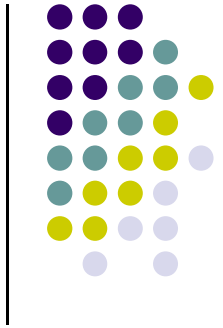


Μηχανές Αναζήτησης

Μηχανές Αναζήτησης

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

Χρήσιμες για τον εντοπισμό σελίδων που περιέχουν συγκεκριμένες λέξεις κλειδιά

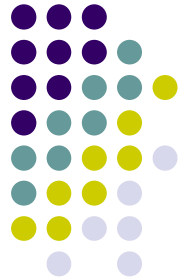
Προβλήματα

- Ένα θέμα μπορεί να περιέχεται σε 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 τι 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

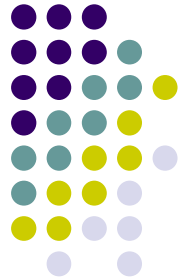
Βασική Ιδέα (συνέχεια)

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

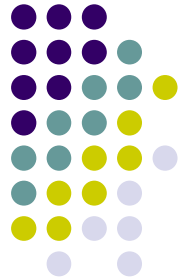
Κάθε σελίδα μια ποσότητα που χαρακτηρίζει τη σημαντικότητα της (αυτή η ποσότητα καλείται **page rank**)

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) *y* έχει *n* outlinks, κάθε link παίρνει *y/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,

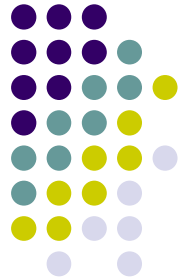
τότε όλα το PageRank PR( ) τους θα μαζευόταν στο A

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

- 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 Α:

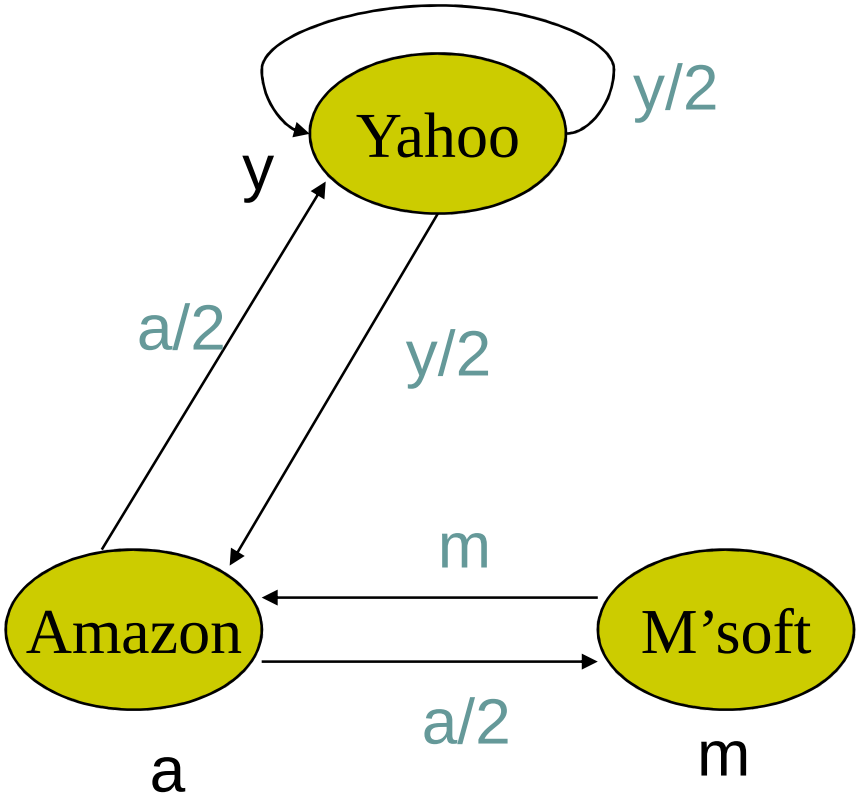
Έστω ότι η Α έχει τις σελίδες T1, ...,Tn που δείχνουν σε αυτήν (δηλαδή, αναφορές)

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

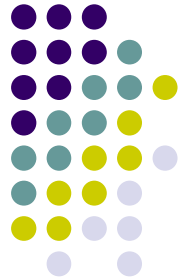
$$PR(A) = PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)$$

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)

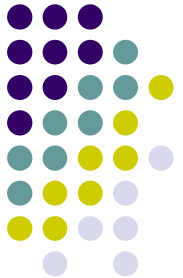
- Επιπρόσθετος περιορισμός για μοναδικότητα της λύσης
  - y+a+m = 1 (το συνολικό PR που μοιράζεται στις σελίδες)
  - y = 2/5, a = 2/5, m = 1/5

Διατύπωση με την μορφή πίνακα

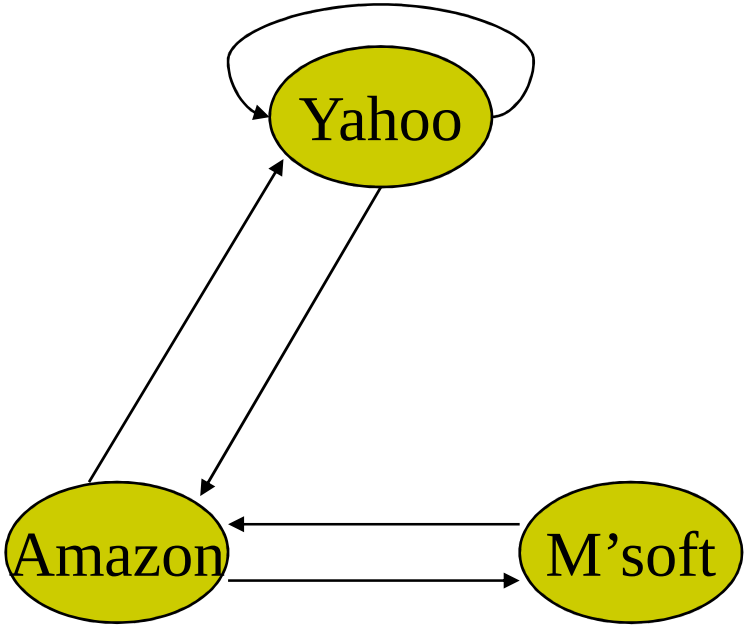
- Ο πίνακας **M** έχει μια γραμμή και μια στήλη για κάθε web σελίδα (πίνακας γειτνίασης)

- Έστω ότι η σελίδα j έχει n outlinks
  - Αν j -> 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

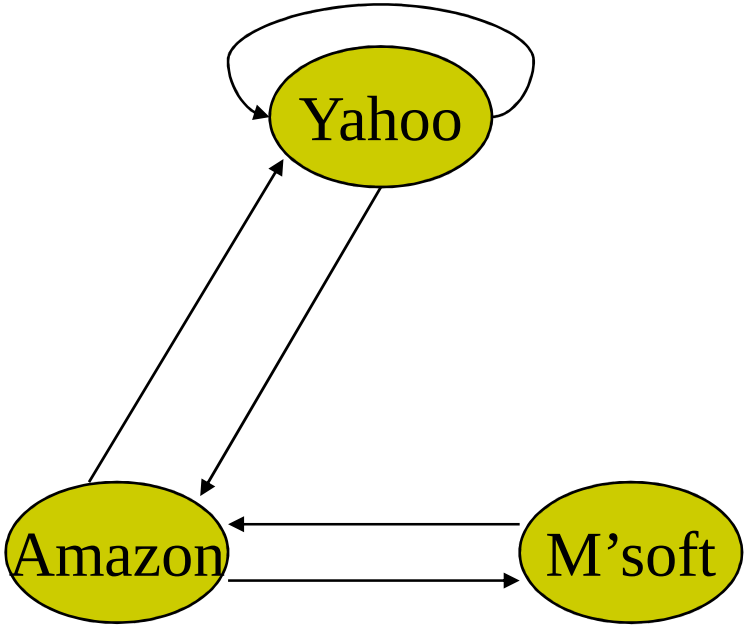
- Έστω **r** ένα διάνυσμα με μια εγγραφή web σελίδα
  - r_i είναι η σημαντικότητα (PR) της σελίδας i
  - r: rank vector

[PR(y)

PR(a)

PR(m)]

PR Διάνυσμα (παράδειγμα)

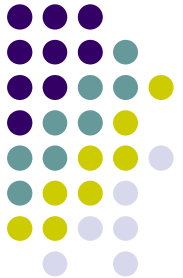


$$y = y/2 + a/2$$
$$a = y/2 + m$$
$$m = a/2$$

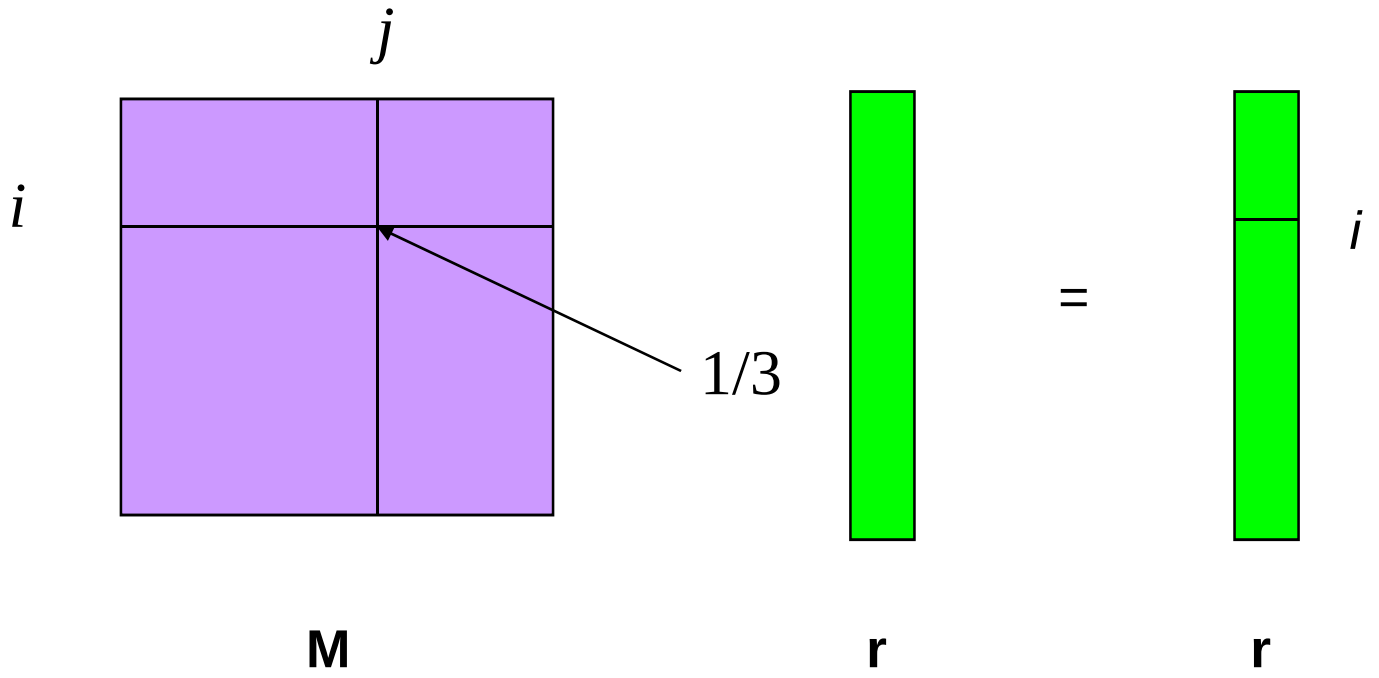
	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

r = Mr

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$



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

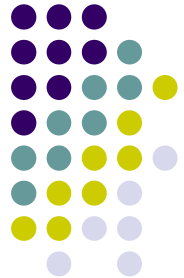


Ιδιοδιανύσματα (eigenvectors)

- 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 λ = 1)



Power Iteration method – Επαναληπτική Μέθοδο

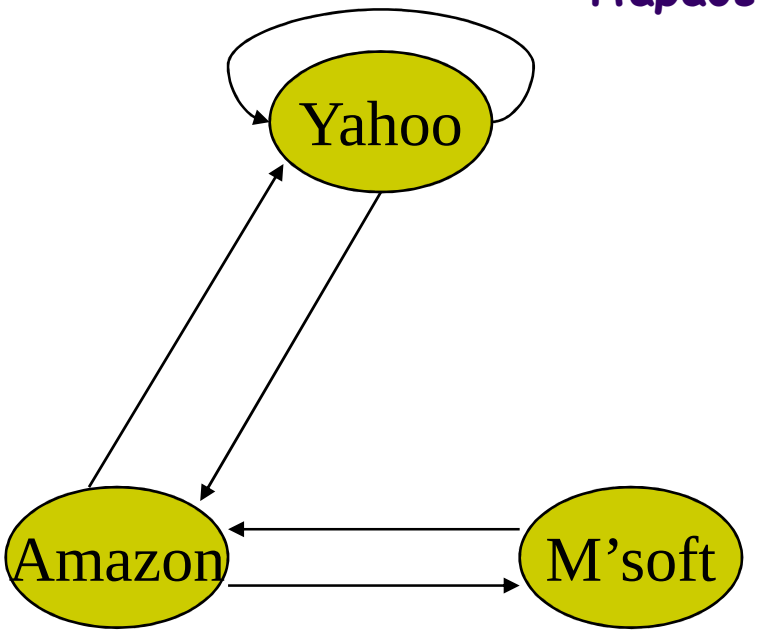
Ένα απλό επαναληπτικό σχήμα (relaxation)
Έστω N web σελίδες

Αρχικοποίηση: $\mathbf{r}^0 = [1/N,\ldots,1/N]^T$
Επανάληψη: $\mathbf{r}^{k+1} = \mathbf{M}\mathbf{r}^k$
Τερματισμός όταν $|\mathbf{r}^{k+1} - \mathbf{r}^k|_1 < \varepsilon$

$|\mathbf{x}|_1 = \sum_{1 \cdot i \cdot N} |x_i|$ 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 **p**(t) το διάνυσμα του οποίου το i-οστό στοιχείο είναι η πιθανότητα ο περιηγητής να είναι στη σελίδα i τη χρονική στιγμή t

p(t) κατανομή πιθανότητας (probability distribution) στις σελίδες

The stationary distribution

- Που είναι ο περιηγητής τη χρονική στιγμή t+1?
  - Ακολουθεί ένα link uniformly at random
  - **p**(t+1) = **M** **p**(t)

- Έστω ότι ο τυχαίος περίπατος φτάνει μια κατάσταση όπου **p**(t+1) = **M** **p**(t) = **p**(t)
  - Τότε **p**(t) ονομάζεται stationary distribution για τον τυχαίο περίπατο

- Επειδή ο πίνακας **r** ικανοποιεί την **r** = **Mr**
  - είναι stationary distribution για τον τυχαίο περιηγητή

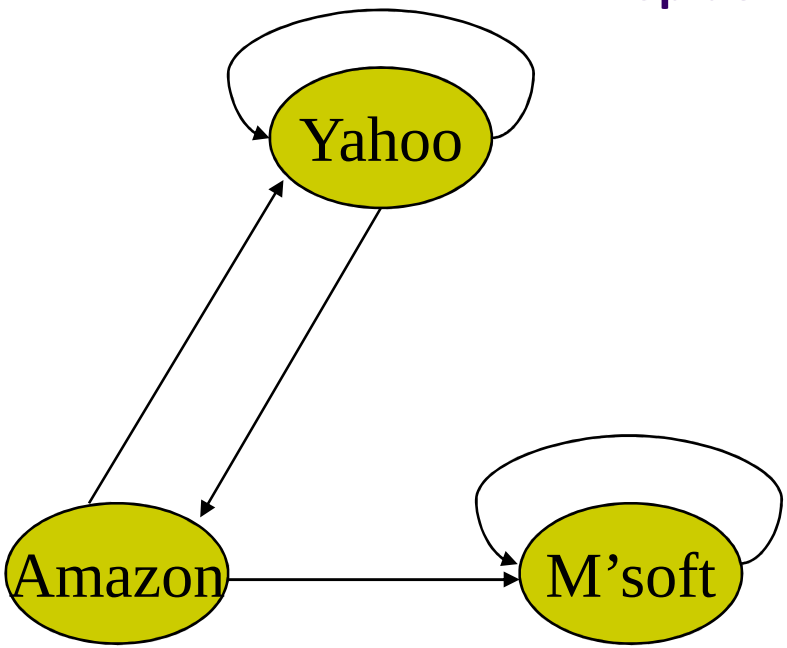
Βασικό αποτέλεσμα από τη θεωρία τυχαίων περιπάτων (και Markov processes):

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 outlink
- Με πιθανότητα 1-β πετάγεται σε κάποια άλλη σελίδα τυχαία (d: dumping factor)
- 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 Α:

PR(A) = PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)

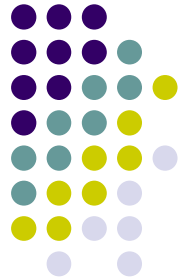
Ορισμός με τον παράγοντας απόσβεσης d (damping factor) μεταξύ του 0 και του 1

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

Ώστε το άθροισμα να είναι 1 –> 1-d/N
Ο πρώτος παράγοντας λέει ότι με την ίδια πιθανότητα διαλέγω οποιαδήποτε σελίδα

όπου d, είναι 1 - β

Επεκτάσεις (τυχαίο άλμα)



- Κατασκευή του NxN πίνακα A

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

- Ο **A** είναι στοχαστικός πίνακας

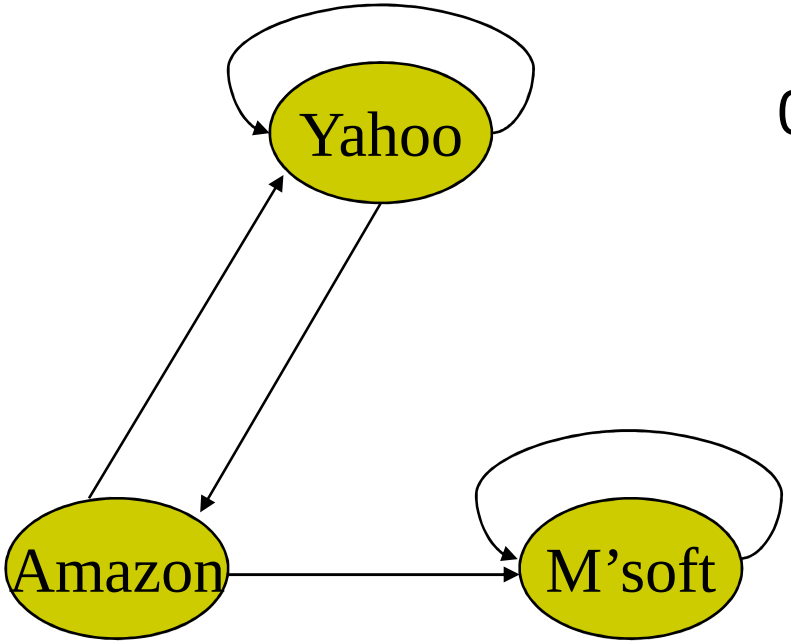
- Το page rank διάνυσμα r είναι το βασικό ιδιοδιάνυσμα αυτού του πίνακα

$$\mathbf{r} = \mathbf{Ar}$$

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

Επεκτάσεις (τυχαίο άλμα)

Παράδειγμα (d=0.8)



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

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

$$\begin{matrix} y \\ a \\ m \end{matrix} = \begin{matrix} 1 & 1.00 & 0.84 & 0.776 & & 7/11 \\ 1 & 0.60 & 0.60 & 0.536 & \ldots & 5/11 \\ 1 & 1.40 & 1.56 & 1.688 & & 21/11 \end{matrix}$$

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

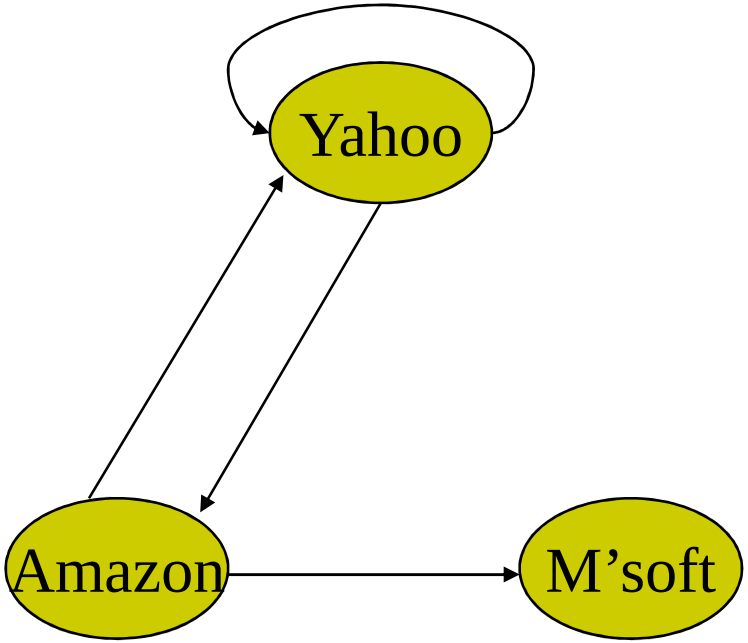
Έστω N σελίδες

- Έστω σελίδα j, με ένα σύνολο outlinks O(j)
  - $M_{ij} = 1/|O(j)|$ αν j -> 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



Επεκτάσεις (αδιέξοδα)

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

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

Μη στοχαστικό!

$$\begin{matrix} y \\ a \\ m \end{matrix} = \begin{matrix} 1 & 1 & 0.787 & 0.648 & & 0 \\ 1 & 0.6 & 0.547 & 0.430 & \ldots & 0 \\ 1 & 0.6 & 0.387 & 0.333 & & 0 \end{matrix}$$

Χειρισμός αδιεξόδων (deadend)

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

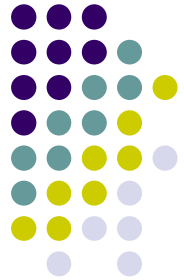
Μια σελίδα μπορεί να έχει υψηλό PR αν:

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

Και οι δύο περιπτώσεις έχουν σημασία:

Πχ στη δεύτερη περίπτωση αν υπάρχει link από πχ Yahoo!

Spam



Content spam – Link spam

Google bombing:

Προσθήκη αναφορών που επηρεάζουν άμεσα το PR

Link farms:

Σελίδες που αναφέρονται η μία στην άλλη

Θεματικό PageRank (Topic-Specific PageRank)

Υπολογισμός δημοτικότητας (popularity) για κάποιο θέμα

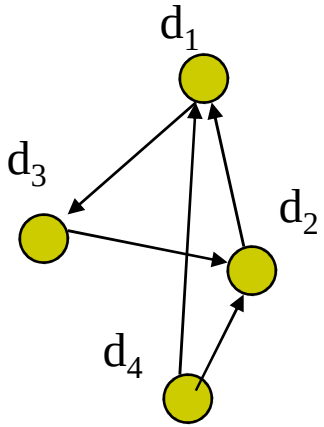
- E.g., computer science, health

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 teleport S, διαφορετικό διάνυσμα $\mathbf{r}_S$

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

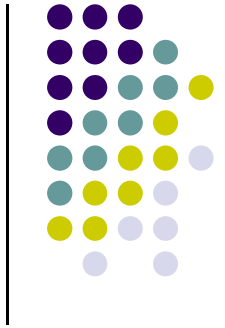
Normalize:

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

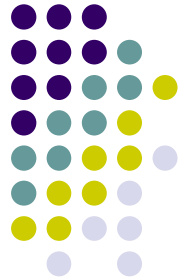
$$\vec{h} = A\vec{a}; \quad \vec{a} = A^T\vec{h}$$

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

Πάλι ιδιοδιανύσματα …



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

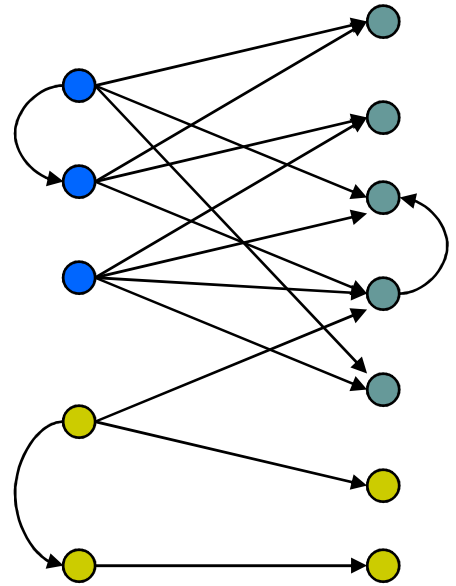
Ο αλγόριθμος HITS (Hyperlink-Induced Topic Search)

Για κάθε θέμα: δύο είδη σελίδων

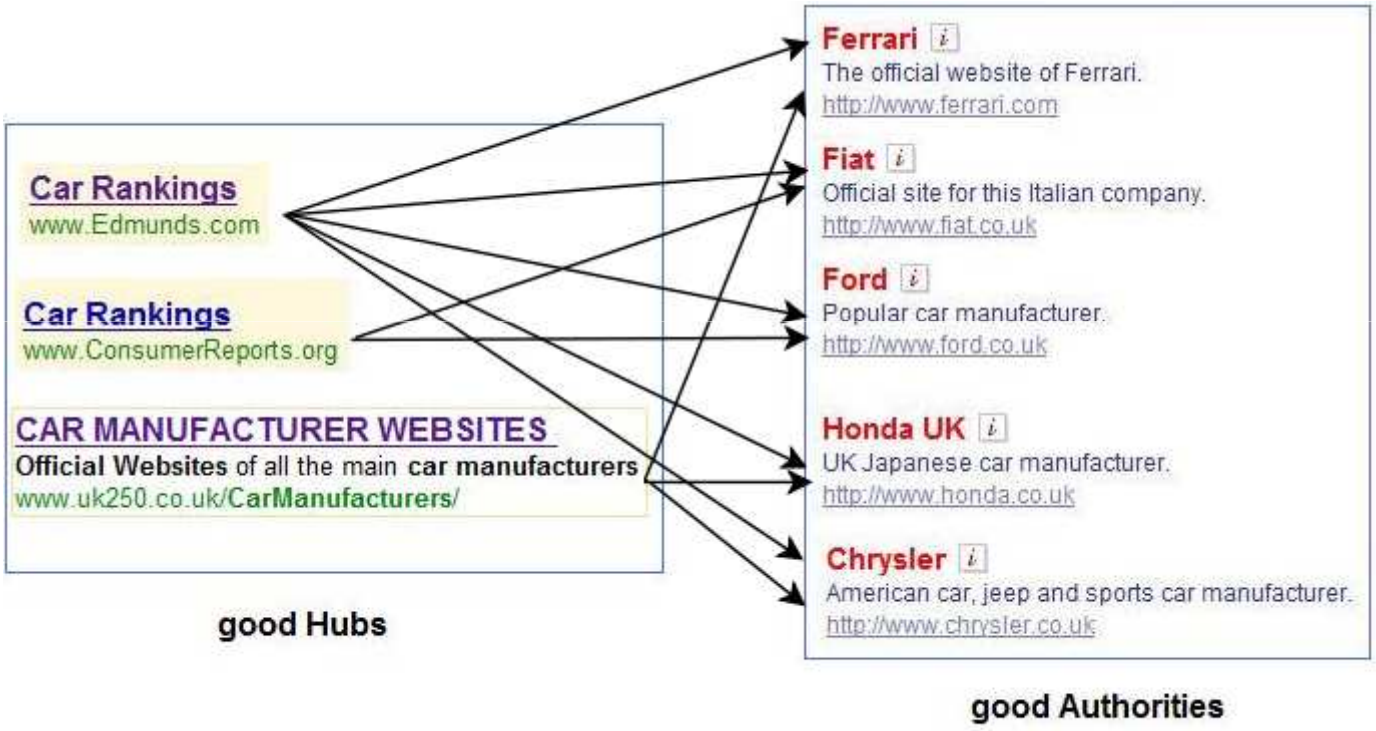
Αυθεντική (authority): Μια σελίδα που είναι αυθεντία σε ένα θέμα και αναγνωρίζεται ως τέτοια από άλλες σελίδες (δηλαδή, υπάρχουν πολλοί σύνδεσμοι σε αυτήν)

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)



Ο Αλγόριθμος HITS



Query: Top automobile makers

Βασική ιδέα του HITS

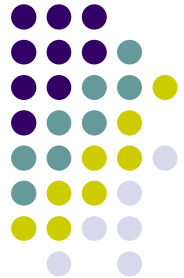
- Καλές αυθεντίες είναι αυτές στις οποίες αναφέρονται καλά κομβικά σημεία

- 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 α (authority)

Ο Αλγόριθμος HITS



Το web ως ένας κατευθυνόμενος γράφος

Κόμβοι: ιστοσελίδες

Ακμή από Α στον Β: η ιστοσελίδα Α έχει έναν υπερ-σύνδεσμο στην ιστοσελίδα Β

Ο αλγόριθμος χωρίζεται σε 2 φάσεις:

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

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

Φάση Ι: 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 Ι: 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 συνδέσμους – *Βασικές Ιστοσελίδες*

Φάση ΙΙ: Ποιες βασικές ιστοσελίδες είναι κόμβοι και αυθεντίες

Κάθε βασική σελίδα p δύο τιμές:

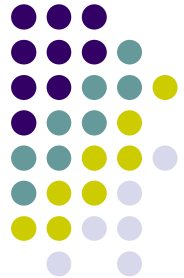
h_p - Συντελεστής Κομβικού Ρόλου (πολλούς δείκτες σε αυθεντικές)

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

Ο Αλγόριθμος HITS

Βασική διαφορά από τον Page Rank

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



Φάση ΙΙ: Ποιες βασικές ιστοσελίδες είναι κόμβοι και αυθεντίες

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

Επαναληπτικά, αυξάνεται

$a_p = \Sigma\ h_q$

Βασικές σελίδες q που δείχνουν στην p

$h_p = \Sigma\ \alpha_q$

Βασικές σελίδες q στις οποίες δείχνει η p

Αναπαράσταση με πίνακες

Έστω το βασικό σύνολο σελίδων {1, 2, ..., n}

Πίνακας Γειτνίασης (adjacency matrix) B: n x n

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

Έστω h = $<h_1, h_2, \ldots, h_n>$ το διάνυσμα συντελεστών κομβικών ρόλων

και α = $<\alpha_1, \alpha_2, \ldots, \alpha_n>$ το διάνυσμα συντελεστών αυθεντικότητας

(αντίστοιχο του r vector)

Ο Αλγόριθμος HITS

Οι κανόνες ενημέρωσης

Αρχικά

$h = B\ a$ $\qquad\qquad$ $a = B^T\ h$

1η επανάληψη

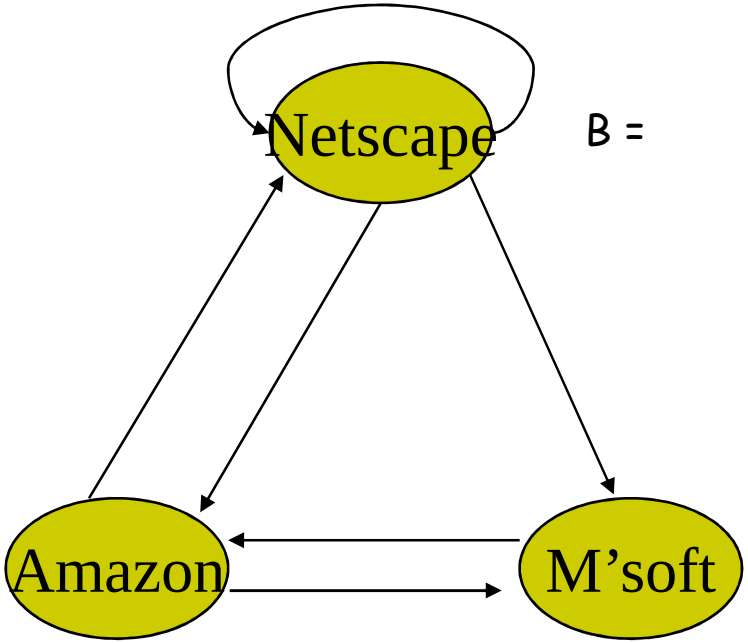
$h = B\ B^T\ h = (B\ B^T)\ h$ $\qquad\qquad$ $a = B^T\ B\ a = (B^T\ B)\ a$

2η επανάληψη

$h = (B\ B^T)^2\ h$ $\qquad\qquad$ $a = (B^T\ B)^2\ a$

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

Διατύπωση με την μορφή πίνακα (παράδειγμα)



B =

n	m	a
1	1	1
0	0	1
1	1	0

B^T =

n	m	a
1	0	1
1	0	1
1	1	0

$B\,B^T$ =

n	m	a
3	1	2
1	1	0
2	0	2

$h = BB^T h$

$$\begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix} \quad \ldots$$

Προβλήματα

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

Anchor Text

Το κείμενο που υπάρχει στα links έχει διαφορετική αντιμετώπιση
Οι περισσότερες μηχανές αναζήτησης το συσχέτιζαν 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