

Θέματα σχετικά με Εξόρυξη από τον Παγκόσμιο Ιστό



Ανάκτηση Κειμένου (εισαγωγικά θέματα)



Βάσεις Κειμένων (document databases)

Μεγάλη συλλογή από κείμενα από διάφορες πηγές
[news articles, research papers, books, digital libraries, e-mail messages, Web pages, blogs, library database, etc.]
Τα δεδομένα δεν ακολουθούν κάποιο αυστηρό μοντέλο - ημι-δομημένα *semi-structured*

Information retrieval - Ανάκτηση Πληροφορίας

Η πληροφορία οργανώνεται σε (ένα μεγάλο αριθμό) από κείμενα - documents

Information retrieval problem: εντοπισμός των *σχετικών* κειμένων (documents) με βάση την είσοδο του χρήστη όπως λέξεις κλειδιά ή παραδείγματα κειμένου



Βασικές έννοιες

Ένα έγγραφο (document) αρχείο κειμένου μπορεί να περιγράψει από ένα σύνολο αντιπροσωπευτικών λέξεων-κλειδιά (keywords) που ονομάζονται **όροι δεικτοδότησης - index terms**.

▪ Διαφορετικοί όροι με διαφορετικό *βαθμό σχετικότητας* μπορούν να χρησιμοποιηθούν για την περιγραφή κειμένων με διαφορετικό περιεχόμενο

Αυτό επιτυγχάνεται με την ανάθεση **αριθμητικών βαρών** (numerical weights) σε κάθε όρο δεικτοδότησης του κειμένου (π.χ.: συχνότητα, tf-idf)

Αναλογία με ΣΔΒΔ:

Όροι Δεικτοδότησης → Γνωρίσματα
Βάρη → Τιμές γνωρισμάτων



Το μοντέλο ενός αρχείου

Επιλέγουμε ένα σύνολο από όρους

Διαδικό (Boolean) μοντέλο:

1: αν ο όρος υπάρχει
0: αν ο όρος δεν υπάρχει

Ερώτηση

$$(t_{11} \vee t_{12} \vee \dots \vee t_{1i1}) \wedge (t_{21} \vee t_{22} \vee \dots \vee t_{2i2}) \wedge \dots \wedge (t_{j1} \vee t_{j2} \vee \dots \vee t_{ji})$$

Όπου τα t_{ij} είναι όροι

Όλα τα έγγραφα που έχουν τους όρους

Ευρετηριοποίηση για την Ανάκτηση Κειμένου



Παράδειγμα

Did	Λέξεις-Κλειδιά
1	agent James Bond
2	agent mobile computer
3	James Madison movie
4	James Bond movie

Παραδείγματα ερωτήσεων:

Agent, James and agent, Agent or James

Ευρετηριοποίηση για την Ανάκτηση Κειμένου



Συνήθως, κατασκευάζονται **ευρετήρια** που περιέχουν ζεύγη <όρος, id-αρχείου> με πιθανών επιπλέον πεδία όπως η συχνότητα εμφάνισης του όρου στο αρχείο

Παρόμοια, ευρετήρια χρησιμοποιούν και οι μηχανές αναζήτησης

Ευρετηριοποίηση για την Ανάκτηση Κειμένου



Μια ταξινομημένη λίστα (**ανεστραμμένη λίστα**) (inverted file, inverted list, inverted index) για κάθε όρο

Παράδειγμα		Agent <1,2>
		Bond <1,4>
Did	Λέξεις-Κλειδιά	Computer <2>
1	agent James Bond	James <1,3,4>
2	agent mobile computer	Madison <3>
3	James Madison movie	Mobile <2>
4	James Bond movie	Movie <3,4>

Παράδειγμα ερωτήσεων

Postings (keyword, DocID)
Ταξινόμηση κάθε λίστας με βάση το DocID

Ευρετηριοποίηση για την Ανάκτηση Κειμένου



Ευρετήριο Λεξιλογίου:

Για τον ταχύτερο εντοπισμό της λίστας για κάθε όρο: Το σύνολο των όρων μπορεί να οργανωθεί με τη χρήση μιας δομής ευρετηρίου (π.χ. Β+-δέντρο)

Στα φύλλα, δείκτες προς την αντίστοιχη ανεστραμμένη λίστα

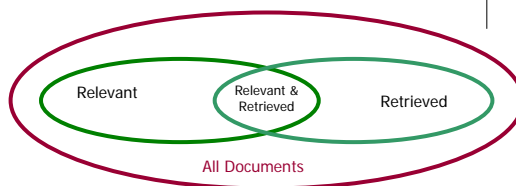
Παράδειγμα

Ένας όρος, σύζευξη, διάζευξη

Ανάκτηση Πληροφορίας



Βασικές Μετρικές



Precision - Ακρίβεια: το ποσοστό των ανακτημένων εγγράφων που είναι σχετικά με την ερώτηση (δηλαδή, το ποσοστό των «σωστών» απαντήσεων)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

Recall - Ανάκληση: το ποσοστό των σχετικών εγγράφων που ανακτούνται

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Μοντέλα Ανάκτηση Κειμένου



Το δυαδικό μοντέλο που είδαμε μέχρι στιγμής θεωρεί ότι οι όροι δεικτοδότησης είτε υπάρχουν είτε δεν υπάρχουν στο αρχείο (κείμενο)

Οι ερωτήσεις είναι όροι συνδεδεμένοι με : *not*, *and*, και *or*
πχ.: *car and repair*, *plane or airplane*

Το δυαδικό μοντέλο προβλέπει ότι ένα αρχείο είναι είτε σχετικό είτε μη σχετικό

Δεν υπάρχει διαβάθμιση (Ranking) - «πόσο» σχετικό

Διανυσματικό μοντέλο -> πάλι ένα έγγραφο περιγράφεται από τους όρους αλλά κάθε όρος με ένα **βάρος** (που σχετίζεται με τη συχνότητα εμφάνισης του όρου στο έγγραφο)

δηλαδή, στο δυαδικό μοντέλο, τα βάρη είναι όλα δυαδικά (0 ή 1)

Μοντέλα Ανάκτηση Κειμένου



Διανυσματικό Μοντέλο

Συχνότητα όρου- term frequency : πόσες φορές εμφανίζεται ένας όρος σε ένα έγγραφο

Κανονικοποιημένο ώστε να αποφύγουμε να δώσουμε μεγαλύτερο βάρος σε μεγάλα έγγραφα

Σημασία του όρου t_i σε ένα έγγραφο

Τοπική
μέτρηση

$$tf_i = \frac{n_i}{\sum_k n_k}$$



Ανεστραμμένη συχνότητα εγγράφου (*inverse document frequency*) μετρά πόσο γενικά σημαντικός είναι ένας όρος

Ολική μέτρηση

$$idf_i = \log \frac{|D|}{|\{d : d \in t_i\}|}$$

|D| αριθμός εγγράφων

Πόσα έγγραφα τον περιέχουν

Έγγραφα στα οποία ανήκει ο όρος t_i



$$tfidf = tf * idf$$

Μεγάλη τιμή όταν

- μεγάλη συχνότητα εμφάνισης (σε ένα συγκεκριμένο έγγραφο) και
- μικρή συχνότητα εμφάνισης του όρου σε όλη τη συλλογή

Βάρος χρήσιμο για να αποφύγουμε κοινούς όρους

Ομοιότητα με την ερώτηση



Τα αρχεία και οι ερωτήσεις αναπαρίστανται ως n -διάστατα διανύσματα, όπου n είναι ο συνολικός αριθμός όρων στη συλλογή

Ο βαθμός ομοιότητας ενός αρχείου d και μιας ερώτησης q υπολογίζεται ως η συνέλιξη τους, χρησιμοποιώντας μετρικές όπως η Ευκλείδεια απόσταση ή το συνημίτονο της γωνίας των δύο διανυσμάτων:

$$\cos(q, d) = \frac{q \cdot d}{|q| |d|}$$

Άλλα Θέματα



- **Ρίζα λέξεων** - Word stem
 - Πολλές λέξεις είναι μικρές παραλλαγές αφού έχουν μια κοινή ρίζα E.g., *drug, drugs, drugged*
- **Συνώνυμα** - Synonymy: Ενώ η λέξη-κλειδί T δεν εμφανίζεται στο κείμενο αν και το κείμενο είναι σχετικό
- **Πολυσημία** - Polysemy: Η ίδια λέξη μπορεί να σημαίνει διαφορετικά πράγματα με βάση τα συμφραζόμενα
- **Stop list**
 - Σύνολο λέξεων που δεν είναι σχετικά αν και εμφανίζονται συχνά, πχ, *a, the, of, for, to, with, etc.*
- **Οντολογίες** - Wordnet



Μηχανές Αναζήτησης



Μηχανές Αναζήτησης

Βασισμένες σε **ευρετήρια**: Αναζητούν σελίδες, τις δεικτοδοτούν και κατασκευάζουν τεράστια ευρετήρια βασισμένα σε λέξεις κλειδιά

Χρήσιμες για τον εντοπισμό σελίδων που περιέχουν συγκεκριμένες λέξεις κλειδιά

Προβλήματα

- Ένα θέμα μπορεί να περιέχει χιλιάδες έγγραφα
- Πολλά σχετικά με κάποιο θέμα έγγραφα μπορεί να μην περιέχουν τις λέξεις κλειδιά που το προσδιορίζουν



Θα δούμε

- Page Rank
- HITS

Και οι δύο εκμεταλλεύονται την ύπαρξη links
- **συνδέσεων** ανάμεσα στις σελίδες



PageRank



PageRank: Capturing Page Popularity (Brin & Page'98)

Ο αρχικός αλγόριθμος του google, παρουσιάστηκε στην κλασική εργασία:
"The Anatomy of a Large-Scale Hypertextual Web Search Engine",
Sergey Brin and Lawrence Page

Η εργασία περιλαμβάνει μια πολύ ενδιαφέρουσα «ιστορικής σημασίας» εισαγωγή

"We chose our system name, Google, because it is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search engines."

The verb, "google", was added to the [Merriam Webster Collegiate Dictionary](#) and the [Oxford English Dictionary](#) in 2006, meaning, "to use the Google search engine to obtain information on the Internet." (source: Wikipedia)



Βασική Ιδέα

Ακόμα και αν ένα τεράστιο ευρετήριο με όλες τις λέξεις και τις σελίδες -> αυτό που έχει σημασία είναι οι σημαντικές σελίδες (*precision vs recall*) τα «10 πρώτα» αποτελέσματα

ΣΤΟΧΟΣ: υπολογισμός μιας τιμής για κάθε σελίδα που να χαρακτηρίζει **πόσο σημαντική** είναι αυτή η σελίδα, η ποσότητα αυτή λέγεται **page rank**

Πότε είναι μια σελίδα σημαντική;



Βασική Ιδέα

- Οι Web pages δεν είναι όλες το ίδιο "σημαντικές"
www.joe-schmoe.com vs www.stanford.edu
- Αναφορές (Inlinks) ως «ψήφοι» - votes
www.stanford.edu 23,400 inlinks
www.joe-schmoe.com 1 inlink

οι συνδέσεις

μια σελίδα που δέχεται πολλές αναφορές περιμένει κανείς να είναι γενικά πιο σημαντική



Βασική Ιδέα (συνέχεια)

Ο PageRank βασίζεται στην «μέτρηση αναφορών» σε μία σελίδα ("citation counting"), αλλά με μια βελτίωση:

Δεν είναι όλες οι αναφορές το ίδιο σημαντικές!

Θεωρεί «έμμεσες αναφορές» "indirect citations":
Αναφορές από σημαντικές σελίδες (δηλαδή, από σελίδες που επίσης έχουν πολλές αναφορές) θεωρούνται πιο σημαντικές

Αναδρομικός ορισμός!



Απλή Αναδρομική Διατύπωση

Κάθε σελίδα μια ποσότητα που χαρακτηρίζει τη σημαντικότητα της (αυτή η ποσότητα καλείται **page rank**)

Αυτή η ποσότητα μοιράζεται ισόποσα στις εξωτερικές ακμές της σελίδας

Συγκεκριμένα:

- Η **ψήφος** κάθε ακμής (αναφοράς) είναι ανάλογη της σημαντικότητας (PR) της σελίδας από την οποία προέρχεται
- Αν μια σελίδα P με σημαντικότητα (PR) γ έχει n outlinks, κάθε link παίρνει γ/n ψήφους



Παράδειγμα

Έστω ότι υπάρχει μια γενική ποσότητα PR που μοιράζεται στις σελίδες του συστήματος.

Έστω 4 σελίδες: A, B, C και D.

Αρχική προσεγγιστική τιμή για καθεμία: PR = 0.25

- Έστω B, C, και D έχουν link μόνο στο A, τότε όλα το PageRank PR() τους θα μαζευόταν στο A

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

- Έστω τώρα ότι η B έχει link στη C, και η D έχει links και στο B και στο C. Η τιμή του PR μιας σελίδας μοιράζεται ανάμεσα στις εξωτερικές ακμές της. Άρα η ψήφος της B έχει αξία για την A 0.125 και 0.125 για την C. Αντίστοιχα, μόνο το 1/3 του PageRank του D μετρά για PageRank του A (περίπου 0.083).

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$



Γενικός ορισμός του PageRank για μια σελίδα A:

Έστω ότι η A έχει τις σελίδες T1, ..., Tn που δείχνουν σε αυτήν (δηλαδή, αναφορές)

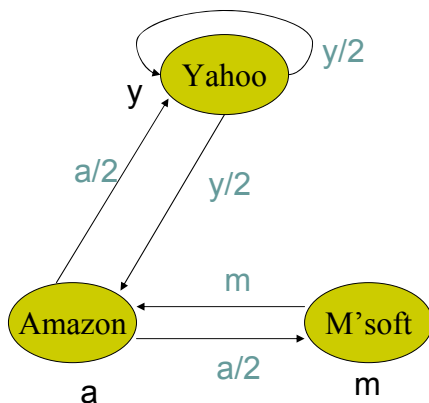
Έστω C(T) ο αριθμός των εξωτερικών ακμών μιας σελίδας T

$$PR(A) = PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn)$$



Απλό μοντέλο «ροής» - "flow" model

Το web το 1839



$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$



Λύση των εξισώσεων ροής

- 3 εξισώσεις, 3 άγνωστοι, όχι σταθερές
 - Μη μοναδική λύση
 - Οι λύσεις ισοδύναμες με κλιμάκωση (scale factor)
- Επιπρόσθετος περιορισμός για μοναδικότητα της λύσης
 - $y+a+m = 1$ (το συνολικό PR που μοιράζεται στις σελίδες)
 - $y = 2/5, a = 2/5, m = 1/5$

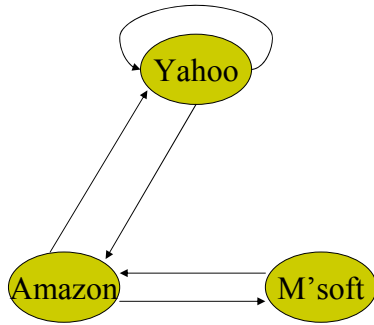


Διατύπωση με την μορφή πίνακα

- Ο πίνακας **M** έχει μια γραμμή και μια στήλη για κάθε web σελίδα (πίνακας γειτνίασης)
- Έστω ότι η σελίδα j έχει n outlinks
 - Αν $j \rightarrow i$, τότε $M_{ij}=1/n$
 - Αλλιώς, $M_{ij}=0$
- **M** είναι column stochastic matrix
 - Οι στήλες έχουν άθροισμα 1



Διατύπωση με την μορφή πίνακα (παράδειγμα)



$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$

	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

↑
Άθροισμα 1 (οι ψήφοι του y)



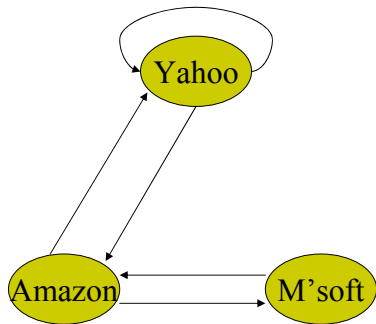
Διατύπωση με την μορφή πίνακα

- Έστω r ένα διάνυσμα με μια εγγραφή web σελίδα
 - r_i είναι η σημαντικότητα (PR) της σελίδας i
 - r : rank vector

[PR(y)
PR(a)
PR(m)]



PR Διάγραμμα (παράδειγμα)



$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$

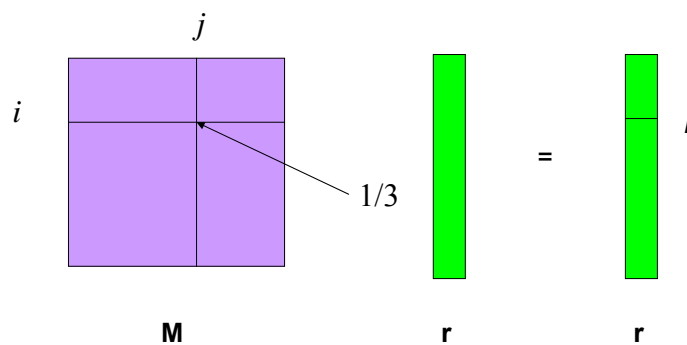
	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

$$r = Mr$$

y	1/2 1/2 0	y
a	1/2 0 1	a
m	0 1/2 0	m



Έστω ότι η σελίδα j έχει links σε 3 σελίδες, συμπεριλαμβανομένου του i





Ιδιοδιανύσματα (eigenvectors)

- Οι εξισώσεις ροής μπορούν να γραφούν

$$r = M r$$

- Δηλαδή, ο rank vector είναι ένα ιδιοδιάνυσμα (eigenvector) του στοχαστικού πίνακα γειτνίασης του web
 - Συγκεκριμένα είναι το βασικό ιδιοδιάνυσμα (αυτό που αντιστοιχεί στην ιδιοτιμή $\lambda = 1$)



Power Iteration method - Επαναληπτική Μέθοδο

Ένα απλό επαναληπτικό σχήμα (relaxation)
Έστω N web σελίδες

Αρχικοποίηση: $r^0 = [1/N, \dots, 1/N]^T$

Επανάληψη: $r^{k+1} = M r^k$

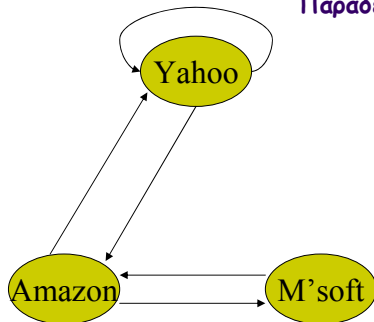
Τερματισμός όταν $|r^{k+1} - r^k|_1 < \epsilon$

$|x|_1 = \sum_{i=1}^N |x_i|$ είναι L_1 norm

Μπορεί να χρησιμοποιηθούν και άλλες μετρικές, πχ Ευκλείδεια



Παράδειγμα



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

y	=	1/3	1/3	5/12	3/8	2/5
a		1/3	1/2	1/3	11/24	2/5
m		1/3	1/6	1/4	1/6	1/5

Συγκλίνει; Μοναδική Λύση;

Μοντέλο Τυχαίου Δικτυακού Περιηγητή - Surfer -
(random walk)

Το PageRank μιας σελίδας μπορεί επίσης να θεωρηθεί ότι εκφράζει την πιθανότητα ένας *τυχαίος περιηγητής να φτάσει σε αυτήν* (δηλαδή, εκφράζει πόσο δημοφιλής είναι)

Ένας *τυχαίος περιηγητής* ξεκινά από μια τυχαία σελίδα και συνεχίζει να κάνει click σε links, χωρίς να επιστρέφει σε προηγούμενη σελίδα

- Τη χρονική στιγμή t , ο περιηγητής είναι σε κάποια σελίδα P
- Τη χρονική στιγμή $t + 1$, ο περιηγητής ακολουθεί ένα εξωτερικό link - outlink του P τυχαία (uniformly at random)
- Φτάνει σε κάποια σελίδα Q του P
 - Συνεχίζει την παραπάνω διαδικασία επ' άπειρων

Έστω $\mathbf{p}(t)$ το διάνυσμα του οποίου το i -οστό στοιχείο είναι η πιθανότητα ο περιηγητής να είναι στη σελίδα i τη χρονική στιγμή t

$\mathbf{p}(t)$ probability distribution - κατανομή πιθανότητας στις σελίδες



The stationary distribution

- Που είναι ο περιηγητής τη χρονική στιγμή $t+1$?
 - Ακολουθεί ένα link uniformly at random
 - $\mathbf{p}(t+1) = \mathbf{M} \mathbf{p}(t)$
- Έστω ότι ο τυχαίος περίπατος φτάνει μια κατάσταση όπου $\mathbf{p}(t+1) = \mathbf{M} \mathbf{p}(t) = \mathbf{p}(t)$
 - Τότε $\mathbf{p}(t)$ ονομάζεται *stationary distribution* για τον τυχαίο περίπατο
- Επειδή ο πίνακας \mathbf{r} ικανοποιεί την $\mathbf{r} = \mathbf{M}\mathbf{r}$
 - είναι *stationary distribution* για τον τυχαίο περιηγητή



Βασικό αποτέλεσμα από τη θεωρία τυχαίων περιπάτων (και Markov processes):

Για γράφους που ικανοποιούν συγκεκριμένες συνθήκες, η stationary distribution είναι μοναδική και τελικά φτάνουμε σε αυτήν ανεξάρτητα από την αρχική κατανομή πιθανότητας τη χρονική στιγμή $t = 0$ (σύγκλιση).

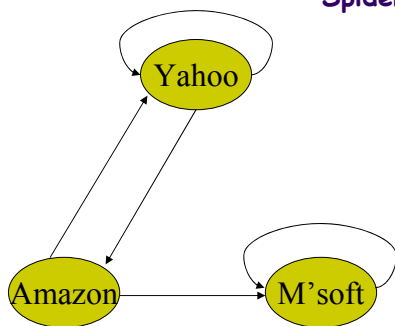


Spider traps

- Μια ομάδα σελίδων είναι μια **αραχνο-παγίδα (spider trap)** αν δεν υπάρχουν ακμές - από την ομάδα σε σελίδες εκτός της ομάδας
 - Ο τυχαίος surfer παγιδεύεται
- Οι συνθήκες που χρειάζονται για το θεώρημα των τυχαίων περιπάτων παύουν να ισχύουν



Spider traps (παράδειγμα)



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

y	=	1	1	3/4	5/8		0
a		1	1/2	1/2	3/8	...	0
m		1	3/2	7/4	2		3



Επέκταση Μοντέλου

Σε κάθε βήμα, ο τυχαίος surfer έχει δύο δυνατότητες:

- Με πιθανότητα β , ακολουθεί ένα τυχαίο link
- Με πιθανότητα $1-\beta$ πετάγεται σε κάποια άλλη σελίδα τυχαία
- Τιμές για το β : 0.8 - 0.9

Καταφέρνει να βγει από την παγίδα μετά από κάποιες χρονικές στιγμές



Επέκταση Μοντέλου

Αρχικός ορισμός του PageRank για μια σελίδα A:

$$PR(A) = PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)$$

Ορισμός με τον παράγοντα απόσβεσης d (damping factor) μεταξύ του 0 και του 1

$$PR(A) = (1-d)/N + d (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

Ώστε το άθροισμα να είναι 1 $\rightarrow 1-d/N$

Ο πρώτος παράγοντας λέει ότι με την ίδια πιθανότητα διαλέγω οποιαδήποτε σελίδα



- Κατασκευή του $N \times N$ πίνακα A

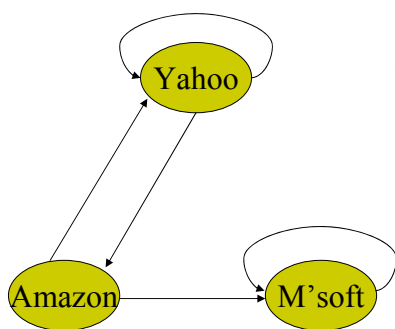
$$A_{ij} = \beta M_{ij} + (1-\beta)/N$$
- Ο A είναι στοχαστικός πίνακας
- Το **page rank διάνυσμα** r είναι το βασικό ιδιοδιάνυσμα αυτού του πίνακα

$$r = Ar$$

Ισοδύναμα, r είναι stationary distribution των τυχαίων περιπάτων με μεταπηδήσεις (random walk with teleports)



Παράδειγμα ($d=0.8$)



$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix}$$

y	1	1.00	0.84	0.776		7/11
a	=	1	0.60	0.60	...	5/11
m		1	1.40	1.56	1.688	21/11



Μοντέλο Τυχαίου Surfer (φυσική ερμηνεία)

Ένας τυχαίος surfer ξεκινά από μια τυχαία σελίδα και συνεχίζει να κάνει click σε links, χωρίς να επιστρέφει σε προηγούμενη σελίδα αλλά τελικά *βαριέται* και ξεκινά από κάποια άλλη τυχαία σελίδα

Το d (ο παράγοντας απόσβεσης) εκφράζει τη πιθανότητα σε κάθε σελίδα ο τυχαίος surfer να βαρεθεί και να αρχίσει από κάποια άλλη τυχαία σελίδα



Διατύπωση της επέκτασης με μεταπηδήσεις με τη μορφή πίνακα

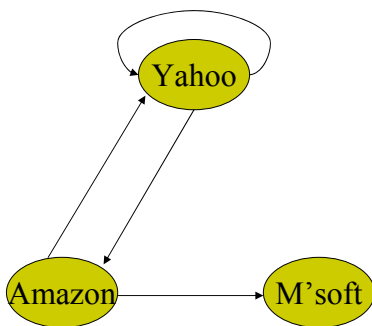
Έστω N σελίδες

- Έστω σελίδα j , με ένα σύνολο outlinks $O(j)$
 - $M_{ij} = 1/|O(j)|$ αν $j \rightarrow i$ and $M_{ij} = 0$ otherwise
- Η τυχαία μεταπήδηση είναι ισοδύναμη με το
 - Να προσθέσουμε ένα τυχαίο link από το j σε οποιαδήποτε άλλη σελίδα με $(1-\beta)/N$
 - Ελάττωση της πιθανότητας να ακολουθήσουμε ένα outlink από $1/|O(j)|$ σε $\beta/|O(j)|$
 - Ή ισοδύναμα: χρέωσε σε κάθε σελίδα ένα ποσοστό $(1-\beta)$ της τιμής της και κάνε κατανομή αυτού ομοιόμορφα



Αδιέξοδα

Οι σελίδες χωρίς outlinks για τον τυχαίο surfer



$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 1/15 \end{bmatrix}$$

y	=	1	1	0.787	0.648	0
a		1	0.6	0.547	0.430	...
m		1	0.6	0.387	0.333	0

Μη
στοχαστικό!



Χειρισμός αδιεξόδων (dead-end)

Μεταπήδηση

- Για αδιέξοδα, ακολούθησε τυχαία μεταπήδηση με πιθανότητα 1
- Τροποποίησε τον πίνακα

Ψαλίδισε τα αδιέξοδα και αναπροσάρμοσε το γράφο

- Προ-επεξεργασία του γράφου για σβήσιμο των αδιεξόδων
 - Πιθανών πολλαπλές επαναλήψεις
- Υπολογισμός $page\ rank$ στον ελαττωμένο γράφο
- Υπολογισμός προσεγγιστικών τιμών για αδιέξοδα μεταφέροντας τις τιμές από τον ελαττωμένο γράφο



Μια σελίδα μπορεί να έχει υψηλό PR αν:

- υπάρχουν πολλές σελίδες που δείχνουν σε αυτήν, ή
- όταν κάποιες σελίδες που δείχνουν σε αυτήν έχουν υψηλό PR

Και οι δύο περιπτώσεις έχουν σημασία:

Τιχ στη δεύτερη περίπτωση αν υπάρχει link από πχ [Yahoo!](#)



Content spam - Link spam

Google bombing:

Προσθήκη αναφορών που επηρεάζουν άμεσα το PR

Link farms:

Σελίδες που αναφέρονται η μία στην άλλη



PageRank συνέχεια ...



Αν δούμε το Web ως γράφο, θέλουμε να βρούμε τους σημαντικούς/κεντρικούς κόμβους



Με βάση το PageRank:

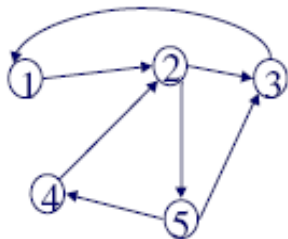
Ένας κόμβος είναι σημαντικός αν συνδέεται με σημαντικούς κόμβους

Μια ποσότητα σε κάθε σελίδα (κόμβο)

Η ποσότητα εξαρτάται από πόσες σελίδες δείχνουν σε αυτήν και μοιράζεται στις σελίδες που δείχνει (αναδρομικός ορισμός)



Παράδειγμα



Κάθε κόμβος μια αρχική τιμή PageRank την οποία μοιράζει ισοδύναμα στους κόμβους στους οποίους δείχνει

Πχ κάθε ακμή

Του κόμβου 2 έχει $\frac{1}{2}$

Του κόμβου 3 έχει 1 κλπ

Ισοδύναμα, η πιθανότητα μετάβασης σε κάποιον κόμβο

Random walks (τυχαίοι περίπατοι)

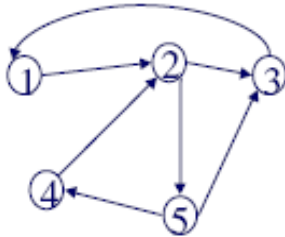
M ο Πίνακας Γειτνίασης (Πίνακας μετάβασης για αλυσίδες Markov)

r το διάνυσμα PageRank

$$r = M r$$



Παράδειγμα



M ο Πίνακας Γειτνίασης, r το διάνυσμα PageRank

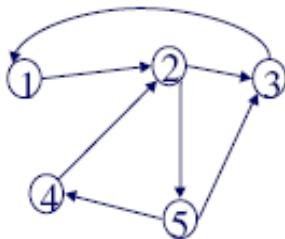
$$r = M r$$

r είναι το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή $\lambda = 1$

(είναι η μεγαλύτερη ιδιοτιμή, επειδή ο πίνακας είναι column-stochastic)



Teleport



Πίνακας A

$$A_{ij} = \beta M_{ij} + (1-\beta)/N$$

Fly-out probability

$$r = A r$$



Θεματικό PageRank (Topic-Specific PageRank)

Υπολογισμός δημοτικότητας (popularity) για κάποιο θέμα

- E.g., computer science, health

Bias the random walk

- Όταν ο τυχαίος περιπατητής teleports, επιλέγει μια σελίδα από ένα σύνολο S σελίδων του παγκόσμιου ιστού
 - S περιέχει μόνο σελίδες που είναι σχετικές με ένα θέμα
Πχ ., Open Directory (DMOZ) σελίδες για κάποιο θέμα (www.dmoz.org)
- Για κάθε σύνολο teleports S , διαφορετικό διάνυσμα r_S



HITS



Προβλήματα με τη χρήση της δομής των συνδέσεων του Web

Δεν αρκεί να δείχνουν πολλές συνδέσεις

- Μια σύνδεση δε σημαίνει απαραίτητα θετική γνώμη (αναγνώριση για τη σελίδα)
(κάποιες συνδέσεις διαφημίσεις, αλλά navigation, κλπ)
- Μια αυθεντία (authority) για κάποιο θέμα σπάνια θα έχει link σε αντίπαλη αυθεντία στον ίδιο τομέα
Οι αυθεντικές σελίδες σπάνια είναι περιγραφικές/αντιπροσωπευτικές



Ο αλγόριθμος HITS (Hyperlink-Induced Topic Search)

Για κάθε **θέμα**: δύο είδη σελίδων

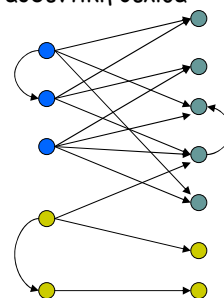
Αυθεντική (authority): Μια σελίδα που είναι αυθεντία σε ένα θέμα και αναγνωρίζεται ως τέτοια από άλλες σελίδες (δηλαδή, υπάρχουν πολλοί σύνδεσμοι σε αυτήν)

Κομβικοί (hubs): Μια σελίδα που αναφέρεται σε μια αυθεντική σελίδα

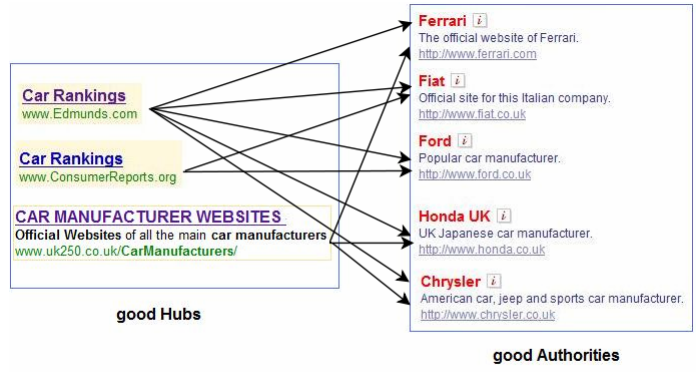
Βασική ιδέα:

Οι σελίδες που αναφέρονται από άλλες σελίδες συχνά πρέπει να είναι αυθεντίες (Authorities)

Οι σελίδες που αναφέρουν πολλές άλλες σελίδες πρέπει να είναι *καλά* κομβικά σημεία (hubs)



Κομβικοί Αυθεντικοί



Query: Top automobile makers



Βασική ιδέα του HITS

- Καλές αυθεντίες είναι αυτές στις οποίες αναφέρονται καλά κομβικά σημεία
- Καλά κομβικά σημεία είναι αυτά τα οποία αναφέρονται σε καλές αυθεντίες

Αναδρομική έκφραση

Σημείωση: Αναθέτει σε κάθε σελίδα δύο τιμές για κάθε θέμα - διάνυσμα h (hub) και a (authority)



Το web ως ένας κατευθυνόμενος γράφος

Κόμβοι: ιστοσελίδες

Ακμή από A στον B: η ιστοσελίδα A έχει έναν υπερ-σύνδεσμο στην ιστοσελίδα B

Ο αλγόριθμος χωρίζεται σε 2 φάσεις:

Φάση I: (δειγματοληπτικό στάδιο) ένα σύνολο σελίδων που αποτελεί το *βασικό σύνολο* για κάποιο θέμα

Φάση II: (επαναληπτικό στάδιο) επεξεργασία του βασικού συνόλου για τον εντοπισμό καλών αυθεντικών και καλών κομβικών ιστοσελίδων



Φάση I: Υπολογισμός βασικού συνόλου

1. Υπολογισμός αρχικού συνόλου: *σύνολο-ρίζα*

Κλασικοί μέθοδοι: πχ ανάκτηση όλων των σελίδων που περιέχουν τις λέξεις κλειδιά

(περιμένουμε ότι θα περιέχει (τουλάχιστον) αναφορές προς σχετικές σελίδες)



**Φάση I: Υπολογισμός βασικού συνόλου
(διεύρυνση του συνόλου ρίζα)**

2. + Σελίδες-σύνδεσμοι:

- Σελίδα που είτε συμπεριλαμβάνει σύνδεσμο που να αναφέρεται σε έναν κόμβο p στο σύνολο ρίζα (p είναι αυθεντία) είτε
- Ένας κόμβος p στο σύνολο ρίζα (p είναι κομβικό σημείο) περιέχει σύνδεσμο που αναφέρεται σε αυτήν

Βασικό Σύνολο: διεύρυνση του συνόλου-ρίζα ώστε να περιλαμβάνει και τις σελίδες συνδέσμων - *Βασικές Ιστοσελίδες*



Φάση II: Ποιες βασικές ιστοσελίδες είναι κόμβοι και αυθεντίες

Κάθε βασική σελίδα p δύο τιμές:

h_p - **Συντελεστής Κομβικού Ρόλου** (πολλούς δείκτες σε αυθεντικές)

a_p - **Συντελεστής Αυθεντικότητας** (πολλοί δείκτες από κομβικές σε αυτήν)



Βασική διαφορά από τον Page Rank

- Δύο τιμές ανά σελίδα (αυθεντία - κομβικό σημείο)
- Θεματικά υποσύνολα του web γράφου - ξεκινάμε από το βασικό σύνολο



Φάση II: Ποιες βασικές ιστοσελίδες είναι κόμβοι και αυθεντίες

Αρχικοποίηση, $\forall p, h_p = 1$ και $a_p = 1$

Επαναληπτικά, αυξάνεται

$$a_p = \sum h_q$$

Βασικές σελίδες q που δείχνουν στην p

$$h_p = \sum a_q$$

Βασικές σελίδες q στις οποίες δείχνει η p



Αναπαράσταση με πίνακες

Έστω το βασικό σύνολο σελίδων $\{1, 2, \dots, n\}$

Πίνακας Γειτνίασης (adjacency matrix) $B: n \times n$

$B[i, j] = 1$ αν η σελίδα i περιέχει σύνδεσμο που δείχνει στη σελίδα j

Έστω $h = \langle h_1, h_2, \dots, h_n \rangle$ το διάνυσμα συντελεστών κομβικών ρόλων
και $a = \langle a_1, a_2, \dots, a_n \rangle$ το διάνυσμα συντελεστών αυθεντικότητας

(αντίστοιχο του r vector)



Οι κανόνες ενημέρωσης

Αρχικά

$$h = B a$$

$$a = B^T h$$

1η επανάληψη

$$h = B B^T h = (B B^T) h$$

$$a = B^T B a = (B^T B) a$$

2η επανάληψη

$$h = (B B^T)^2 h$$

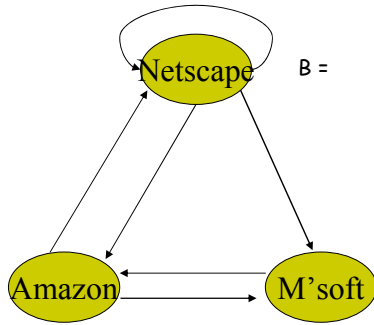
$$a = (B^T B)^2 a$$

Σύγκλιση στα ιδιοδιανύσματα του BB^T και B^TB αν κανονικοποιηθούν αρχικά οι συντελεστές

Ο Αλγόριθμος HITS



Διατύπωση με την μορφή πίνακα (παράδειγμα)



$$B = \begin{matrix} & n & m & a \\ \begin{matrix} 1 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 1 \\ 0 \end{matrix} \end{matrix}$$

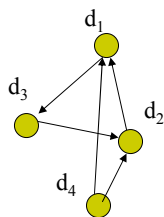
$$B^T = \begin{matrix} & n & m & a \\ \begin{matrix} 1 \\ 1 \\ 1 \end{matrix} & \begin{matrix} 0 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 1 \\ 0 \end{matrix} \end{matrix}$$

$$B B^T = \begin{matrix} & n & m & a \\ \begin{matrix} 3 \\ 1 \\ 2 \end{matrix} & \begin{matrix} 1 \\ 1 \\ 0 \end{matrix} & \begin{matrix} 2 \\ 1 \\ 0 \end{matrix} & \begin{matrix} 2 \\ 0 \\ 2 \end{matrix} \end{matrix}$$

$$h = B B^T h$$

$$\begin{matrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{matrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix} \dots$$

Ο Αλγόριθμος HITS



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

Πίνακας Γειτνίασης

$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$

$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

Αρχικές Τιμές: $a=h=1$

Iterate

Normalize:

$$\bar{h} = A\bar{a}; \quad \bar{a} = A^T\bar{h}$$

$$\bar{h} = AA^T\bar{h}; \quad \bar{a} = A^T A\bar{a}$$

$$\sum_i a(d_i)^2 = \sum_i h(d_i)^2 = 1$$

Πάλι ιδιοδιανύσματα ...



Προβλήματα

- **Drifting**: όταν ένα κομβικό σημείο περιέχει πολλά θέματα
- **Topic hijacking**: όταν πολλές σελίδες από το ίδιο web site δείχνουν στο ίδιο δημοφιλές κόμβο

Λίγα ακόμη για τις μηχανές
αναζήτησης





Anchor Text

Το κείμενο που υπάρχει στα links έχει διαφορετική αντιμετώπιση
Οι περισσότερες μηχανές αναζήτησης το συσχετίζουν με τη σελίδα στην οποία εμφανίζεται

Google και με τη σελίδα στην οποία δείχνει

- Πιο ακριβείς πληροφορίες για τις σελίδες που δείχνουν παρά για τις σελίδες στις οποίες εμφανίζονται
- Μπορεί να δείχνουν σε σελίδες που δεν έχουν κείμενο αλλά εικόνες, προγράμματα, κλπ



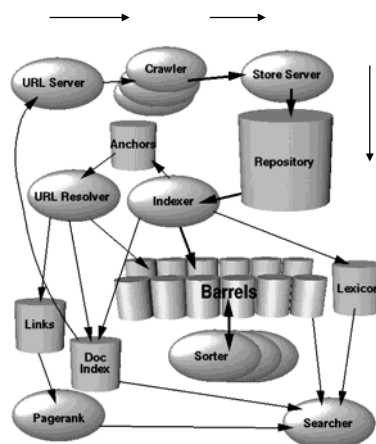
Most of Google is implemented in C or C++ for efficiency and can run in either Solaris or Linux.

The **web crawling** (downloading of web pages) is done by several distributed crawlers.

There is a **URLserver** that sends lists of URLs to be fetched to the crawlers.

The web pages that are fetched are then sent to the **storeserver**.

The **storeserver** then compresses and stores the web pages into a **repository**.



Google: Αρχιτεκτονική



Every web page has an associated ID number called a **docID** which is assigned whenever a new URL is parsed out of a web page.

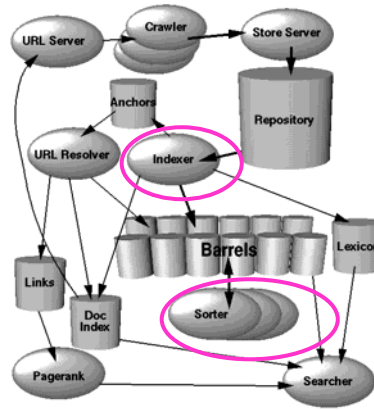
The indexing function is performed by the indexer and the sorter.

The **indexer** reads the repository, uncompresses the documents, and parses them.

document → a set of word occurrences called **hits**.

Hits: word, position in document, an approximation of font size, and capitalization.

The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index.



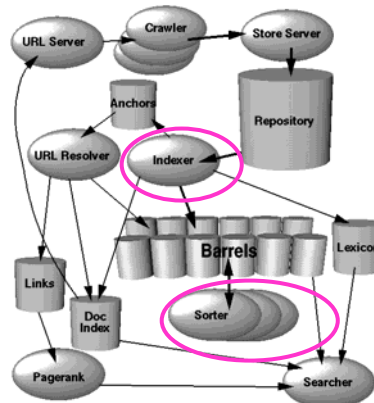
Google: Αρχιτεκτονική



Indexer:

It parses out all the links in every web page and stores important information about them in an anchors file.

This file contains enough information to determine where each link points from and to, and the text of the link.



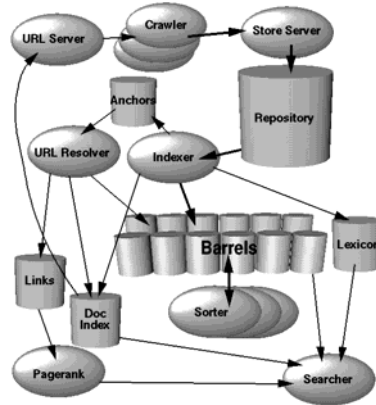
Google: Αρχιτεκτονική



URLresolver relative URLs ->
absolute URLs -> docIDs.

The sorter takes the barrels,
which are sorted by docID
and resorts them by wordID
to generate the inverted
index.

+ lexicon

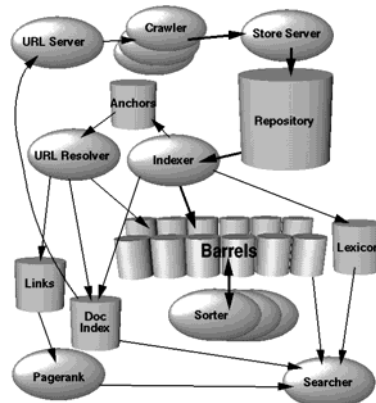


Google: Αρχιτεκτονική

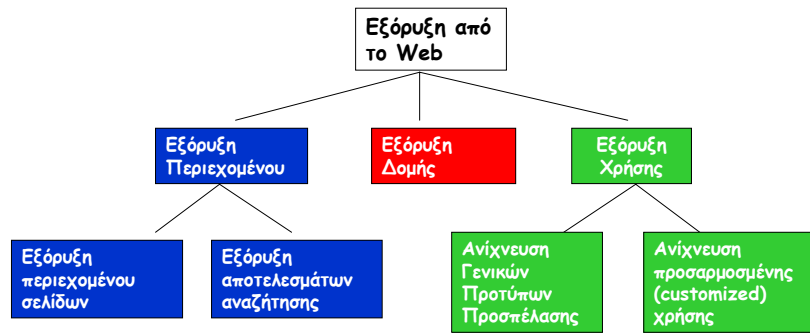


The searcher is run by a web
server

uses the lexicon built by
DumpLexicon together with
the inverted index and the
PageRanks to answer queries.



Κατηγορίες Εξόρυξης από το Web

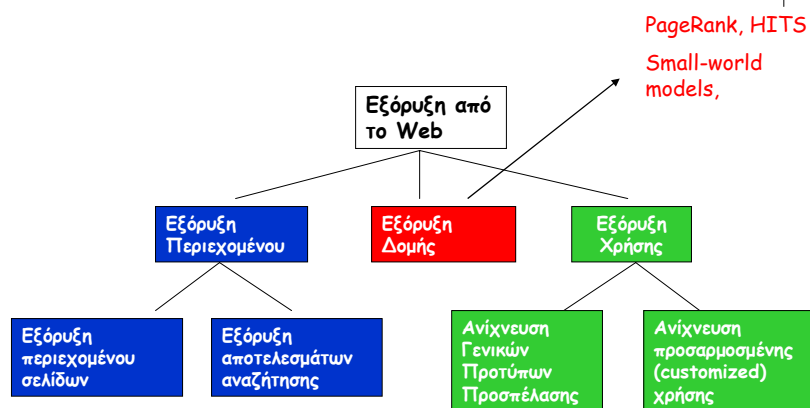


Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ

83

Κατηγορίες Εξόρυξης από το Web



Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΠΑΓΚΟΣΜΙΟΣ ΙΣΤΟΣ

84



Τέλος