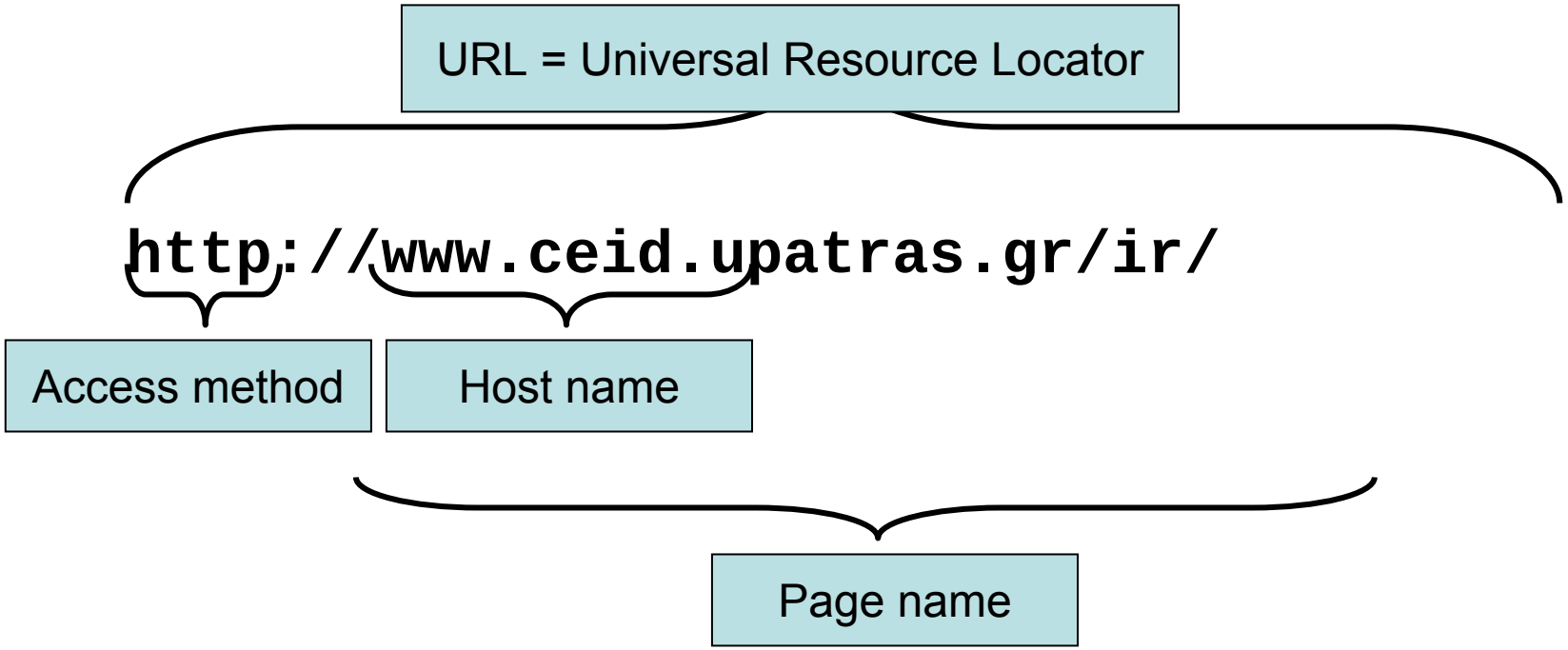
Αναζήτηση στο Διαδίκτυο

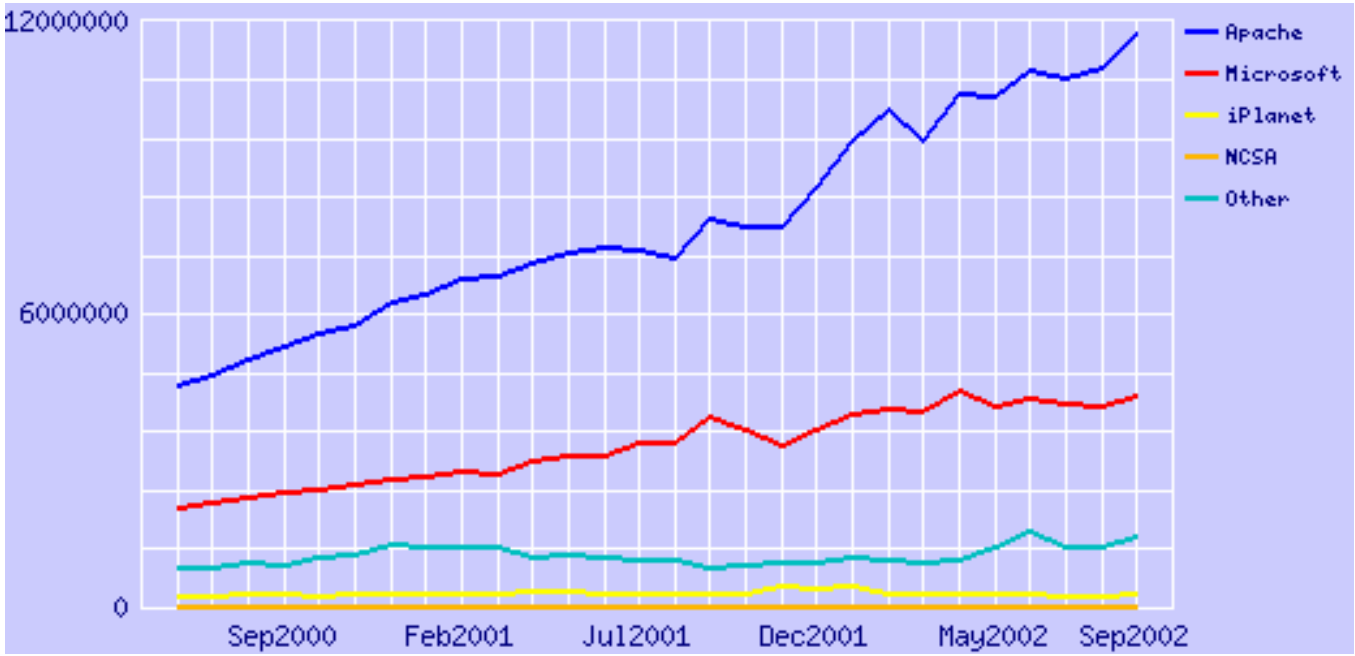
Εισαγωγή


www
url
καταλόγους
(π.χ. Yahoo)
Μηχανές
Αναζήτησης
τεράστιος, μη ομογενής
μεταβάλλεται ραγδαία
επικοινωνιακό κόστος


URL = Universal Resource Locator
http://www.ceid.upatras.gr/ir/
Access method
Host name
Page name

Εισαγωγή – Web

- Τεράστιο μέγεθος
  - 2-10Β στατικές σελίδες, διπλασιαζόμενες κάθε 8-12 μήνες
  - Μέγεθος Λεξικού: 10-100άδες εκατομμύρια λέξεις



http://www.netcraft.com/Survey

Εισαγωγή – Web

- Γλώσσες/Κωδικοποιήσεις:
  - Εκατοντάδες γλώσσες, W3C κωδικοποιήσεις: 55 (Ιουλ. 01)
  - Σελίδες (1997): Αγγλικές 82%, Επόμενες 15: 13%
- Μεγάλος Ρυθμός Αλλαγής στις Σελίδες
- Ανομοιογένεια στη μορφή:
  - Εκατομμύρια άνθρωποι δημιουργούν σελίδες με τη δικιά τους γραμματική, λεξικό, στυλ
  - Πολλές φορές οι σελίδες εξυπηρετούν εμπορικούς σκοπούς (marketing)
- Μεγάλος Ρυθμός Αλλαγής στις Σελίδες
- Επανάληψη της ίδιας πληροφορίας
  - Συντακτική επανάληψη (30-40% πανομοιότυπες)
  - Σημασιολογική ομοιότητα?
- Υψηλή Συνεκτικότητα
  - Κατά μέσο όρο ~8 υπερδεσμοί/σελίδα
- Πολύπλοκη τοπολογία γράφου
  - Bow-tie τοπολογία

Εισαγωγή – Web

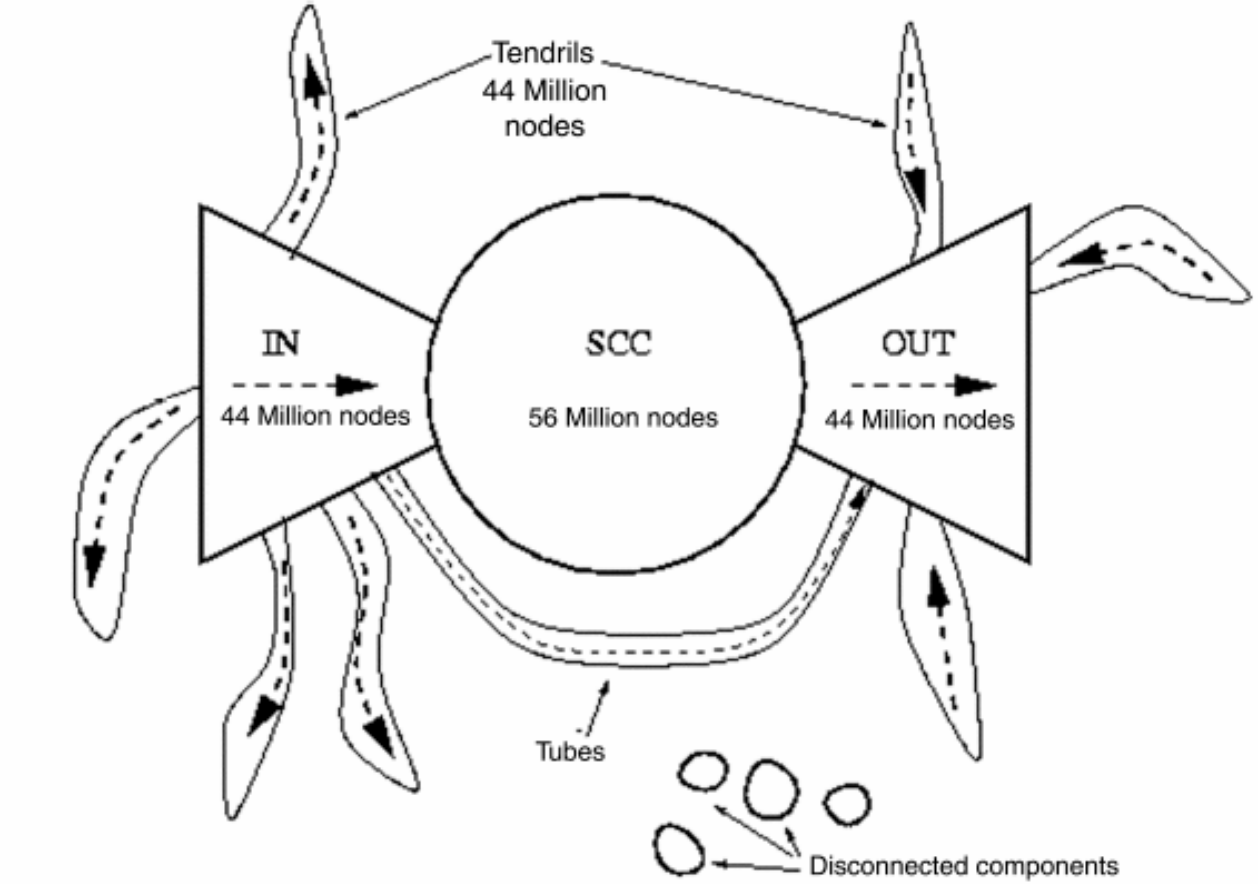
- Κακώς σχηματισμένες ερωτήσεις
  - Μικρές σε πλήθος όρων
  - Ανακριβείς όροι
  - Μη βέλτιστη σύνταξη (80% ερωτήματα χωρίς τελεστή)
  - Χαμηλή προσπάθεια
- Μεγάλη απόκλιση σε
  - ανάγκες
  - Επίπεδα αναμονής
  - Γνώση
  - Bandwidth
- Τυπική συμπεριφορά
  - Εστίαση στην πρώτη οθόνη, όχι feedback, ακολούθηση υπερδεσμών

Ποσότητες που μπορούν να μετρηθούν

- Το σχετικό μέγεθος των μηχανών αναζήτησης
  - προβλήματα
    - Επέκταση κειμένων: π.χ. το Google δεικτοδοτεί σελίδες που δεν έχουν γίνει crawl δεικτοδοτώντας anchor-text.
    - Περιορισμός στα κείμενα: Μερικές μηχανές περιορίζουν το τι δεικτοδοτείται (πρώτες *n* λέξεις, μόνο σχετικές λέξεις κ.λ.π.)
- Η κάλυψη μίας μηχανής σε σχέση με κάποια άλλη διεργασία crawling.

Τεχνικές Μέτρησης

- Random queries
- Random searches
- Random IP addresses
- Random walks



Graph structure in the Web, *Computer Networks,2000.*

Andrei Broder, Ravi Kumar, et al.

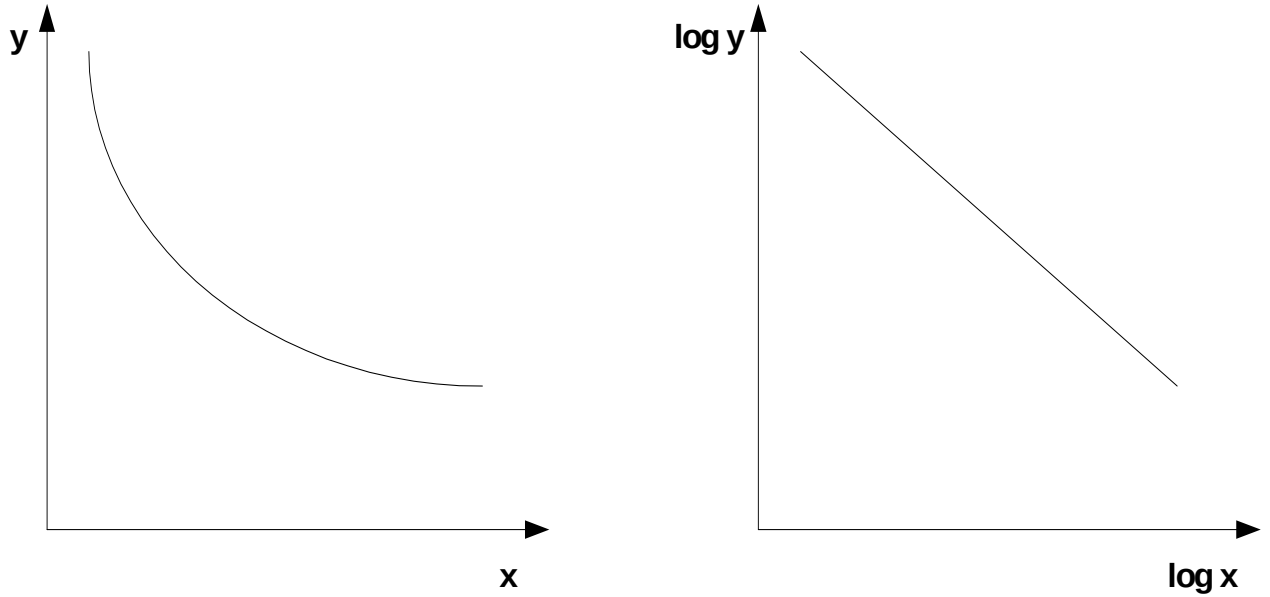
Εισαγωγή - Web

- Για τυχαίες σελίδες *p1*,*p2*:
  - Pr[*p1* να προσπελαύνεται από *p2*] ~ 1/4
- Μέγιστη απόσταση μεταξύ 2 SCC κόμβων: >28
- Μέση κατευθυνόμενη απόσταση μεταξύ 2 κόμβων: ~16
- Μέση μη κατευθυνόμενη απόσταση: ~7

Power Laws - Γενικά

- Δύο ποσότητες x και y συνδέονται με έναν power law όταν

$$y \approx x^{-c} \Leftrightarrow \log y = -c * \log x$$



Ένας γνωστός power law

- Κατανομή Zipf

  y : συχνότητα λέξης σε κείμενο

  x : ο x-οστός πιο συχνός όρος

  Power law για c=1

$$y \approx 1/x$$

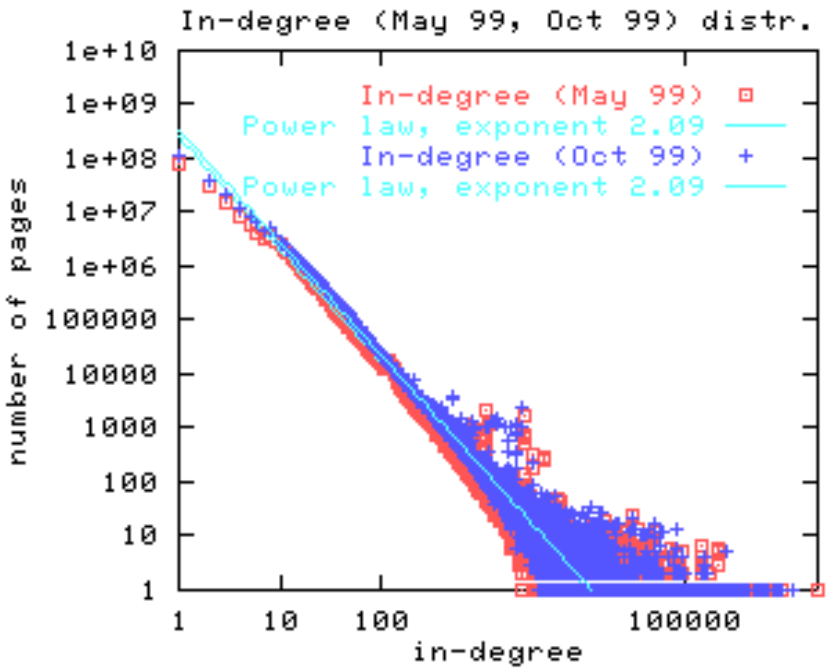
Power laws και στο Web?

- Broder et. al. 1999

  y = #links που εισέρχονται σε σελίδα *i*

  x = #σελίδων με d εισερχόμενα links

$$y \approx x^{-2.09}$$



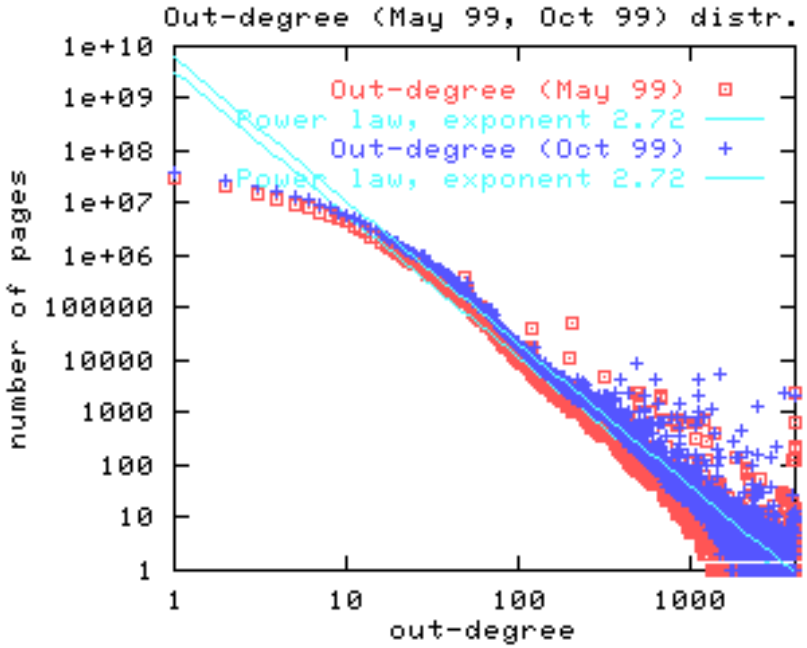
Power laws και στο Web?

(συνέχεια)

y = #links που εξέρχονται από σελίδα *i*

x = #σελίδων με d εξερχόμενα links

$$y \approx x^{-2.72}$$

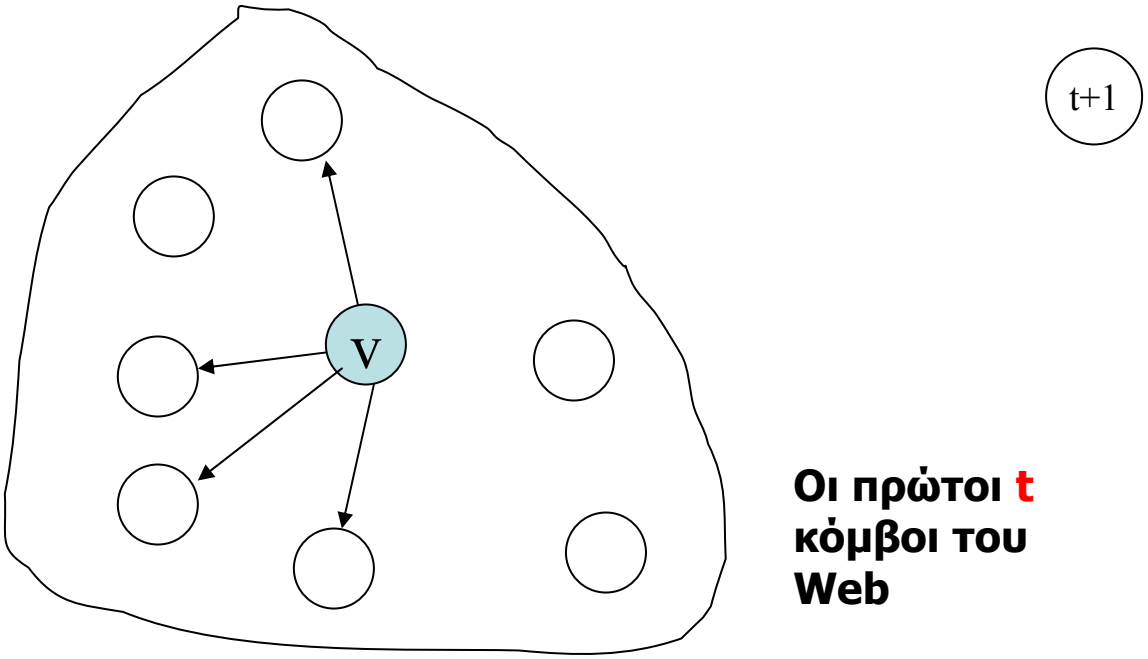


Χρησιμότητα Παρατήρησης

- Βοηθάει στην κατανόηση και πρόβλεψη της εξέλιξης του Web
- Βοηθάει στην κατασκευή νέων αλγορίθμων ταξινόμησης
- Εκτέλεση προσομοιώσεων σε σχέση με το Web
- Μοντελοποίηση του Web

Μοντελοποίηση Γραφήματος του Web

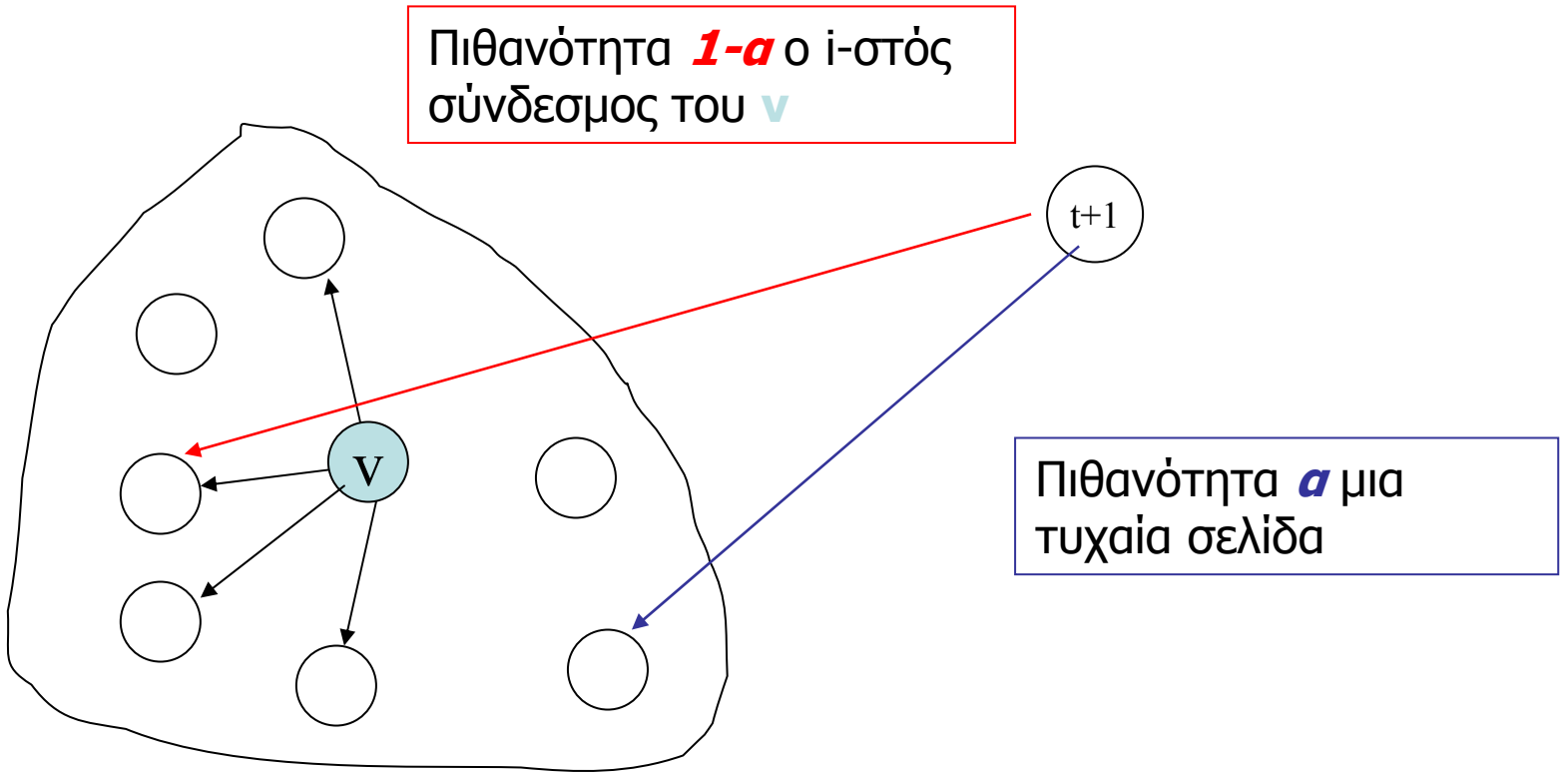
- Kumar et. al. *Stochastic models for the Web Graph, FOCS 2000*



Οι πρώτοι t κόμβοι του Web

Μοντελοποίηση Γραφήματος του Web

- Για τον t+1 φτιάξε d συνδέσμους d>1
- Πως επιλέγεται ο ι-στος σύνδεσμος?



Μοντελοποίηση Γραφήματος του Web

- Όταν δημιουργείται μια σελίδα αυτή ανήκει σε ένα θέμα.
  - Μας ενδιαφέρει να αντιγράψουμε τους συνδέσμους ενός Hub στο θέμα
  - Ή να εισάγουμε νέες ιδέες

- Το μοντέλο ακολουθεί Power laws!
  - Το μέσο πλήθος των σελίδων με βαθμό d είναι:

$$\Theta\left(d^{-(2-\alpha)/(1-\alpha)}\right)$$

Επεκτάσεις

- Εμπορικά πιο σημαντικές εφαρμογές:
  - Enterprise search
  - Peer-2-Peer (P2P) search

Peer-to-Peer Δίκτυα

- Όχι κεντρικός δεικτοδοτητής
- Κάθε κόμβος στο διαδίκτυο κτίζει και διαχειρίζεται το δικό του δείκτη

Παραδείγματα

- Gnutella
- Kazaa
- Bearshare
- Aimster
- Grokster
- Morpheus

Μηχανές Αναζήτησης

- Πρώτη γενιά - χρήση μόνο “on page” δεδομένων κειμένου
  - Συχνότητα λέξεων, γλώσσα
- Δεύτερη γενιά -- χρήση off-page, web-specific δεδομένων
  - Link (ή connectivity) ανάλυση
  - Click-through δεδομένα (σε ποια αποτελέσματα γίνεται click on)
  - Anchor-text (πως οι άνθρωποι αναφέρονται σε δεδομένα)
- Τρίτη γενιά “καταγραφή ανάγκης πίσω από ερώτημα”
  - Σημασιολογική ανάλυση – σε τι αναφέρεται?
  - Εστίαση σε ανάγκες χρηστών και όχι ερωτήματα
  - Προσδιορισμός context
  - Βοήθεια στο χρήστη
  - Ολοκλήρωση ψαξίματος και ανάλυσης κειμένου

Μηχανές Πρώτης Γενιάς

- Επεκταμένο Boolean μοντέλο
  - Ταιριάσματα: exact, prefix, phrase,…
  - Τελεστές: AND, OR, AND NOT, NEAR, …
  - Πεδία: TITLE:, URL:, HOST:,…
  - Συνήθως ο τελεστής AND υλοποιείται πιο εύκολα, και πιθανώς να είναι προτιμητέα ως η εκ των προτέρων επιλογή για μικρά ερωτήματα

- Διάταξη
  - TF παράγοντες: TF, άμεσα keywords, λέξεις σε τίτλους, άμεση έμφαση (headers), κ.λ.π.
  - IDF παράγοντες: IDF, συνολικός αριθμός λέξεων στο corpus, συχνότητα στο query log, συχνότητα στη γλώσσα

Μηχανές Δεύτερης Γενιάς

- Κατάταξη
  -- χρήση off-page, web-specific δεδομένων
  - Link (ή connectivity) ανάλυση
  - Click-through δεδομένα (σε ποια αποτελέσματα οι άνθρωποι εστιάζουν)
  - Anchor-text (πως οι άνθρωποι αναφέρονται σε μία σελίδα)

- Crawling
  - Αλγόριθμοι δημιουργίας του καλύτερου δυνατού corpus

Μηχανές Τρίτης Γενιάς

- Query language determination and different ranking
  - (if query Japanese do not return English)
- Integration of Search and Text Analysis

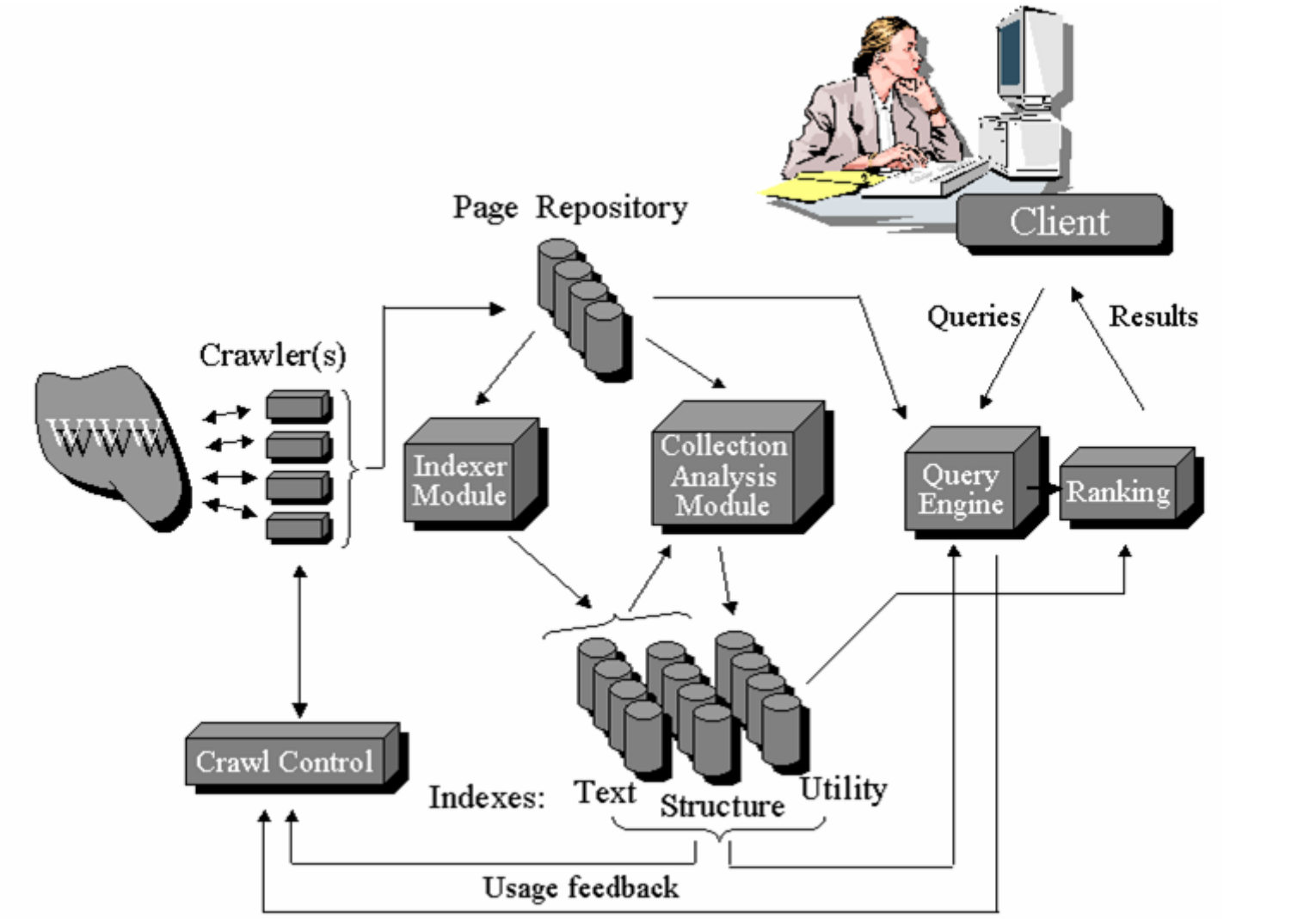
- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (vertical search, family friendly)
  - implicit (use AltaVista from AltaVista France)

- Context use
  - Result restriction
  - Ranking modulation

Βοήθημα στο Χρήστη

- spell checking
- query refinement
- query suggestion
- context transfer

Μηχανές Αναζήτησης



Διαπερνώντας το διαδίκτυο (Crawling)

$$S_0 = \{url_1, url_2, ..., url_k\}$$

- ποιες σελίδες πρέπει να προσπελαστούν ;
- τι γίνεται όταν το περιεχόμενο των σελίδων μεταβάλλεται ; (refresh policy)
- πως ελαχιστοποιείται ο φόρτος ;
- πως η διαδικασία διαπέρασης γίνεται παράλληλα ;

Interest Driven

$$\text{Pages} = \{P_1, P_2, ..., P_n\}$$

$$P_i =< w_1, w_2, ..., w_m >$$

$$P_j =< w'_1, w'_2, ..., w'_m >$$

$$w_k = \begin{cases} 0 & : \quad w_i \notin P_i \\ \#\text{εμφ}_{P_i} * idf & : \quad w_i \in P_i \end{cases}$$

$$idf_{w_i} = \frac{1}{\#\text{εμφ}_{Pages}}$$

Interest Driven & Ομοιότητα Κειμένων

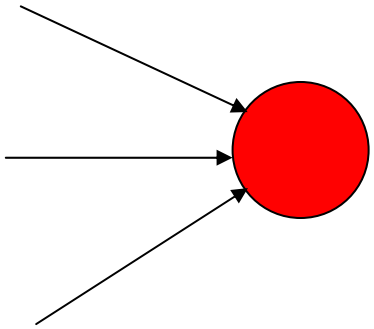
$$\text{Ομοιότητα}(P_i, P_j) = \frac{\vec{P}_i \cdot \vec{P}_j}{|\vec{P}_i| * |\vec{P}_j|} = \cos(\angle(\vec{P}_i, \vec{P}_j))$$

$$I_1(P_i) = \text{Ομοιότητα}(P_i, Q)$$

$$I'_1(P_i) \cong I_1(P_i)$$

"A new approach to topic-specific web resource discovery" Chakrabarti et al. 8th WWW conference 1999

Popularity Driven



$$I_2(P_i) = \#in_{links}(P_i)$$

$$I_2'(P_i) \cong I_2(P_i)$$

Location Driven

$$I_3(P_i) = f(url_{P_i})$$

$$I(P) = \alpha I_1(P) + \beta I_2(P) + \gamma I_3(P)$$

Crawling - Ανανέωση Σελίδων

- f=σταθερή
- f=F(λ_i)

$$\forall i,j : \frac{\lambda_i}{f_i} = \frac{\lambda_j}{f_j}$$

$$Pages_{local} = \{P_1, P_2, ..., P_n\}$$

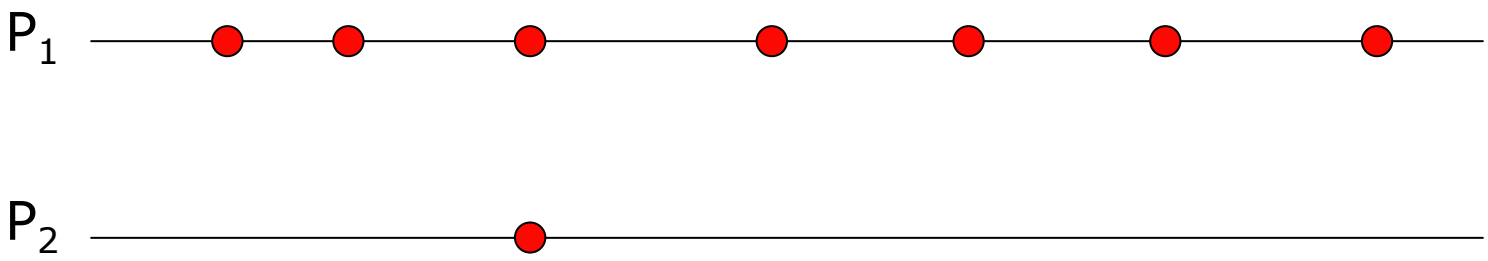
$$F_{P_i}(t) = \begin{cases} 1 & : \quad P_i \quad up\_to\_date \\ 0 & : \quad P_i \quad else \end{cases}$$

$$F(t) = \frac{1}{n} \sum_{i=1}^{n} F_{P_i}(t)$$

Crawling - Ανανέωση Σελίδων

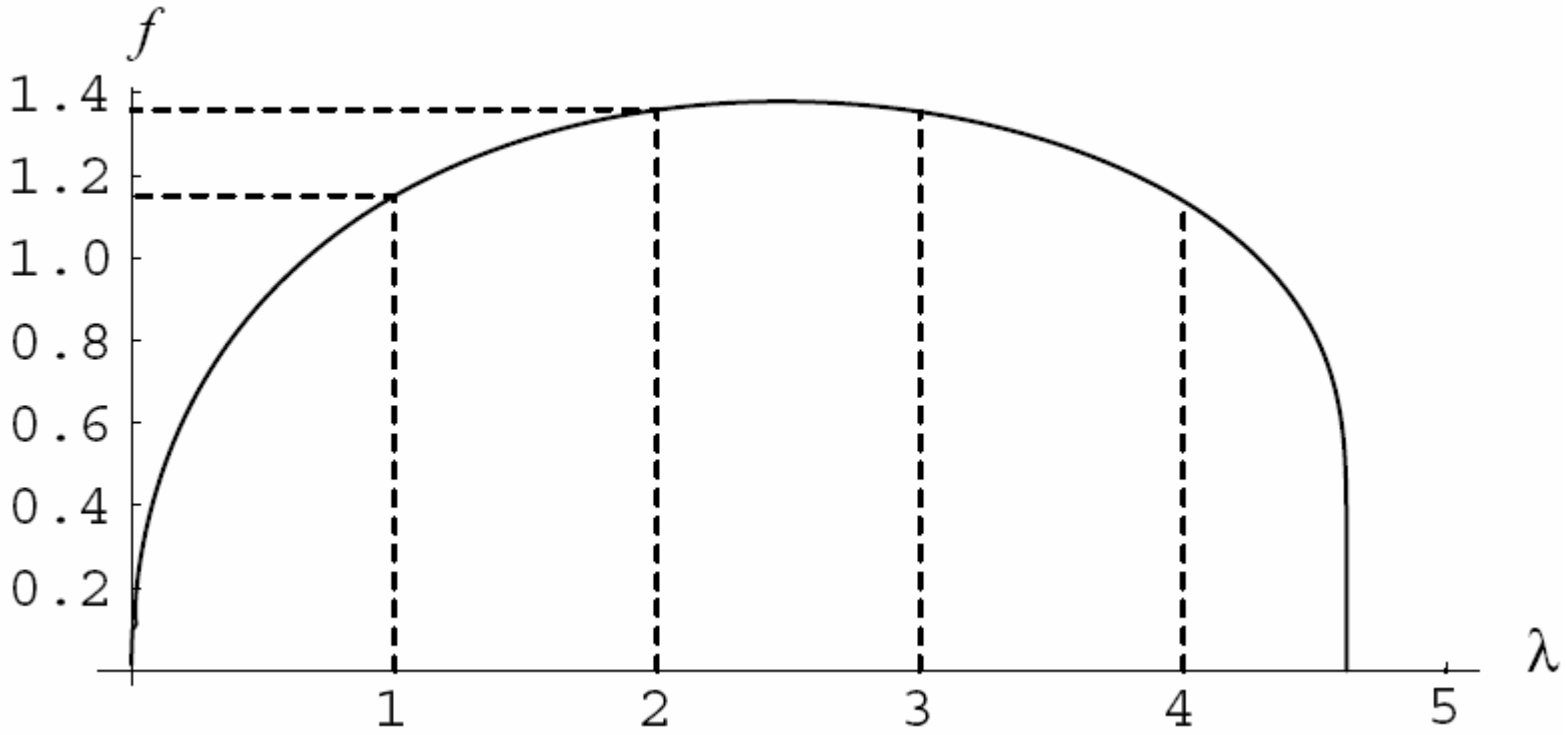
$$A_{P_i}(t) = \begin{cases} 0 & : \quad P_i \quad up\_to\_date \\ t - t' & : \quad P_i \quad else \end{cases}$$

$$A(t) = \frac{1}{n} \sum_{i=1}^{n} A_{P_i}(t)$$

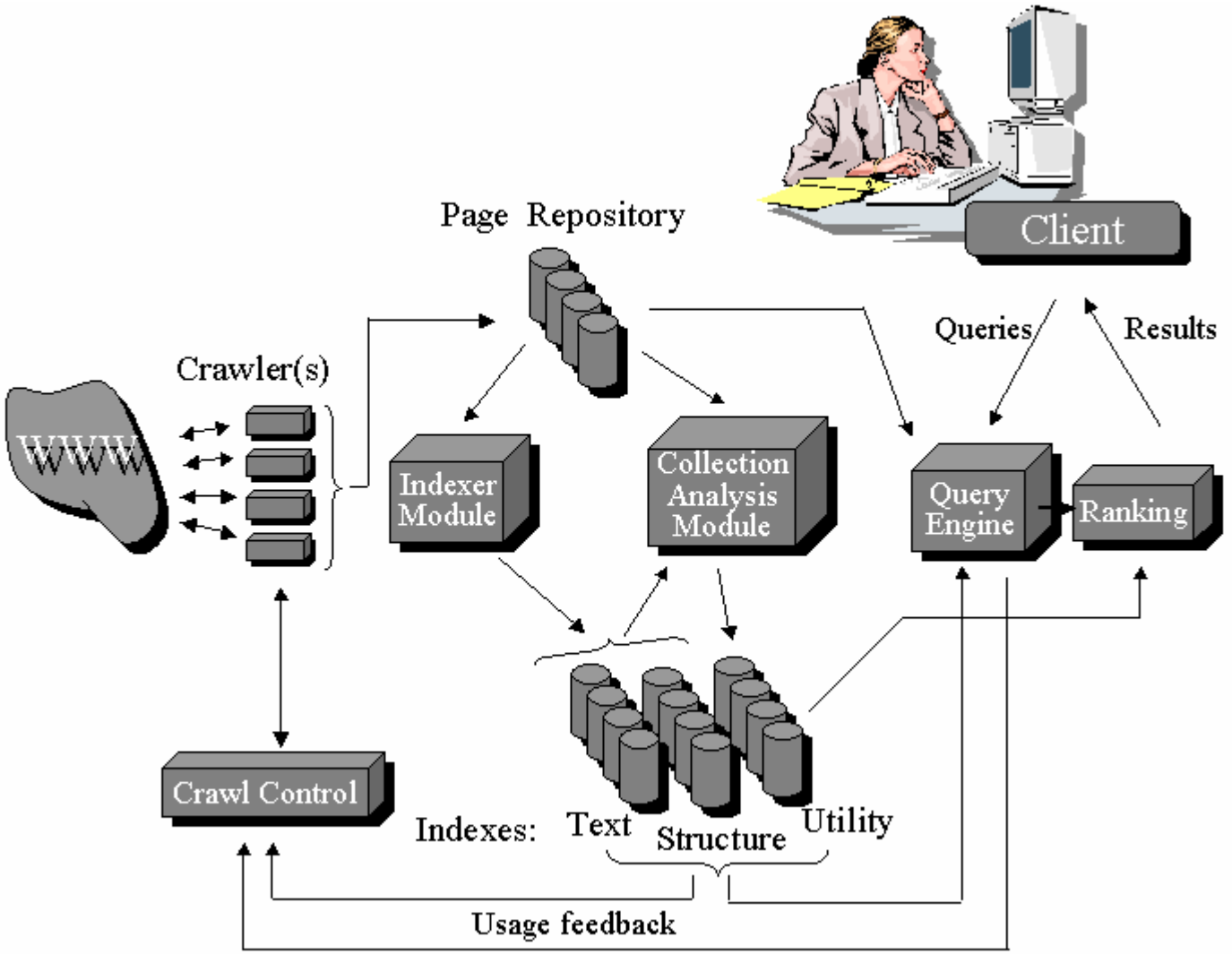


Crawling - Ανανέωση Σελίδων

"Synchronizing a database to improve freshness." Cho, Molina. In *Pro-ceedings of the International Conference on Management of Data*, 2000.

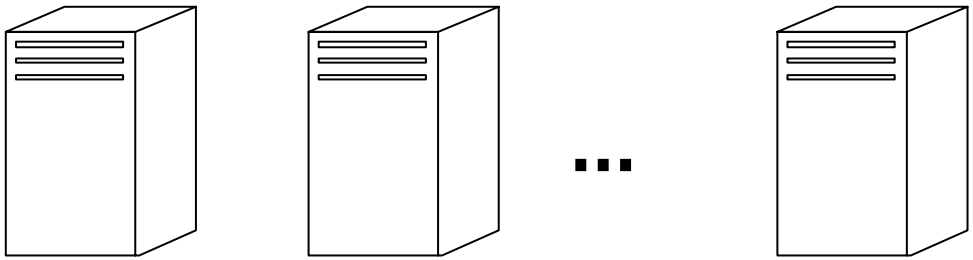


Αποθήκευση – Page Repository



Αποθήκευση – Page Repository

- Κατανεμημένο και αυξομειώσιμο

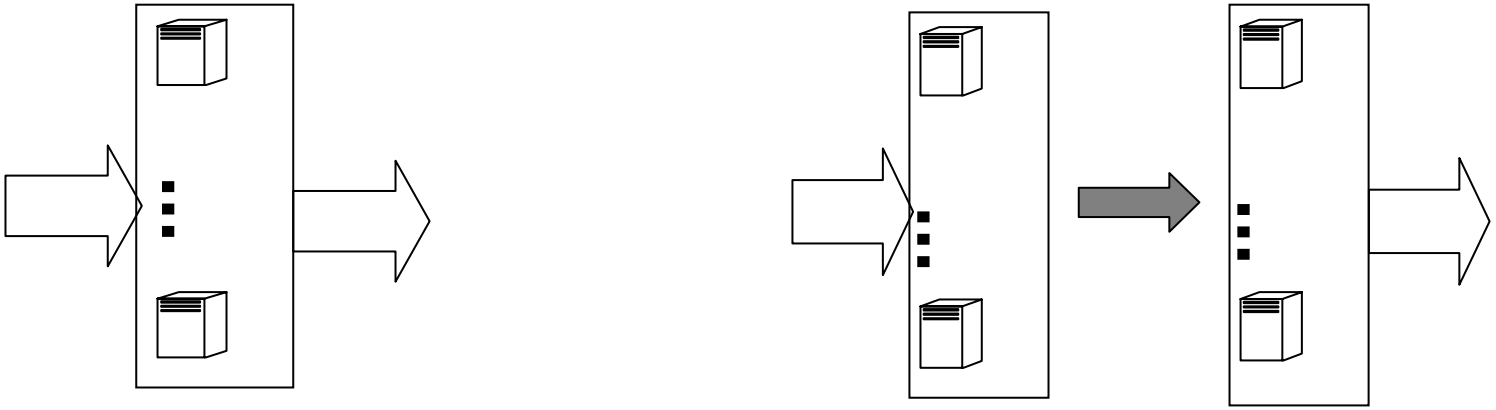


- Φυσική Οργάνωση : αποδοτικό RPA και Streaming Access

	Log	Hash	Hash-Log
Streaming Access	+!	-!	+
RPA	~	+!	~
Page Addition	+!	-!	~

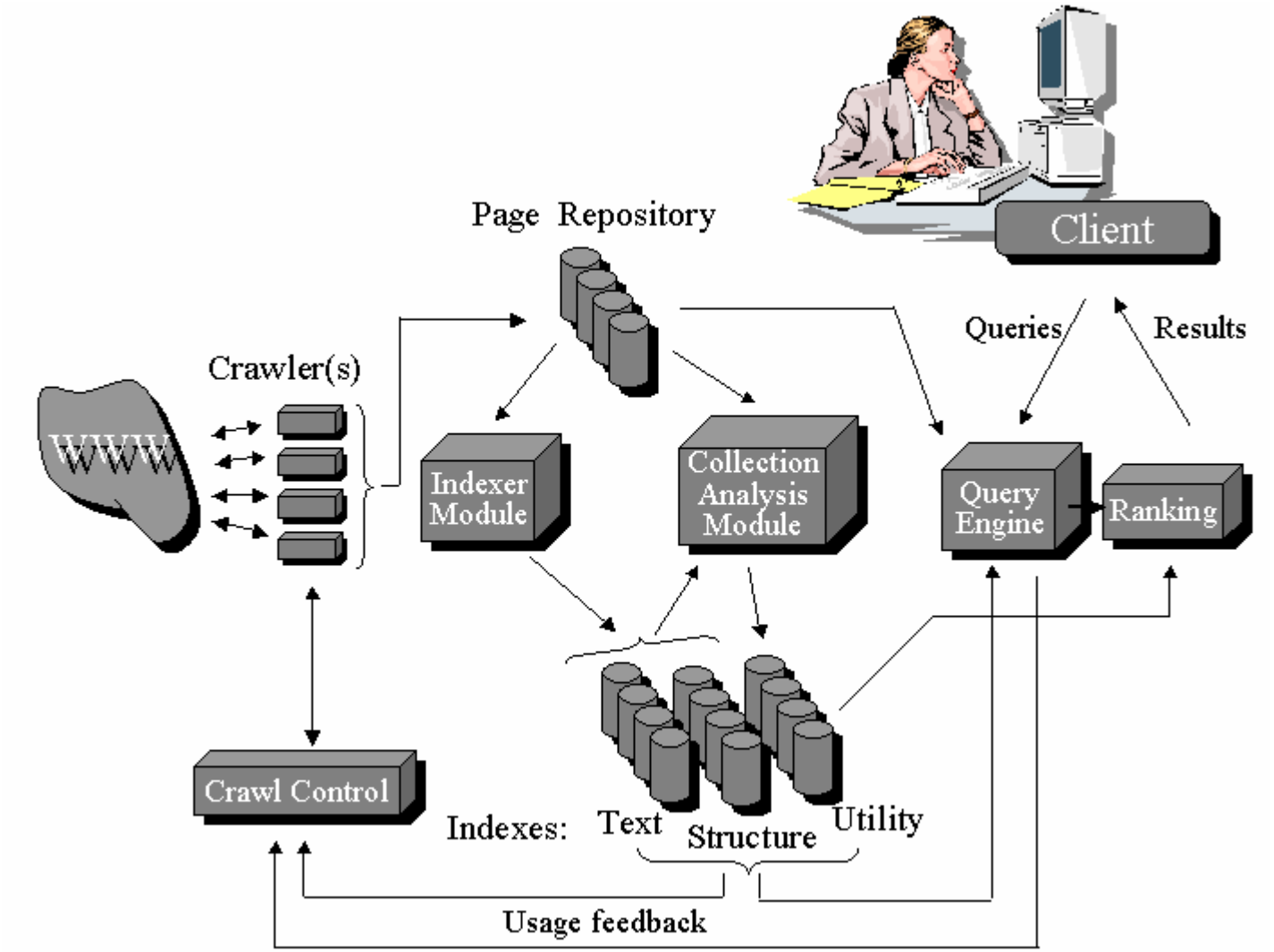
Αποθήκευση – Page Repository

- conflicts vs. freshness



- obsolete pages : μηχανισμός διαγραφής

Δημιουργία Ευρετηρίων – Indexing



Δημιουργία Ευρετηρίων – Indexing

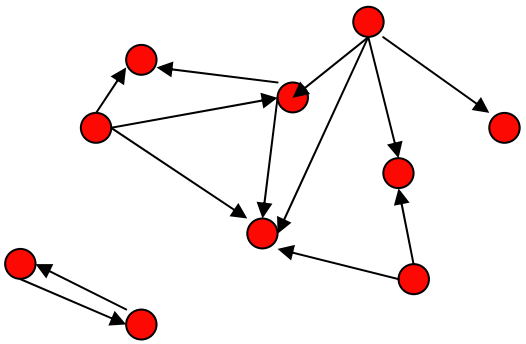
text index

- inverted files
- suffix arrays
- signature files

→

- κατανεμημένο
- συμπιεσμένο

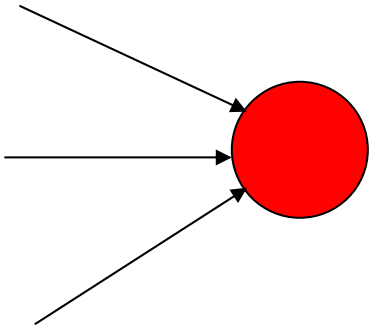
structure (link) index



: link
: site

utility index

Ο τρόπος διασύνδεσης των σελίδων μπορεί να μας δώσει σημαντική επιπλέον πληροφορία !

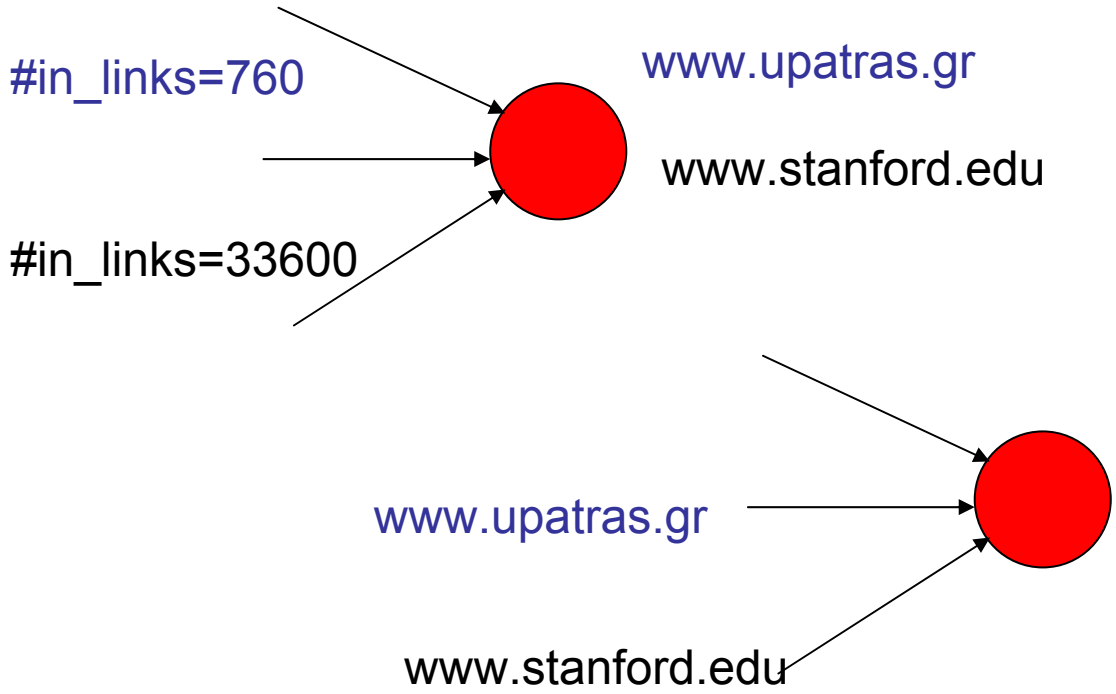


• **PageRank** : "The pagerank citation ranking:Bringing order to the web". Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. Technical report, Computer Science Department, Stanford University,1998. (**Google)**

• **HITS: "**Authoritative sources in a hyperlinked environment". Jon Kleinberg. *Journal of the ACM*, 46(5):604-632, November 1999. (**Clever** – IBM).

PageRank

- Κάθε σελίδα λαμβάνει μία βαθμολογία που εκφράζει την «σημαντικότητα» της.



$$\text{Pages} = \{P_1, P_2, ..., P_n\}$$

$$N(i) = \#\text{Outgoing links}$$

$$B(i) = \{x : x \overset{link}{\rightarrow} i\}$$

strongly connected graph

$$Rank(i) = \sum_{j \in B(i)} \frac{Rank(j)}{N(j)}$$

PageRank

$$\mathsf{A} = \begin{matrix} n \\ \end{matrix} \overset{n}{\begin{pmatrix} & \\ A_{u,v} & \\ & \end{pmatrix}} \qquad A_{u,v} = \begin{cases} \frac{1}{N(u)} & : \quad u \overset{link}{\rightarrow} v \\ 0 & : \quad else \end{cases}$$

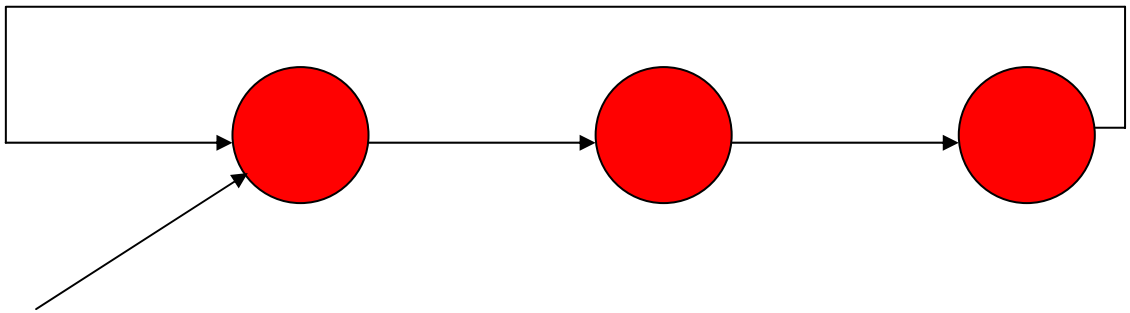
$$\mathsf{R} = \begin{matrix} n \\ \end{matrix} \overset{1}{\begin{pmatrix} \\ \\ \end{pmatrix}}$$

$$\mathsf{R} = \mathsf{AR}$$

$$\lambda R = AR, \lambda = 1$$

- random surfer model

PageRank



$$Rank(i) = d * \sum_{j \in B(i)} \frac{Rank(j)}{N(j)} + \frac{1-d}{n}$$

$$0 \leq d \leq 1$$

- random surfer model

Λεπτομέρειες Υπολογισμού (1)

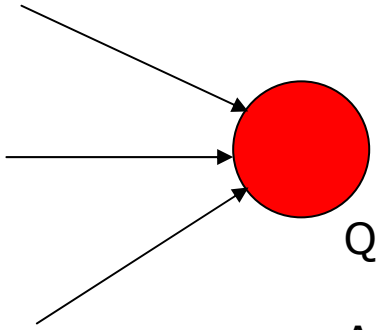
- Μία αλυσίδα Markov αποτελείται από *n* καταστάσεις, και ένα *n*×*n* πιθανοτικό πίνακα μεταβάσεων **P**.
- Σε κάθε βήμα, είμαστε σε μία μόνο από τις καταστάσεις.
- Για $1 \leq i,j \leq n$, το στοιχείο *Pij* μας δίνει τη πιθανότητα το *j* να βρίσκεται στην επόμενη κατάσταση, υποθέτοντας ότι βρισκόμαστε στην κατάσταση i.
- Μία Markov chain είναι εργοδική εάν
  - Υπάρχει μονοπάτι από κάθε κατάσταση σε άλλη
  - Μπορούμε να βρισκόμαστε σε κάθε κατάσταση κάθε στιγμή με μη μηδενική πιθανότητα.

Λεπτομέρειες Υπολογισμού (2)

- Για κάθε εργοδική Markov αλυσίδα, υπάρχει μία *Steady-state distribution*.
- Έστω a = (a1, … an) το row vector με τις steady-state πιθανότητες.
- Εάν η τρέχουσα θέση περιγράφεται με a, τότε η επόμενη περιγράφεται με aP.
- Άρα a=aP, και συνεπώς
  - το a είναι το (αριστερό) ιδιοδιάνυσμα του P.
  - (αντιστοιχεί στο “βασικό” ιδιοδιάνυσμα του P με τη μεγαλύτερη ιδιοτιμή.)

Hypertext Induced Topic Search (HITS)

- Χρησιμοποιεί μηχανισμό αξιολόγησης που εξαρτάται από ένα ερώτημα Q.



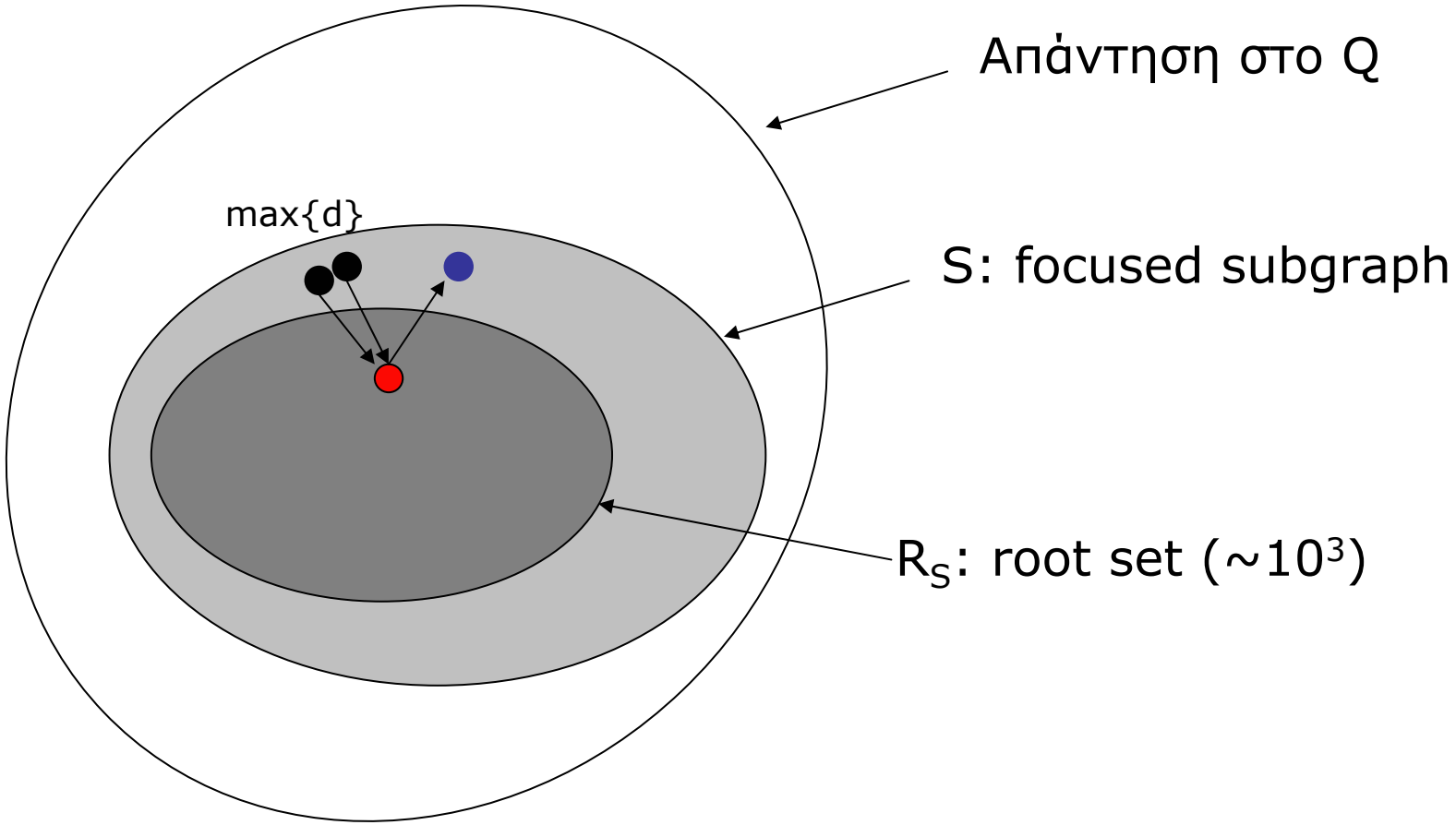
authority
hub

Q="greek university"

Authority : www.upatras.gr
www.auth.gr
students.ceid.upatras.gr/~antonid

Hub: www.gunet.gr
Universities Worldwide
http://geowww.uibk.ac.at/univ/world.html
www.cti.gr

Hypertext Induced Topic Search (HITS)



Hypertext Induced Topic Search (HITS)

$$\mathsf{S} = \{P_1, P_2, ..., P_n\}$$

$$B(i) = \{x : x \overset{link}{\longrightarrow} i\}$$

$$F(i) = \{x : i \overset{link}{\longrightarrow} x\}$$

$$Authority_i = A(i)$$

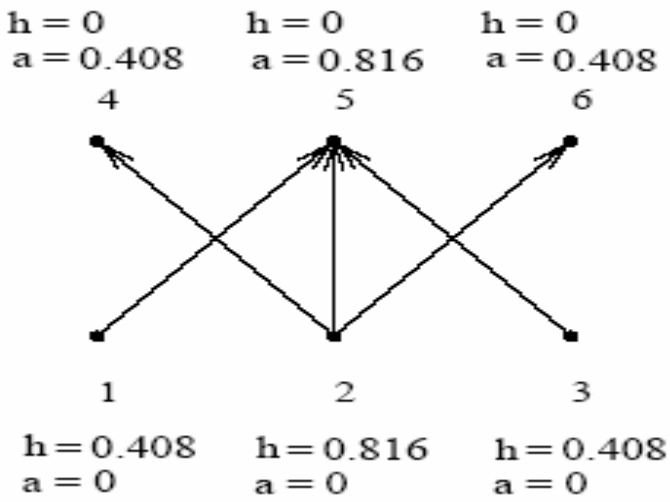
$$Hub_i = H(i)$$

Hypertext Induced Topic Search (HITS)

I-step $\forall i : A(i) = \sum_{j \in B(i)} H(j)$

O-step $\forall i : H(i) = \sum_{j \in F(i)} A(j)$

$\sum_i A(i)^2 = 1, \sum_i H(i)^2 = 1$



Hypertext Induced Topic Search (HITS)

$$\mathsf{A} = \begin{matrix} n \\ \end{matrix}\overset{n}{\begin{pmatrix} & \\ A_{u,v} & \\ & \end{pmatrix}} \qquad A_{u,v} = \begin{cases} 1 & : \quad u \overset{link}{\rightarrow} v \\ 0 & : \quad else \end{cases}$$

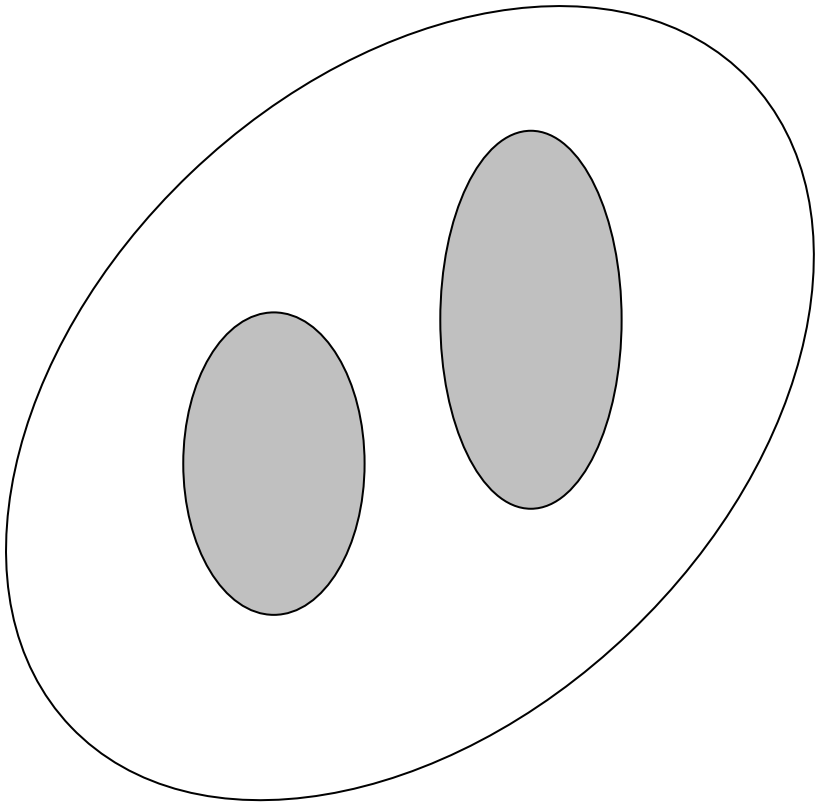
$$\mathsf{a} = \begin{matrix} n \\ \end{matrix}\overset{1}{\begin{pmatrix} \\ \\ \end{pmatrix}} \qquad \mathsf{h} = \begin{matrix} n \\ \end{matrix}\overset{1}{\begin{pmatrix} \\ \\ \end{pmatrix}}$$

$$a = Ah \qquad h = A^T a$$

$$a = AA^T a \qquad h = A^T Ah$$

Hypertext Induced Topic Search (HITS)

Πολλαπλά σύνολα



- jaguar
- randomized algorithms
- abortion

Tag/position heuristics

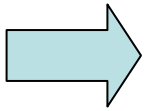
- Αύξησε βάρη όρων
  - σε τίτλους
  - σε tags
  - Κοντά στην αρχή του κειμένου, στα κεφάλαια και sections

Χρήσεις του Anchor Text

- Όταν δεικτοδοτείται μία σελίδα, να δεικτοδοτείται επίσης και το anchor text των υπερδεσμών που δείχνουν σε αυτή.
- Για να μπαίνουν βάρη στον αλγόριθμο hubs/authorities.
- Το Anchor text συνήθως είναι ένα παράθυρο μεγέθους 6-8 λέξεων, γύρω από ένα link anchor.

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$

$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$

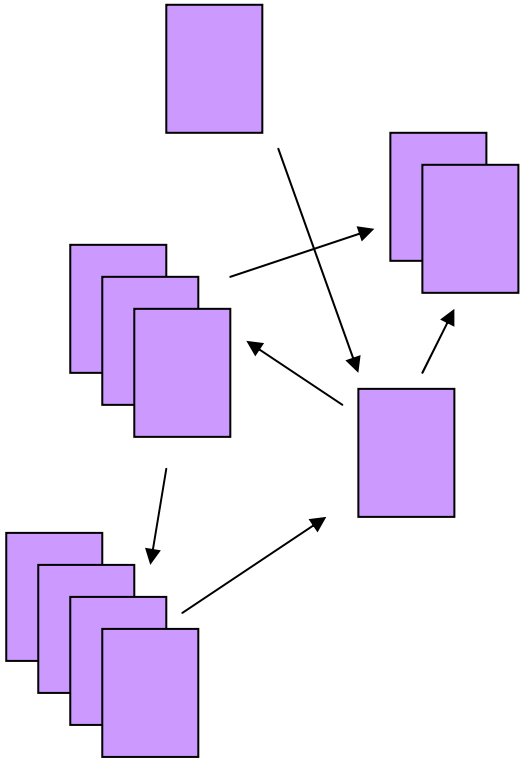


$$h(x) = \sum_{x \mapsto y} w(x,y) \cdot a(y)$$

$$a(x) = \sum_{y \mapsto x} w(x,y) \cdot h(y)$$

Web sites, όχι σελίδες

- Οι σελίδες σε ένα site δίνουν πληροφορίας για παραλλαγές ίδιου θέματος



Αναφορές

→ **Σημειώσεις :** http://mmlab.ceid.upatras.gr/ir

→ "**Searching the WEB**" Arasu, Cho, Molina, Paepcke, Raghavan. ACM Transactions on Internet Technology,2001. http://dbpubs.stanford.edu/pub/2000-37

→ "**The pagerank citation ranking:Bringing order to the web**". Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. Technical report, Computer Science Department, Stanford University,1998.

→ **"Authoritative sources in a hyperlinked environment**". Jon Kleinberg. *Journal of the ACM*, 46(5):604-632, November 1999.