αριθμητικό κάθε κόμβου και πληροφορία η λίστα των γειτόνων του. Για τις εισερχόμενες ακμές έχουμε το ίδιο γράφημα αλλά σε ανάστροφη μορφή.

Η εσωτερική αναπαράσταση του γραφήματος στη LEDA έχει τη μορφή:

```
h_array <long, list<long> > webgraph  
h_array <long, list<long> > revwebgraph
```

δηλαδή χρησιμοποιούμε δυο πίνακες κατακερματισμού για το γράφημα του διαδικτύου. Η μορφή των πινάκων επιτρέπει σταθερή σε χρόνο πρόσβαση στους κόμβους και γραμμική αναζήτηση ακμών.



Στατιστικά Στοιχεία

Γενικά στατιστικά στοιχεία γραφήματος

Ένα από τα πρώτα στατιστικά στοιχεία για τον ιστό είναι το μέγεθός του σε κόμβους και ακμές. Ο ιστός στην Ελλάδα έχει τα ακόλουθα στοιχεία :

Ακμές	7.724.508
Κόμβοι	1.119.029
Ακμές ανά Κόμβο	6,9

Από αυτά τα πρώτα στοιχεία βλέπουμε ότι το γράφημα στην Ελλάδα είναι αραιό. Οι σελίδες έχουν υποστεί επεξεργασία και για αυτό το λόγο ο αριθμός τους δεν είναι τα 2.000.000 που έχει βάλει στο ευρετήριό της η AltaVista. Η επεξεργασία έγινε γιατί στα web έγγραφα η AltaVista περιλαμβάνει και μη στατικές σελίδες (π.χ. cgi, jhtml κλπ). Επιπλέον μπορούμε να φίλτραρουμε το γράφημα και να σβήσουμε από το γράφημα τις ακμές οι οποίες είναι για πλοήγηση μέσα στον ίδιο δικτυακό τόπο. Σαν δεύτερο βήμα σβήνουμε όλους τους κόμβους με έσω-βαθμό και έξω βαθμό 0. Μετά από αυτό το φίλτρο τα στοιχεία αλλάζουν ως εξής :

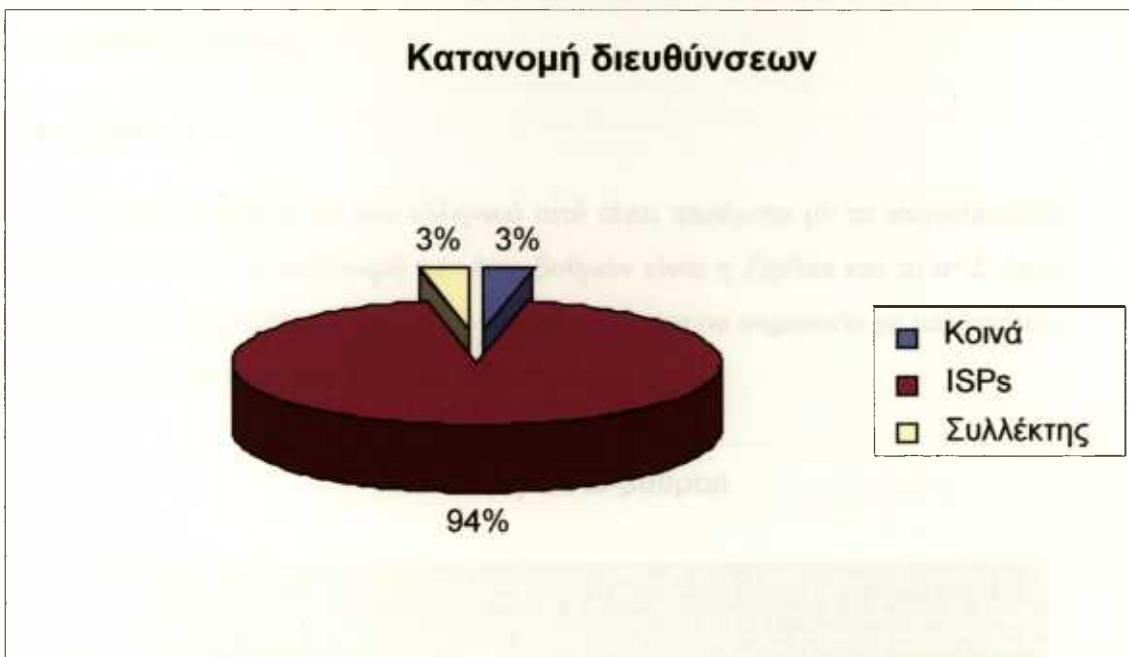
Ακμές	3.708.619
Κόμβοι	857.201
Ακμές ανά Κόμβο	4,3

Παρατηρούμε μια μεγάλη μείωση του αριθμού των ακμών στο ποσοστό του 52% και μια μικρότερη μείωση του αριθμού των κόμβων στο 76,6% του αρχικού. Αυτό μας δίνει μια πρώτη εικόνα για την διασύνδεση των δικτυακών τόπων στην Ελλάδα.

Θα μπορούσε να περάσει το γράφημα και από ένα άλλο φίλτρο. Σε μια σελίδα μπορεί να υπάρχει μια διαφήμιση η οποία ανήκει σε άλλο δικτυακό τόπο. Συνήθως οι σελίδες που αποτελούν διαφήμιση έχουν μεγάλο έσω-βαθμό. Το φίλτρο θα μπορούσε να αφαιρεί με τυχαίο τρόπο ακμές από αυτή τη σελίδα έως ότου φτάσει ο αριθμός των έσω-βαθμών σε μια ελάχιστη προκαθορισμένη τιμή. Σύμφωνα με πειράματα που έχουν γίνει σε όλο τον κόσμο, ένα τέτοιο φίλτρο δεν αλλάζει σχεδόν καθόλου τη

μορφή του γραφήματος, ούτε επηρεάζονται τα παρακάτω αποτελέσματα από την εφαρμογή του ή μη.

Ένα άλλο στοιχείο που μπορούμε να εξάγουμε είναι η χρησιμοποίηση των δεσμευμένων διευθύνσεων στον ελληνικό ιστό. Όπως φαίνεται στο παρακάτω διάγραμμα, η χρησιμοποίηση είναι μόλις στο 3%. Σύμφωνα με μία άλλη έρευνα, η οποία δεν ακολούθησε τον τρόπο της συλλογής με συλλέκτη, ο αριθμός των ενεργών δικτυακών τόπων είναι περίπου στις 11.000. Το νούμερο αυτό δεν μπορεί να χαρακτηριστεί αξιόπιστο γιατί δεν, αφενός, δεν είναι γνωστός ο τρόπος με τον οποίο έχει γίνει αυτή η έρευνα και, αφετέρου, γιατί το μεγαλύτερο ποσοστό των IP διευθύνσεων είναι παράνομες. Το πιο πιθανό είναι να ανήκουν σε χρήστες με σύνδεση μέσω τηλεφώνου και οι οποίοι έχουν έναν προσωπικό εξυπηρετητή, η για διευθύνσεις μέσα σε firewall, οι οποίες κατάφερνουν να είναι ορατές προς τα έξω.



Κατανομές έσω-βαθμού και έξω-βαθμού

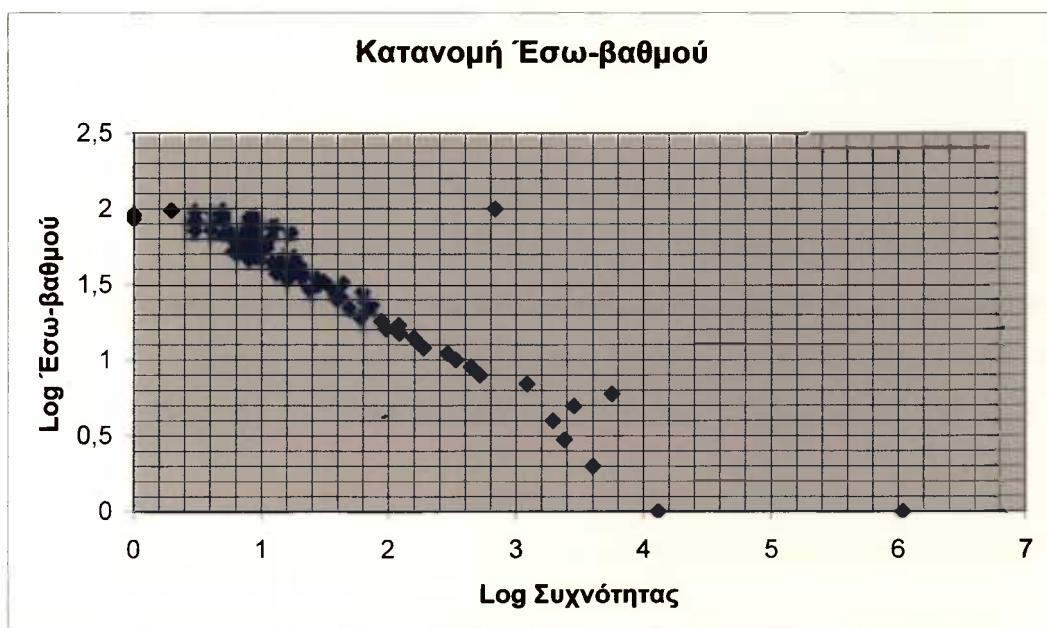
Έχει παρατηρηθεί πως η κατανομή των έσω-βαθμών σε δείγματα από τον παγκόσμιο ιστό ακολουθεί την κατανομή Zipfian. Η εναλλακτική κατανομή που προτείνεται είναι η εκθετική τρίτου βαθμού. Εάν λάβουμε υπόψη τον παράγοντα της κλίμακας, εκθετική τρίτου βαθμού φαίνεται πως είναι καλύτερη προσέγγιση. Από την άλλη πλευρά, αν γίνει επεξεργασία χωρίς να ληφθεί υπόψη η κλίμακα – αφού οι κόμβοι μετρούνται σε εκατομμύρια και οι έσω βαθμοί σε εκατοντάδες το πολύ – η κατανομή Zipfian δίνει καλύτερες τιμές. Η κατανομή Zipfian είναι της μορφής

$$f(x) = k \frac{1}{x^\alpha}$$

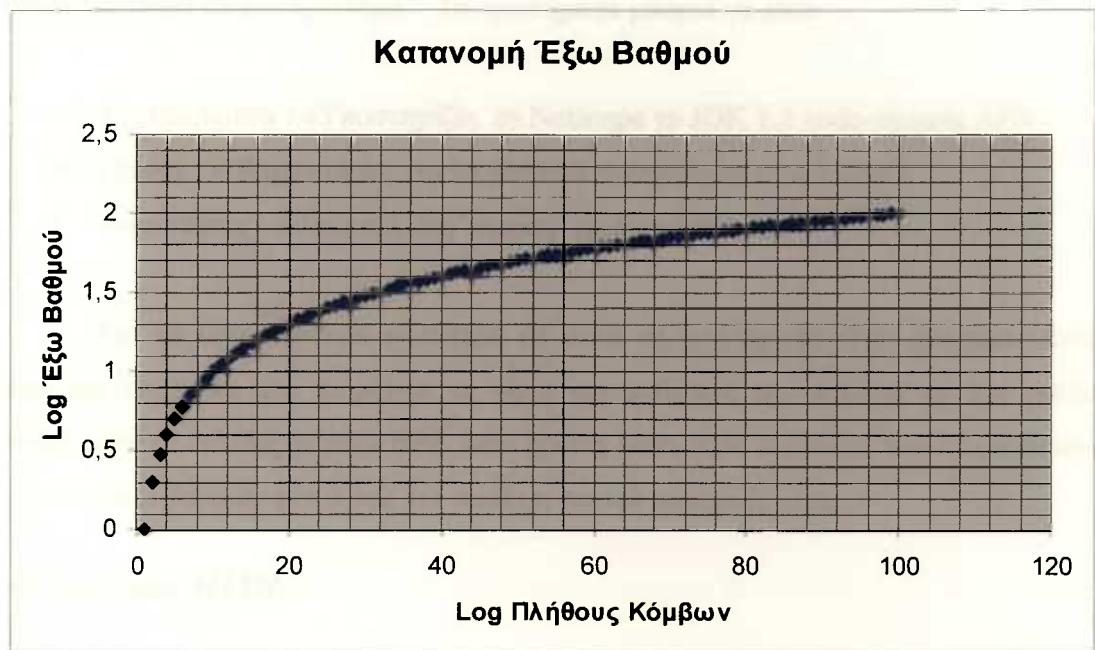
Το k είναι ο παράγοντας της κλίμακας, ενώ το $\alpha=2$, σύμφωνα με τις προαναφερθείσες έρευνες.

Αποτελέσματα

Τα αποτελέσματα για τον ελληνικό ιστό είναι παρόμοια με τα αποτελέσματα των άλλων ερευνών. Η κατανομή των έσω-βαθμών είναι η Zipfian και το $\alpha=2$, όπως φαίνεται και στο παρακάτω διάγραμμα, το οποίο έρχεται σε συμφωνία με μελέτες που έχουν γίνει σε άλλα δείγματα του παγκόσμιου ιστού.



Επίσης μπορούμε να δούμε και την κατανομή των έξω-βαθμών. Σε αυτή τη περίπτωση η κατανομή δεν ακολουθεί την κατανομή Zipfian στην αρχή, όμως, όσο το πλήθος των κόμβων τείνει στο άπειρο, η κατανομή φαίνεται να προσεγγίζεται από τη Zipfian.



Πηγές πληροφορίας

Ένα περιβάλλον συνδέσμων μπορεί να δώσει επιπλέον πληροφορία για το περιεχόμενο του, αν αναλύσουμε τον τρόπο με τον οποίο είναι δικτυωμένο. Με αυτό τον τρόπο μπορούμε να απαντήσουμε καλύτερα σε ερωτήσεις που γίνονται για εύρεση πληροφορίας στον ιστό. Το ερώτημα είναι : “πώς μπορούμε να απαντήσουμε αποδοτικότερα σε ένα ερώτημα”. Τα ερωτήματα μπορεί να είναι :

- Συγκεκριμένα : «Υποστηρίζει το Netscape το JDK 1.1 code-signing API»
- Γενικά : «Πληροφορίες για τη γλώσσα προγραμματισμού Java»
- Ομοιότητας : «Βρες σελίδες ‘όμοιες’ με java.sun.com»

Για να απαντήσουμε καλύτερα σε αυτά τα ερωτήματα είναι χρήσιμο εκτός από την ανάλυση του κειμένου να γίνει και ανάλυση του τρόπου με τον οποίο συνδέονται οι σελίδες μεταξύ τους. Μία πρώτη λύση είναι η μελέτη των έσω-βαθμών και των έξω-βαθμών που όμως δεν παρέχει σωστά αποτελέσματα.

Αλγόριθμος HITS

Ο Kleinberg στο paper του [Kleinberg97] εισάγει την έννοια των αξιόπιστων δεικτών (hubs) και πηγών πληροφορίας (authorities) πάνω σε κάποια ερώτηση (query), δηλαδή σε κάποιο θέμα. Κάθε κόμβος του γραφήματος «ζυγίζεται» με ένα hub, authority βάρος. Με αυτόν τον τρόπο προκύπτουν οι σελίδες που μπορούν να δώσουν την πιο αρμόδια πληροφορία για κάποιο θέμα. Το υπολογιστικό βήμα του Kleinberg που «ζυγίζει» τους κόμβους το εφαρμόσαμε όχι με κάποια συγκεκριμένη ερώτηση αλλά στο ελληνικό δίκτυο στο σύνολο του ψάχνοντας εκεί για τις πιο αρμόδιες σελίδες. Η ισχύς του αλγορίθμου έγκειται και στον παράγοντα άνθρωπο. Με αυτό εννοούμε πως ο Kleinberg κάνει την παραδοχή πως οι σύνδεσμοι υπάρχουν γιατί εξυπηρετούν και τους έχουν βάλει άνθρωποι. Δηλαδή η λογική των συνδέσμων σε επίπεδο έννοιας έγκειται στους αγθρώπους.

Υποθέτουμε ότι έχουμε μία ερώτηση, η οποία προσδιορίζεται από μία συμβολοσειρά σ. Θέλουμε να προσδιορίσουμε τις πηγές πληροφορίας, εφαρμόζοντας μια ανάλυση στη δομή των συνδέσμων. Πρώτα θα πρέπει να προσδιορίσουμε το

υπογράφημα του ιστού που θα εργαστούμε, έτσι ώστε να μειωθεί ο υπολογιστικός φόρτος μόνο στις σελίδες που είναι σχετικές με το ζήτημα. Θα μπορούσαμε να θεωρήσουμε σαν υπογράφημα Q_σ όλες τις σελίδες όπου αναφέρεται η συμβολοσειρά S . Το μειονέκτημα σε αυτή την περίπτωση είναι ότι, από τη μία πλευρά, οι σελίδες μπορεί να ανέρχονται σε εκατομμύρια και από, την άλλη, καμία υπεύθυνη σελίδα να μην ανήκει σε αυτό το σύνολο. Θα πρέπει η συλλογή S_σ να πληροί κάποιες ιδιότητες :

- Το S_σ να είναι σχετικά μικρό, ώστε να μειωθεί ο υπολογιστικός φόρτος.
- Το S_σ να είναι πλούσιο σε σχετικές σελίδες, ώστε να βρούμε καλές υπεύθυνες σελίδες.
- Το S_σ να περιέχει τις περισσότερες ή αρκετές από τις καλύτερες υπεύθυνες σελίδες.

Για να πετύχουμε τους παραπάνω στόχους ακολουθούμε την παρακάτω διαδικασία. Επιλέγουμε ένα t , έτσι ώστε να συλλέξουμε τις t καλύτερες σελίδες για το ερώτημα σ από μια μηχανή αναζήτησης. Ονομάζουμε αυτό το αρχικό σύνολο των σελίδων R_σ . Το σύνολο αυτό έχει τις ιδιότητες (i) και (ii). Επίσης να σημειώσουμε πως το σύνολο R_σ έχει λίγες ακμές και είναι αδόμητο. Η ιδιότητα (iii) συνήθως δεν ικανοποιείται, αφού και το υπερσύνολο $Q_\sigma \supseteq R_\sigma$ δεν την ικανοποιεί.

Μπορούμε από το σύνολο R_σ να παράγουμε το W_σ έτσι ώστε να ικανοποιείται και η ιδιότητα (iii). Μπορεί οι υπεύθυνες σελίδες να μην ανήκουν στο R_σ , αλλά υπάρχει σίγουρα μία σελίδα $p \in R_\sigma$ ή οποία δείχνει σε μία υπεύθυνη σελίδα. Για αυτό επεκτείνουμε το σύνολο R_σ με κόμβους που δείχνουν στο R_σ και με κόμβους που δείχνονται από το R_σ . Σε μετρήσεις που έχουν γίνει, ο αριθμός των κόμβων του R_σ αυξάνει κατά 250%.

Η επόμενη φάση του αλγορίθμου είναι επαναληπτική και χρησιμοποιούνται δύο λειτουργίες I και η O . Η πρώτη υπολογίζει τα βάρη των πηγών και η δεύτερη τα βάρη των δεικτών σύμφωνα με τα παράκατω :

x_p : authoritative weight

y_p : hub weight

$$y_p = \sum_{q \text{ such that } p \rightarrow q} x_q$$

$$x_p = \sum_{q \text{ such that } q \rightarrow p} y_q$$

To Web ως γράφημα

Η διαδικασίες είναι αναδρομικές και θα μπορούσαν να γραφούν ως εξής :

$$I : x \leftarrow A^T y \leftarrow A^T Ax \leftarrow (A^T A)x$$
$$O : y \leftarrow Ax \leftarrow AA^T y \leftarrow (AA^T)y$$

Μετά από αυτό το βήμα ακολουθεί μια κανονικοποίηση των βαρών στα

$$x \leftarrow A^T y \quad \& \quad y \leftarrow Ax$$

διανύσματα x και y.

Από τη γραμμική άλγεβρα είναι γνωστό ότι :

$$\lim_{n \rightarrow \infty} ((A^T A)y)^n = P(A^T A)$$

$$P(A^T A) : \text{φασματική ακτίνα } A^T A$$

Παρατηρούμε πως το αποτέλεσμα δεν εξαρτάται από την αρχικοποίηση των x και y αλλά από τη μήτρα γειτνίασης. Θα πρέπει τα διανύσματα x και y να μην είναι μηδενικά. Για αυτό το λόγο κάνουμε μια παραδοχή και αρχικοποιούμε τα διανύσματα με 1. Μετά από 50 επαναλήψεις έχουμε μια πολύ καλή προσέγγιση της φασματικής ακτίνας.

Μετά από το επαναληπτικό βήμα γίνεται ταξινόμηση με φθίνουσα σειρά των βαρών στα διανύσματα x και y και έτσι λαμβάνουμε ως αποτέλεσμα τις πηγές και τους δείκτες.

Μετρήσεις

Θα μπορούσαμε να φανταστούμε μια ερώτηση σε μια μηχανή αναζήτησης η οποία θα μας επέστρεφε ολόκληρο τον ελληνικό ιστό, όπως π.χ. “gr domain”. Η μήτρα γειτνίασης είναι όλο το γράφημα του ελληνικού ιστού που έχουμε. Με βάση λοιπόν τον ελληνικό ιστό που έχουμε, ο αλγόριθμος HITS έδωσε τα ακόλουθα αποτελέσματα.

Πηγές πληροφορίας

www.smart.gr
www.forthnet.gr
www.mfa.gr
www.mfa.gr/ggae
www.westnet.gr
www.skynet.gr
www.koz.forthnet.gr/net99/network
www.otenet.gr
www.hol.gr
pathfinder.gr

Δείκτες Πληροφορίας

spoc.hol.gr/HOLTools/grzone.html
www.gamecenter.gr/HOLTools/grzone.html
www.developer.gr/HOLTools/grzone.html
www.ntsupport.gr/HOLTools/grzone.html
www.pathfinder.gr/HOLTools/grzone.html
www.magenta.gr/otherlinks_gr.html
www.vavouras.gr/links.html
www.elka.gr/linksengcont.htm
www.elka.gr/links.htm
users.otenet.gr/~tiresias/links/links.html

Ακολουθούν οι σελίδες σε φθίνουσα σειρά σύμφωνα με τον έσω-βαθμό και των έξω-βαθμώ:

Έσω-βαθμός

www.in.gr
signup.in.gr
info.in.gr/help.htm
info.in.gr/contact.htm
info.in.gr/advertise.htm
info.in.gr/about.htm
www.hellasnet.gr
www.phaistosnetworks.gr
chat.in.gr
info.in.gr/newsite.htm

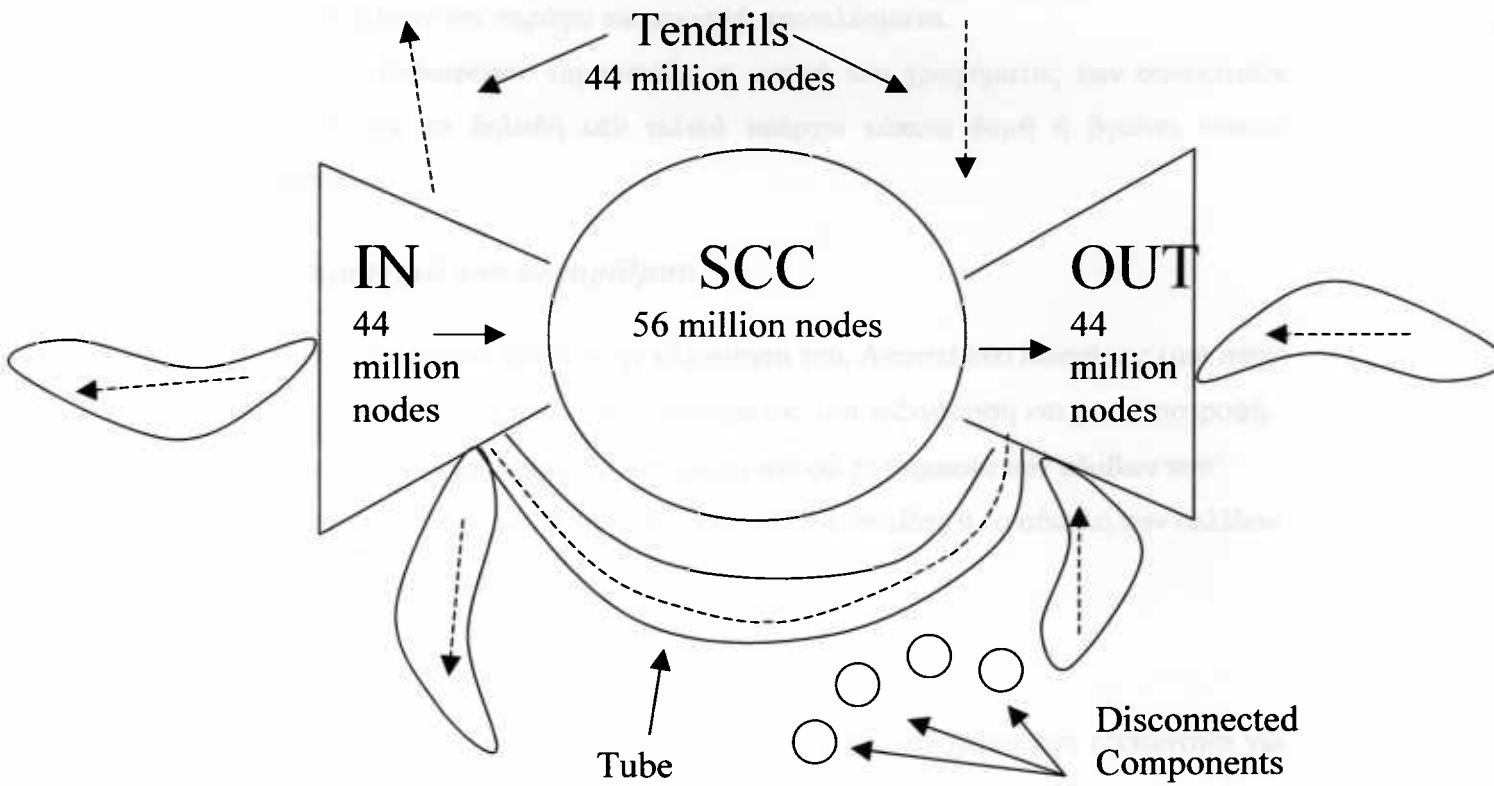
Έξω-βαθμός

homepages.pathfinder.gr/all_users.php
spoc.hol.gr/HOLtools/grzone.html
www.gamecenter.gr/HOLtools/grzone.html
www.ntsupport.gr/HOLtools/grzone.html
www.developer.gr/HOLtools/grzone.html
www.snowreport.gr/archive.htm
www.eone.gr/portal/depts/dept.asp
genesis.ee.auth.gr/dimakis/poets.htm
www.edu.uch.gr/usage/page.html
www.protoporia.gr/protoporia/subject_new.asp

Παρατηρούμε πως ο έσω-βαθμός δίνει αποτελέσματα που είναι λάθος για να απαντήσουν στο ερώτημα “gr domain”. Ωα πρέπει να παρατηρήσουμε επιπλέον πως στους δείκτες υπάρχουν στοιχεία με τους μεγαλύτερους έξω-βαθμούς. Ουσιαστικά πρόκειται για σελίδες που έχουν όλα τα δεσμευμένα “.gr” ονόματα.

Ισχυρές συνεκτικές συνιστώσες

Ένα ενδιαφέρον θέμα είναι το κατά πόσο υπάρχουν συνεκτικές συνιστώσες στον ιστό, και αν υπάρχει πυρήνας. Σε μελέτες που έχουν γίνει έχει παρατηρηθεί πως το γράφημα των συνιστωσών του ιστού έχει το παρακάτω σχήμα :



Ο ιστός σε παγκόσμιο επίπεδο έχει μια κεντρική συνιστώσα η οποία αποτελεί το 40% περίπου του συνολικού ιστού. Υπάρχουν άλλα δύο μέρη του ιστού : ο παλιός ιστός και ο νέος ιστός. Ο νέος ιστός είναι αυτός που δείχνει στον πυρήνα αλλά αυτός ακόμα δεν δείχνει σε αυτόν. Ο παλιός ιστός είναι αυτός που τον δείχνει ο πυρήνας αλλά πρόκειται για σελίδες οι οποίες δεν συντηρούνται πλέον και το περιεχόμενό τους δεν έχει καμία αξία. Το επόμενο βήμα στο κύκλο ζωής μιας σελίδας είναι να ανήκει στα σκουπίδια ή αλλιώς στις συνιστώσες που δεν έρχονται σε επαφή με κανένα μέρος του τωρινού ιστού. Μπορεί ο παλιός και ο νέος ιστός να συνδέονται με σωληνώσεις (tubes), οι οποίες δεν περνάνε από τον πυρήνα. Ο νέος και ο παλιός πυρήνας αποτελούνται από μια ή και περισσότερες συνιστώσες (tendrils).

Η εύρεση των συνεκτικών συνιστωσών εξερευνεί το γράφημα και ψάχνει για συνεκτικά κομμάτια του. Δίνει με αυτό τον τρόπο μια πιο μακροσκοπική εικόνα των δεδομένων από την απλή αναπαράσταση κόμβων και ακμών.

Ένα ανοικτό θέμα στον αλγόριθμο είναι το τι θα θεωρήσουμε κόμβο γραφήματος, τις απλές σελίδες ή τους web κόμβους (hosts) που αντιπροσωπεύουν πολλές σελίδες μαζί. Από την άποψη της αναπαράστασης (visualization) και της απόδοσης αδιαμφισβήτητα συμφέρει η δεύτερη λύση που επεξεργάζεται μικρότερο μέγεθος δεδομένων και παράγει πιο προστιά αποτελέσματα.

Ενδιαφέρον παρουσιάζει η μορφή του γραφήματος των συνεκτικών συνιστωσών για το δηλαδή εάν τελικά υπάρχει κάποια δομή ή βγαίνει κάποιο συμπέρασμα.

Χαρακτηριστικά του αλγορίθμου

- Είναι εξαιρετικά απλός στην υλοποίηση του. Απαιτεί δυο διασχίσεις (μια στην αρχή και μια στο τέλος) του γραφήματος, μια ταξινόμηση και μια αναστροφή.
- Όπως ήδη αναφέρθηκε μια απόφαση αφορά τη σημασία των κόμβων που μπορούν να αντιπροσωπεύουν απλές HTML σελίδες ή το σύνολο των σελίδων ενός τόπου.

Βήματα

Βήμα 1^o : Διέσχισε το γράφημα και υπολόγισε το χρόνο που εξετάστηκε για τελευταία φορά κάθε κόμβος.

Βήμα 2^o : Ανέστρεψε το γράφημα.

Βήμα 3^o: Ταξινόμηση κατά φθίνουσα σειρά τη χρόνο τελευταίας εξέτασης των κόμβων.

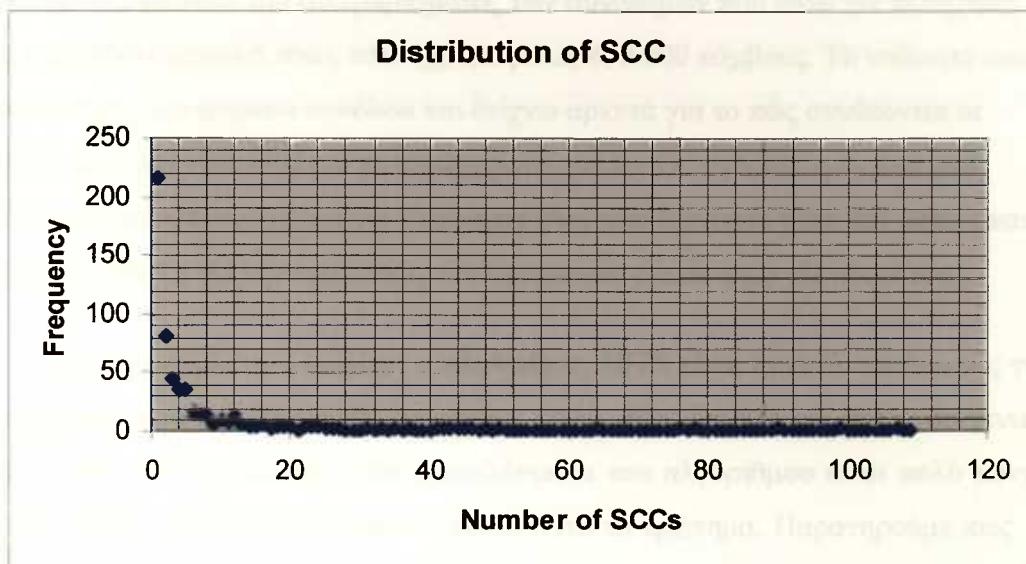
Βήμα 4^o: Με βάση τους χρόνους, άρχισε την εξερεύνηση με τον πρώτο κόμβο στην ταξινομημένη λίστα.

Κόστος υλοποίησης

Βήμα / Διαδικασία	Κόστος
Βήματα 1,2,4	$O(E)$
Βήμα 3	$O(ElogE)$

Αποτελέσματα

Η κατασκευή του γραφήματος των συνεκτικών συνιστωσών στον ελληνικό ιστό έγινε στο σύνολό του, πριν το φίλτραρισμα των εσωτερικών συνδέσμων. Η μεγαλύτερη συνιστώσα είχε μέγεθος 7864 κόμβων, αριθμός που είναι πολύ μικρός για το συνολικό αριθμό των κόμβων (0,702%). Η επόμενη συνιστώσα είχε μέγεθος 108 κόμβων. Το πιο ενδιαφέρον στοιχείο είναι πως 1.107.300 κόμβοι δεν ανήκουν σε καμία συνιστώσα. Σε επίπεδο εξυπηρετητών, η συνεκτική συνιστώσα περιελάμβανε 589 από το σύνολο των 2899 – σε ποσοστό δηλαδή γύρω στο 20,3%. Στον πίνακα φαίνεται η κατανομή των συνιστωσών ανάλογα με το μέγεθός τους. Ο άξονας των x έχει το μέγεθος των συνιστωσών σε κόμβους και ο y το πλήθος που παρατηρήθηκαν αυτές.



Από τα παραπάνω φαίνεται πως ο ελληνικός ιστός δεν έχει σχηματίσει ακόμα πυρήνα και μάλλον βρίσκεται σε πρώιμα στάδια της ανάπτυξής του. Έχει πολύ μεγάλο ενδιαφέρον να μελετηθεί ο τρόπος εξέλιξης του ιστού και ο τρόπος με τον οποίο φτάνει στο επίπεδο του γραφήματος των συνεκτικών συνιστωσών που ισχύει για τον ιστό στο σύνολό του.

Επίλογος – Συμπεράσματα

Η εξέταση του ελληνικού ιστού γίνεται για πρώτη φορά και δεν υπάρχουν αναφορές στις οποίες μπορούμε να στηριχθούμε για να επιβεβαιώσουμε τις μετρήσεις που έγιναν. Η συλλογή που έγινε έδειξε ότι μόνο ένα μικρό μέρος του ιστού χρησιμοποιείται ενεργά – μόλις το 3% - και πως το μεγαλύτερο μέρος είναι δεσμευμένο μελλοντική χρήση.

Ο μόνος περιορισμός της συλλογής ήταν να σαρώσει μόνο διευθύνσεις οι οποίες περιέχονται στον ελληνικό ιστό. Από αυτές – αρχικά 2.000.000 περίπου – οι στατικές σελίδες (htm, html, jsp, asp, jhtml, php) που μας ενδιαφέρουν είναι μόλις 1.100.000 περίπου. Αν αφαιρέσουμε τους κόμβους με μηδενικούς έσω-βαθμούς και έξω-βαθμούς, κατόπιν του φιλτραρίσματος των συνδέσμων που είναι για πλοήγηση μέσα στον ίδιο δικτυακό τόπο, τότε έχουμε μόλις 850.000 κόμβους. Το νούμερο αυτό είναι στο 40% του αρχικού συνόλου και δείχνει αρκετά για το πώς συνδέονται οι σελίδες στον ελληνικό ιστό και το περιεχόμενο αυτών. Οι περισσότεροι από τους συνδέσμους χρησιμοποιούνται για πλοήγηση στον ίδιο δικτυακό τόπο και αυτό είναι μια πρώτη ένδειξη για την οργάνωση των δικτυακών τόπων στον ελληνικό ιστό.

Η πληροφορία που παρέχει ο αλγόριθμος HITS είναι αρκετά αποδοτικός για τον ελληνικό ιστό και δείχνει ότι μπορεί να δώσει πληροφορίες για ένα τόσο γενικό ερώτημα όσο το “gr domain”. Τα αποτελέσματα του αλγόριθμου είναι πολύ κοντά στις απαντήσεις που θα ήθελε όποιος έκανε αυτό το ερώτημα. Παρατηρούμε πως η πλειοψηφία των σελίδων-πηγών αναφέρεται σε δικτυακά θέματα και μεγάλους ISPs. Επίσης υπάρχει η σελίδα του Υπουργείου Εξωτερικών και η σελίδα του Υπουργείου Εξωτερικών για τον Απόδημο Ελληνισμό. Η πρώτη πηγή είναι μια χρηματιστηριακή εταιρία. Αυτό είναι είτε αποτέλεσμα της έλλειψης του φίλτρου στο γράφημα για σελίδες διαφημίσεις, είτε είναι μια ένδειξη για το τι έχει άνθιση στην Ελλάδα τον τελευταίο καιρό, αν υποθέσουμε ότι ο ιστός είναι καθρέφτης μιας χώρας. Στην περίπτωση των δεικτών παρατηρούμε μια ταύτιση σε μεγάλο ποσοστό με το σύνολο των σελίδων με μεγάλο έξω-βαθμό. Αυτό είναι αποτέλεσμα του γεγονότος ότι οι σελίδες με τους μεγαλύτερους έξω-βαθμούς είχαν συνδέσμους σε όλες τις δεσμευμένες διευθύνσεις.

Το γράφημα των συνιστώσων δεν έδωσε τα αποτελέσματα που έχουν δώσει μελέτες για τον παγκόσμιο ιστό. Ο ελληνικός ιστός φαίνεται πως είναι σε πρώιμα στάδια ανάπτυξης, λαμβάνοντας υπόψη πως η μεγαλύτερη συνιστώσα του είναι της τάξης του 0,702% των συνολικών κόμβων. Από αυτή την άποψη έχει αρκετό ενδιαφέρον να μελετηθεί η εξέλιξη του ελληνικού ιστού.

Σε ένα επόμενο στάδιο θα ήταν χρήσιμο να μελετηθεί ποιες από τις σελίδες που είναι εκτός ελληνικού ιστού δείχνουν στον ελληνικό ιστό και σε ποιες σελίδες δείχνουν κατά κύριο λόγο. Ήα μπορούσε με αυτό τον τρόπο να επεκταθεί το ήδη υπάρχον γράφημα στον ελληνικό ιστό και με εξωτερική πληροφορία και να δώσει καινούρια αποτελέσματα σε σχέση με το ποιες είναι οι πηγές και ποιοι οι δείκτες στον ελληνικό ιστό, σύμφωνα με τον αλγόριθμο HITS.

Ο ελληνικός ιστός βρίσκεται στα αρχικά στάδια της εξέλιξής του και είναι πολλές πληροφορίες που μπορούν να εξαχθούν τόσο για την ποιότητα της πληροφορίας του, όσο και για τον τρόπο με τον οποίο εξελίσσεται.

Βιβλιογραφία

[Adamic99] L. Adamic, The Small World Web, *In Proc. of ECDL'99, September 22-24, 1999, Paris, France.*

[Altavista] www.altavista.com

[AVSE] Compaq. AltaVista Search Intranet Version 3 Administration Handbook, June 2000.

[HM99] Allan Heydon and Marc Najork. Mercator: A Scalable, Extensible Web Crawler, Compaq Systems Research Center, June 1999.

[Google] www.google.com

[Infoseek] www.infoseek.com

[KKRRT] Jon M. Kleinberg¹, Ravi Kumar², Prabhakar Raghavan², Sridhar Rajagopalan² and Andrew Tomkins². The web as a graph : measurements, models and methods, ¹Department of Computer Science, Cornell University, ²IBM Almaden Research Center.

[Kleinberg97] J.M.Kleinberg, Authoritative sources in a hyperlink environment, IBM Research Report RJ 10076 (91892), May 1997. Also, in Proc of ACM-SIAM Symposium on Discrete Algorithms, 1998

[Kleinberg00] J. Kleinberg, The Small-World Phenomenon: An Algorithmic Perspective, *In Proc. of STOC 2000, May 21-23, 2000, Portland, Oregon.*

[KPR] J. Kleinberg, C. Papadimitriou, and P. Raghavan. A Microeconomic View of Data Mining.

[KRRT] R.Kumar, P.Raghavan. S.Rajagopalan, and A.S.Tomkings, Trawling the web for emerging cyber-communities, *In Proc. 8th WWW Conference, 1999.*

[LEDA] <http://www.mpi-sb.mpg.de/LEDA>

[Mercator] Allan Heydon and Marc Najork. Mercator: A Scalable, Extensible Web Crawler, Compaq Systems Research Center, June 1999.

[MN] K.Mehlhorn and S.Näher, *The LEDA Book*, Cambridge University Press.

[Perl] <http://www.perl.com/pub>

[Yahoo] www.yahoo.com

[YBCFW00] Ziv Bar-Yossef, Alexander Berg, Steve Chien, Jittat Fakcharoenphol, and Dror Weitz, Approximating Aggregate Queries about Web Pages via Random Walks, In Proc. of VLDB 2000.

Περίληψη

Η αναζήτηση πληροφορίας στον παγκόσμιο ιστό είναι ένα από τα πιο ενδιαφέροντα και προκλητικά πεδία έρευνας. Η ανακάλυψη πληροφορίας σχετικά με ένα θέμα ή ακόμα και η ακριβής στατιστική μελέτη των ιστοσελίδων είναι ζητήματα αρκετά δύσκολα λόγω της χαοτικής εξέλιξης του ιστού και της έλλειψης οργάνωσης αυτού.

Για την αποδοτικότερη μελέτη του παγκόσμιου ιστού μπορούμε να τον θεωρήσουμε ως ένα προσανατολισμένο γράφημα. Η κάθε σελίδα είναι ένας κόμβος στο γράφημα και ένας σύνδεσμος από μια σελίδα ρ σε μια σελίδα q είναι μια προσανατολισμένη (p,q) ακμή του γραφήματος. Προς αυτή τη θεώρηση του ιστού έχουν προχωρήσει όλες οι ερευνητικές προσπάθειες.

Το πρώτο βήμα είναι να συλλέξουμε τις σελίδες από τον παγκόσμιο ιστό. Για αυτό το σκοπό χρησιμοποιούνται ειδικά προγράμματα που ονομάζονται συλλέκτες (crawlers). Οι συλλέκτες διακρίνονται, ανάλογα με τη χρησιμότητά τους σε δύο τύπους : τους προσανατολισμένους σε δικτυακούς τόπους (site-specific) και στους γενικού σκοπού (general-purpose). Οι αρχιτεκτονικές των δύο τύπων αυτών είναι διαφορετικές στο τρόπο με τον οποίο μαζεύουν τα δεδομένα, αλλά όλοι χρησιμοποιούν στοίβες για την υλοποίησή τους.

Στο επόμενο βήμα γίνεται η κατασκευή του γραφήματος των συνδέσμων. Το μοναδικό εργαλείο που υπάρχει αυτή τη στιγμή είναι ο Connectivity Server της Compaq, ο οποίος συνεργάζεται με την AltaVista Search Engine. Για τους δικούς μας σκοπούς αναπτύχθηκε ένα εργαλείο το οποίο έχει την ίδια λειτουργία με τον Connectivity Server. Η υλοποίηση έγινε χρησιμοποιώντας το AltaVista SE SDK, τις βιβλιοθήκες LEDA και τη γλώσσα προγραμματισμού C++.

Τα πρώτα στατιστικά στοιχεία αφορούν το ίδιο το γράφημα. Αρχικά υπολογίστηκε το πλήθος των ακμών και των κόμβων. Στην περίπτωση του ελληνικού ιστού είναι περίπου 7 εκατομμύρια ακμές και 1,1 εκατομμύρια κόμβοι. Κατόπιν επεξαργασίας για τη διαγραφή των συνδέσμων που εξυπηρετούν ανάγκες πλοήγησης

σε ένα δικτυακό τόπο οι ακμές μειώθηκαν στα 3,5 εκατομμύρια και οι κόμβοι στις 850.000. Το πλήθος των δεσμευμένων διευθύνσεων είναι 30000, από τις οποίες γίνεται χρήση μόνο το 3% στον ελληνικό ιστό. Η κατανομή των έσω-βαθμών και των έξω-βαθμών ακμών ακολουθεί την Zipfian κατανομή, με συντελεστή, όπως και στον παγκόσμιο ιστό, $\alpha=2$.

Για την εύρεση των πηγών (authorities) στον ελληνικό ιστό χρησιμοποιήθηκε ο αλγόριθμος HITS, οποίος χρησιμοποιεί τη πληροφορία της διασύνδεσης των σελίδων. Σύμφωνα με τον αλγόριθμο, κάθε σελίδα έχει δύο βάρη : ένα ως πηγή και ένα ως δείκτης (hub). Οι πηγές είναι σελίδες οι οποίες είναι προσανατολισμένες στι θέμα, ενώ οι δείκτες περιέχουν συνδέσμους σε σχετικές με το θέμα σελίδες. Στο πρώτο βήμα του αλγόριθμου, το οποίο είναι αναδρομικό, ενημερώνονται τα βάρη των ακμών και των κόμβων και κανονικοποιούνται οι τιμές. Στο δεύτερο βήμα έχουμε ταξινόμηση των κόμβων σύμφωνα με τα βάρη. Οι πηγές στον ελληνικό ιστό είναι δικτυακοί τόποι ISPs και του Υπουργείου Εξωτερικών, ενώ οι δείκτες είναι σελίδες οι οποίες έχουν ως συνδέσμους όλες τις δεσμευμένες διευθύνσεις του ελληνικού ιστού.

Στο τελευταίο βήμα έγινε προσπάθεια σκιαγράφησης της δομής του ελληνικού ιστού με το γράφημα των συνεκτικών συνιστώσων του. Ο ελληνικός ιστός δυστυχώς δεν έχει κάποιο πυρήνα αξιόλογο. Η μεγαλύτερη σε πλήθος κόμβων συνεκτική συνιστώσα έχει 7864 κόμβους, περίπου το 0,7% του συνόλου, ενώ η επόμενη σε μέγεθος έχει 108 κόμβους. Αξιόλογο είναι το γεγονός ότι 1.107.300 κόμβοι δεν ανήκουν σε καμία συνιστώσα. Αυτό είναι μια ένδειξη της μορφής του ιστού όταν είναι στα αρχικά στάδια εξέλιξής του.

Περαιτέρω δουλεία μπορεί να γίνει στο τομέα των κοινοτήτων (cyber communities). Αντί, δηλαδή, να προκύψουν οι πηγές και οι δείκτες, να προκύψουν οι θεματικές ενότητες του ελληνικού ιστού. Επίσης θα ήταν χρήσιμο να επεκταθεί το γράφημα και με όλους τους κόμβους που δείχνουν στον ελληνικό ιστό. Λόγω της νηπιακής ηλικίας του ιστού στην Ελλάδα, μπορούμε να μελετήσουμε τον τρόπο με τον οποίο εξελίσσεται ένας ιστός.

Executive Summary

Searching the World Wide Web for information is one of the most fascinating and challenging domains of research. Either the reveal of information subject to a topic or precise statistical information about the pages is a very tough issue. This is due to the chaotic evolution of the WWW and the lack of structure that WWW has.

In order to observe WWW in a more efficient way, we can consider a directed graph, where nodes of the graph are the pages of the WWW and the directed edges (p,q) are the hyperlinks between the pages p and q . All research efforts are in this direction.

The first step is to collect the pages from the web. In order to achieve this, we use crawlers. There are two different kinds of crawlers: site-specific and general-purpose crawlers. Both of the two architectures use stacks but they collect the pages with different mechanisms. Site-specific crawlers are using multithreading mechanisms whereas general-purpose crawlers use agents.

The follow-up of the crawling is to build the link graph. One, if not the only tool yet, is Compaq's Connectivity Server, which co-operates with AltaVista Search Engine. For our purpose we develop a tool, with the same functionality as the Compaq's Connectivity Server, using AltaVista SDK, LEDA libraries and C++.

Elementary statistics for the graph are the number of nodes and edges. We found 7 million edges and 1,1 million nodes. After further filtering for intrinsic links we had a decrease at the number of edges at half and the number of nodes were 850,000. The number of the reserved DNS names are approximately 30,000 and it is being used only the 3%. The distribution of in-degree and out-degree is a Zipfian one, with $a=2$.

In order to find the authorities and the hubs of the .gr domain we have used the HITS algorithm, which is using the link information to extract information from the web. Each page has two weights: an authoritative and a hub one. Authoritative pages are focused on the topic whereas hubs pages contain links to useful, relevant pages on

the topic. The first step of the algorithm is an iterative step, which consists of two operations, each of which updates the authoritative and hub weight of each page. After each iterative step we have a normalization of the two vectors. In the second step we sort the nodes in a decreasing order according to the weights. The authoritative pages in .gr are ISPs pages and hub pages are pages we have lists of links with every reserved DNS name in .gr.

Finally we tried to build the .gr graph according to its strongly connected components. Unfortunately, .gr has not a core. The biggest component has 7864 nodes – 0,7% of the total number of nodes. One more interesting result is that there are 1,107,300 nodes that are not part of any strongly connected component. This is a clue that .gr web is in its early stages.

Further work can be done in the area of cyber communities – a reversed view of the HITS algorithm for authoritative and hub pages. We could also expand our graph including pages, which are not in .gr, and they link to .gr. Due to the fact that .gr web is at its early stages, we can study the evolution of the web.



80025 75540

