ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Αποθήκες Δεδομένων και Εξόρυξη Γνώσης από Δεδομένα

Web Data Mining: Link Analysis

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Web Data Mining: Link Analysis

Περιεχόμενα

- Εισαγωγή
- Social Network Analysis
- Co-Citation & Bibliographic Coupling
- PageRank
- HITS
- Community Discovery

Εισαγωγή

- Μέχρι το 1996 οι μηχανές αναζήτησης χρησιμοποιούσαν κλασσικούς αλγορίθμους Ανάκτησης Πληροφορίας
  - Πολλοί τεχνηέντως μπορούσαν να «κλέψουν» εισάγοντας διπλότυπες ή παραπλήσιες λέξεις.
- Για να το αντιμετωπίσουμε, βασιζόμαστε όχι μόνον στις λέξεις αλλά στα Links
- Μια σελίδα που τη δείχνουν πολλά links έχει περισσότερες πιθανότητες να περιέχει έγκυρη και δημοφιλή πληροφορία
  - 2 αλγόριθμοι άλλαξαν το τοπίο στις μηχανές αναζήτησης
    - PageRank
    - HITS

Εισαγωγή

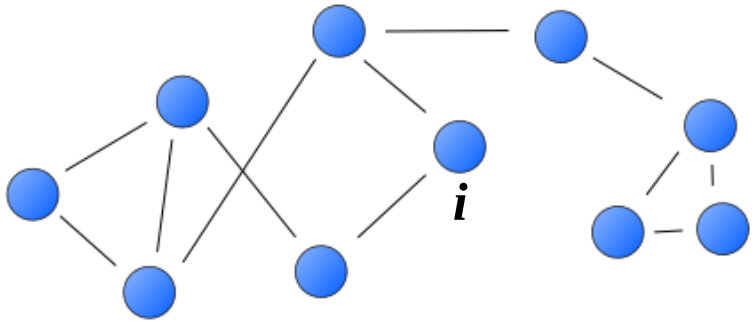
- Και οι 2 αλγόριθμοι βασίζονται στην έννοια του
  - Social Network Analysis
- Αξιοποιούν 2 κριτήρια για τις ιστοσελίδες
  - Centrality
  - Prestige ή Authority

Social Network Analysis

- Η μελέτη των κοινωνικών οντοτήτων (ανθρώπων σε οργανισμούς που καλούνται actors) και των μεταξύ τους αλληλεπιδράσεων.
- Οι σχέσεις μπορούν να περιγραφούν ως γράφος
- Η μελέτη αυτή είναι χρήσιμη για το Web μιας και το ίδιο είναι ένα εικονικό Social Network
  - Σελίδες ως οντότητες και σύνδεσμοι ως σχέση μεταξύ τους
- Μέτρηση σημαντικότητας ενός actor
  - Centrality
  - Prestige

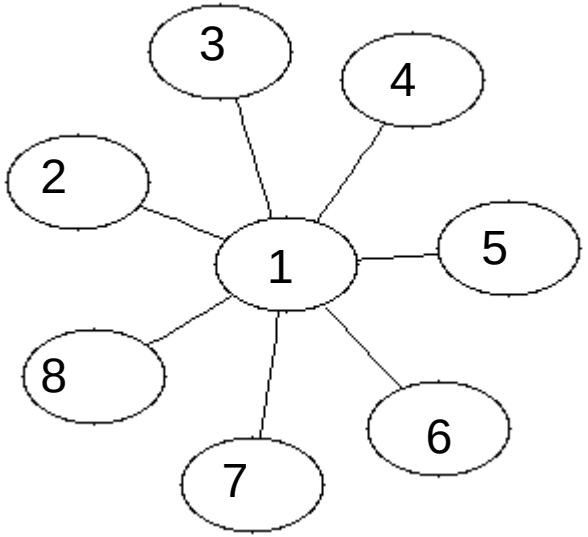
Centrality

- Δείχνει πόσο επικοινωνεί
  ένας κόμβος στο συνολικό δίκτυο
- Degree Centrality
  - $C_D(i)=d(i)/n-1$
    - d(i)=ακμές του κόμβου I
    - n=βαθμός του δικτύου
  - Η βασική ιδέα είναι πως ένας κόμβος είναι κεντρικός αν έχει πολλές συνδέσεις με άλλους
- Closeness Centrality
  - $C_C(i)=(n-1)/\Sigma_n d(i,j)$
    - d(i,j)=το συντομότερο μονοπάτι μεταξύ των κόμβων i και j
  - Η βασική ιδέα είναι πως ένας κόμβος είναι κεντρικός αν αλληλεπιδρά εύκολα με τους υπόλοιπους



Centrality (2)

- Betweenness Centrality
  - Αν 2 μη γειτονικοί κόμβοι j,k θέλουν να αλληλεπιδράσουν και ο κόμβος i είναι στο μονοπάτι τους, τότε ο i παίζει ρόλο στην επικοινωνία τους
  - Δηλ. αν ο i είναι στο μονοπάτι πολλών επικοινωνιών, τότε ο i θεωρείται κεντρικός κόμβος
  - $C_B = \sum_{j<k} \frac{p_{jk(i)}}{p_{jk}}$
  - Αριθμητής=αριθμός συντομότερων μονοπατιών που περνάν από το i
  - Παρονομαστής = αριθμός συντομότερων μονοπατιών όλων των κόμβων που δεν περιλαμβάνουν το i.
  - Μπορεί να υπολογιστεί ακόμη και σε μη-συνδεδεμένο γράφο



C(1)=(n-1)(n-2)/2=7*6/2=21

Prestige

- Διάκριση ανάμεσα σε
  - Outlinks
  - Inlinks
- Το Prestige θεωρεί μόνο τα Inlinks
  - Άρα δουλεύει μόνο για κατευθυνόμενους γράφους
- Centrality→Outlinks
- Prestige→Inlinks

- Degree Prestige
  - $C_D(i)=d_{In}(i)/n-1$
    - $d_{In}(i)$=αριθμός inlinksτου i
    - n=αριθμός actors του δικτύου
- Proximity Prestige
  - $P_P(i) = \frac{|I|/(n-1)}{\sum_{j \in I_i} d(j,i)/|I|}$
    - $|I|$=influence domain, δηλ. πόσοι κόμβοι μπορούν να φτάσουν στον i

Proximity Prestige

vertex	ID	$ID_r = \frac{ID}{9}$	$\overline{d}$	$PP = \frac{ID_r}{\overline{d}}$
1.	9	1.00	2.00	0.500
2.	0	0.00	?	0.000
3.	3	0.33	1.00	0.333
4.	0	0.00	?	0.000
5.	0	0.00	?	0.000
6.	0	0.00	?	0.000
7.	3	0.33	2.00	0.167
8.	2	0.22	1.50	0.148
9.	1	0.11	1.00	0.111
10.	0	0.00	?	0.000

Rank Prestige

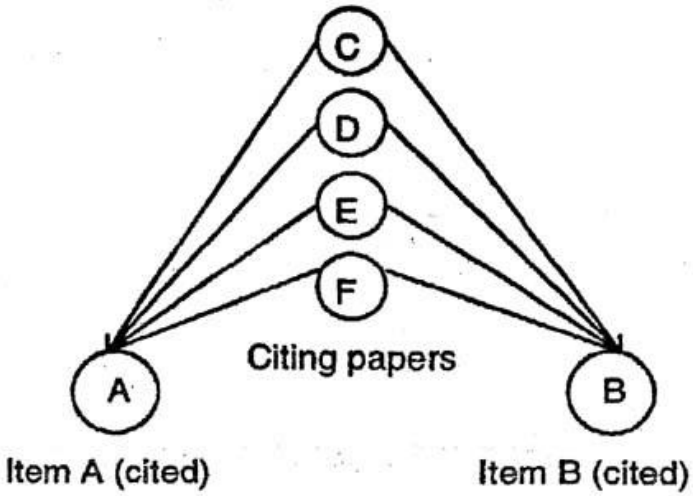
- Τα 2 προηγούμενα μέτρα prestige βασίζονται σε inlinks και αποστάσεις
- Ένα σημαντικό χαρακτηριστικό που δεν λαμβάνουν υπόψη είναι το κύρος των actors
  - Π.χ. σε μια εταιρία, το αν σε προτείνει ο CEO ή ένας απλός υπάλληλος δεν έχει την ίδια βαρύτητα.
- Άρα το rank prestige ενός actor είναι μια συνάρτηση όλων των rank prestiges των υπολοίπων actor.
- $P_R(i)=A_{1i}PR(1)+A_{2i}PR(2)+\ldots+A_{ni}PR(n)$
  - $A_{ni}=1$ αν ο n δείχνει στον i.
  - n εξισώσεις με n αγνώστους➔αναπαράσταση με πίνακες
- $P=A^TP$
  - Η βασική ιδέα για τον PageRank και τον HITS

Co-Citation & Bibliographic Coupling

- Citation=αναφορά ενός επιστημονικού άρθρου σε ένα άλλο
  - Επίσης μπορεί να περιλαμβάνει:
    - Αναφορά ενός συγγραφέα σε άλλον (ους)
- 2 τύποι ανάλυσης
  - Co-Citation
  - Bibliographic

Co-Citation

- Αν 2 άρθρα i,j αναφέρονται από ένα άρθρο k, ίσως να έχουν σχέση ακόμη και αν μεταξύ τους δεν υπάρχει άμεση αναφορά
  - Όσο περισσότερα άρθρα αναφέρουν τα i,j τόσο πιο δυνατή είναι η συσχέτιση τους



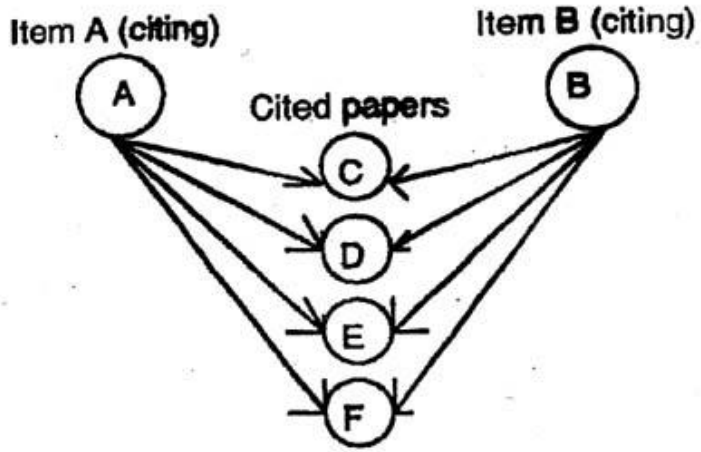
Papers A and B are associated because they are both cited by papers C, D, E and F.

Co-Citation

- L=citation matrix (L_{ij}=1 αν το άρθρο i αναφέρει το j)
- Co-citation (πόσα άρθρα συν-αναφέρουν τα i,j):
  - $C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$
    - n=αριθμός συνολικών άρθρων

Bibliographic Coupling

- Το ανάστροφο πρόβλημα με το co-citation
  - Αν 2 άρθρα i,j αναφέρουν ένα άρθρο k, ίσως να έχουν σχέση ακόμη και αν μεταξύ τους δεν υπάρχει άμεση αναφορά
  - Όσο περισσότερα άρθρα αναφέρονται από τα i,j τόσο πιο δυνατή είναι η συσχέτιση τους



Citing papers A and B are related because they cite papers C, D, E, and F.

Bibliographic Coupling

- Bibliographic Coupling (πόσα άρθρα συν-αναφέρονται από τα i,j):
  - $B_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$
    - n=αριθμός συνολικών άρθρων

PageRank

- Το 1998 ήταν μια σημαντική χρονιά για το Web
  - HITS
    - Kleinberg (ερευνητής στην IBM και στο Cornell)
  - PageRank
    - Brin & Page (ιδρυτές της Google)
  - Αν και έχουν πολλές ομοιότητες ο PageRank έχει υπερισχύσει γιατί
    - Αξιολογεί τις σελίδες ανεξάρτητα του ερωτήματος του χρήστη
    - Αντιμετωπίζει το spamming (το να κλέψει ο ιδιοκτήτης της σελίδας)
  - Ο PageRank αντιμετωπίζει το Web με την αρχή της δημοκρατίας
    - Link(x→y)=Vote(x→y)
    - Ωστόσο λαμβάνει υπόψη το κύρος των σελίδων που ψηφίζουν (όμοιο με το Prestige Rank του Social Analysis)

Εισαγωγικά για τον PageRank

- Inlinks μιας σελίδας i:
  - Οι υπερσύνδεσμοι που δείχνουν στη σελίδα i από άλλες σελίδες
    - Συνήθως οι σύνδεσμοι από το ίδιο site δεν λαμβάνονται υπόψη
- Outlinks μιας σελίδας i:
  - Οι υπερσύνδεσμοι της i που δείχνουν σε άλλες σελίδες
    - Συνήθως οι σύνδεσμοι από το ίδιο site δεν λαμβάνονται υπόψη

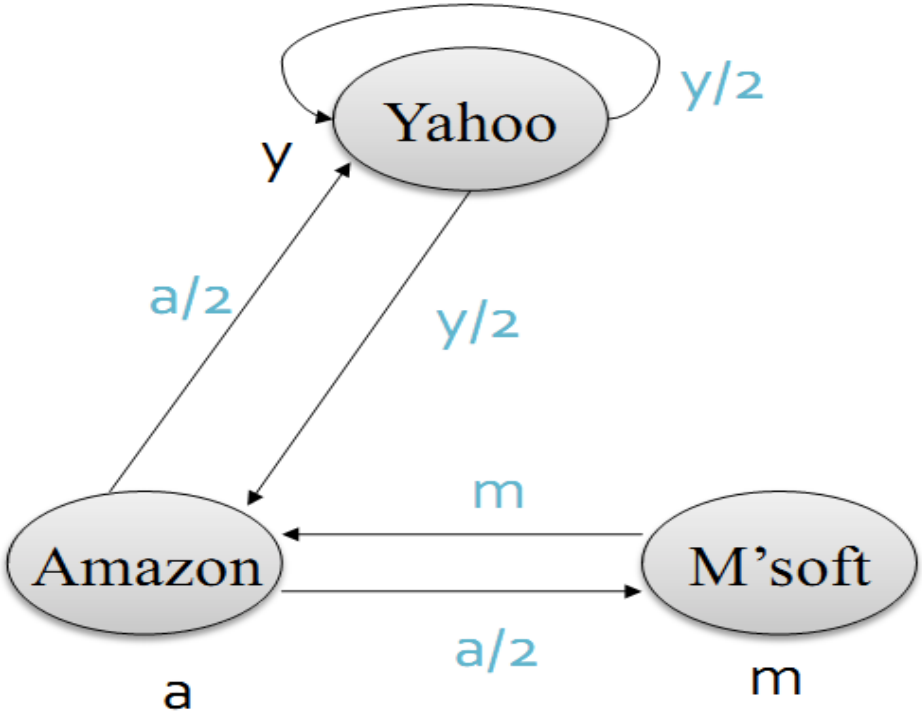
Εισαγωγικά για τον PageRank

- Ένας σύνδεσμος από μια σελίδα σε μια άλλη είναι ένας έμμεσος παράγοντας κύρους προς τη σελίδα που δείχνει.
  - Όσο περισσότερα inlinks έχει η σελίδα τόσο μεγαλύτερο το prestige της.
- Οι σελίδες που δείχνουν στην σελίδα i έχουν το δικό τους prestige
  - Μια σελίδα με υψηλό prestige που δείχνει στην i είναι πιο σημαντική από μια σελίδα με χαμηλότερο prestige που δείχνει στην i
    - Άρα μια σελίδα είναι σημαντική αν την δείχνουν σημαντικές σελίδες

Μαθηματική Μοντελοποίηση

- Το web ως γράφος G=(V,E)
  - V=σελίδες
  - E=σύνδεσμοι
- PageRank
  - Αν η σελίδα P με σημαντικότητα x έχει n outlinks, κάθε link παίρνει x/n ψήφους
  - Η σημαντικότητα μιας σελίδας είναι το άθροισμα των ψήφων στα inlinks που έχει
    - $P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j}$
    - O_j=outlinksτης σελίδας j.
- Με πίνακες
  - $P=A^TP$
  - $A_{ij} = \begin{cases} \frac{1}{O_j}, \alpha\nu\ (i,j) \in E \\ 0, \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\alpha \end{cases}$
    - Δεν μπορεί να λυθεί αποτελεσματικά για το Web→μετάβαση σε αλυσίδες Markov (Markov Chains)
      - Θα το δούμε σε λίγο….

Απλό παράδειγμα



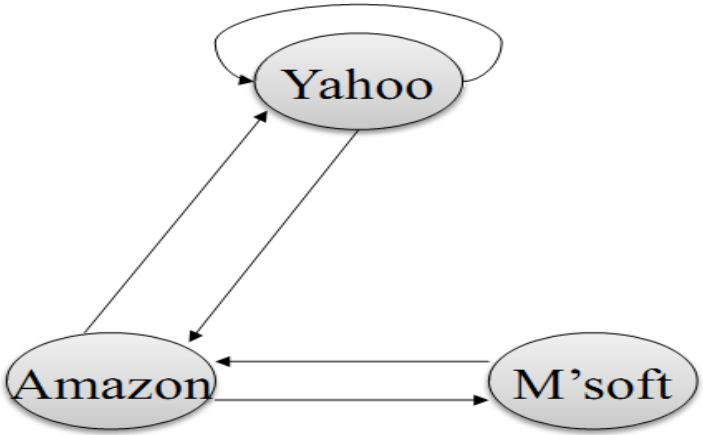
$$y = y/2 + a/2$$
$$a = y/2 + m$$
$$m = a/2$$

Απλό παράδειγμα

- Δεν υπάρχει μοναδική λύση
  - Όμως
    - $y+a+m = 1$
  - Άρα
    - $y = 2/5, a = 2/5, m = 1/5$

Απλό παράδειγμα

- Με πίνακες:



	Y!	A	MS
Y!	½	½	0
A	½	0	1
MS	0	½	0

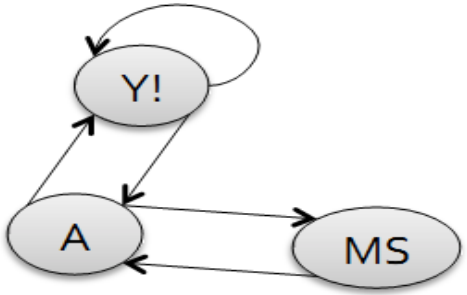
r = Mr

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

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 1 \\ 0 & ½ & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

Λύση για λίγα δεδομένα

- Power iteration:
  - Set $r_i=1/n$
  - $r_i=\sum_j M_{ij}r_j$
  - And iterate



	Y!	A	MS
Y!	½	½	0
A	½	0	1
MS	0	½	0

- Example:

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

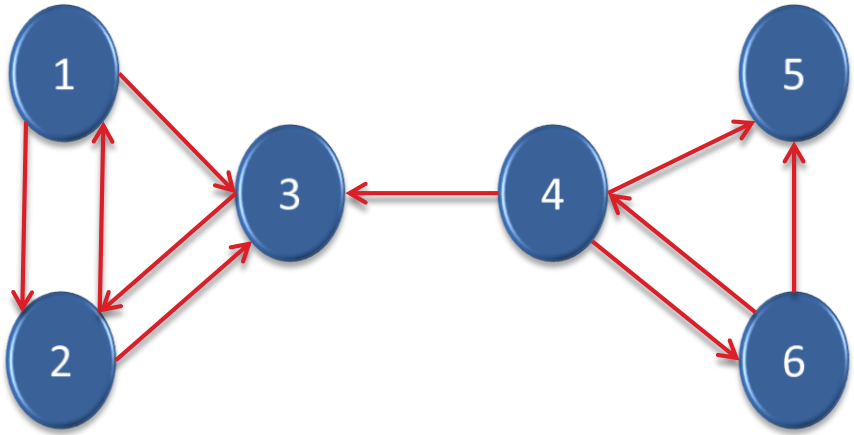
Για το WEB→Markov Chains

- Θεωρία
  - Κάθε σελίδα είναι state
  - Κάθε σύνδεσμος είναι transition (από state σε state με κάποια πιθανότητα)
    - Άρα το web surfing είναι μια στοχαστική διαδικασία
      - Θυμηθείτε ότι O_j είναι τα outlinks άρα $1/O_j$ η πιθανότητα ενός χρήστη να κάνει click σε έναν σύνδεσμο με βάση την uniform κατανομή, χωρίς να πατάει back και χωρίς να πληκτρολογεί URL.

Markov Chains

- Σε μια αλυσίδα Markov το κύριο ερώτημα είναι:
  - Δοσμένης μια αρχικής κατανομής πιθανότητας p0 στην αρχή, ποια είναι η πιθανότητα ότι m βήματα αργότερα η αλυσίδα θα είναι στην κατάσταση j?
    - $p_1=A^T p_0$ για το πρώτο βήμα και…
    - $p_m=A^T p_{m-1}$ μετά από m βήματα (μοιάζει πολύ με την αρχική εξίσωση του PageRank)
      - Από τη θεωρία των MC, μπορούμε να φτάσουμε σε μια στατική κατανομή P ανεξάρτητα από την αρχική επιλογή του p_0.
        - Δηλ. $\lim_{k\to\infty} p_k = P$
    - Επομένως η παραπάνω εξίσωση για m βήματα γράφεται:
      - $P=A^T P$

Παράδειγμα

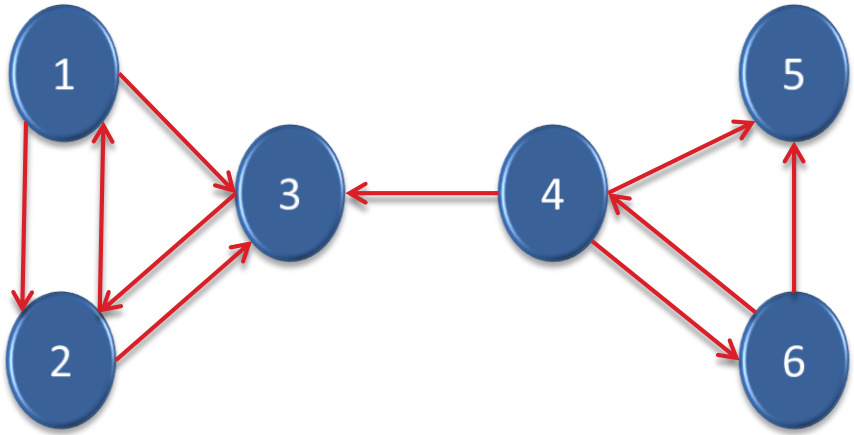


- Για random surfing, ο πίνακας A γίνεται:

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

- Επειδή η $5^{η}$ γραμμή είναι 0 ο A δεν είναι στοχαστικός
  - 2 τρόποι επίλυσης
    - Αφαίρεση του κόμβου
    - Ισοπίθανη κατανομή για τον προβληματικό κόμβο

Παράδειγμα



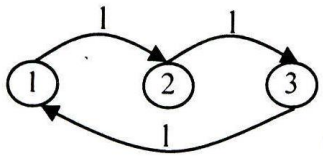
- Συνήθως η 2η λύση ειναι πιο αποδοτική, επομένως ο Α γίνεται

$$\overline{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}.$$

Markov Chains

- Για να ισχύει η στατική κατανομή πρέπει να ικανοποιούνται 2 ιδιότητες για τον A:
  - Irreducible
    - Να είναι strongly connected
      - Δηλ να υπάρχει μονοπάτι για κάθε ζεύγος κόμβων
  - Aperiodic
    - Δειτε το παράδειγμα
      - Όλοι οι κύκλοι έχουν την ίδια περίοδο (k=3)

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



Πως το διορθώνουμε στον PageRank?

- Προσθέτουμε ένα σύνδεσμο από κάθε σελίδα σε κάθε άλλη και δίνουμε στον κάθε σύνδεσμο μια μικρή πιθανότητα μετάβασης, ελεγχόμενη από μια παράμετρο d
- Με το παραπάνω ο A γίνεται:
  - Irreducible
    - Γιατί πλέον είναι strongly connected
  - Aperiodic
    - Γιατί ένας random surfer δεν χρειάζεται πλέον να διασχίζει έναν συγκεκριμένο κύκλο για κάθε κατάσταση
- Άρα ένας random surfer έχει 2 επιλογές
  - Με πιθανότητα d επιλέγει τυχαία ένα outlink για να ακολουθήσει
  - Με πιθανότητα 1-d μεταπηδά σε μια τυχαία σελίδα χωρίς να ακολουθήσει σύνδεσμο

PageRank τελικώς….

- $P(i) = (1-d) + d\sum_{j=1}^{n} A_{ji}P(j)$
- d=dampling factor (από 0 έως 1) συνήθως d=0.85
- Αλγόριθμος:
  - P(0)←1/n
  - K←1
  - Repeat
    - $P(k) = (1-d) + dATPk_{-1}$
    - k←k+1
  - Until $||P_k - P_{k-1}||_1 < \varepsilon$
  - Return P_k

Στην πράξη συγκλίνει πολύ γρήγορα
Για 322M Links θέλει μόλις 52 επαναλήψεις!

Γενικά συμπεράσματα για τον PageRank

- Θετικά
  - Αντιμετωπίζει το spam
    - Είναι πολύ δύσκολο για τον spammer να βρει Inlinks από prestigious σελίδες
  - Δεν εξαρτάται από το query
    - Ο PageRank αξιολογεί μια σελίδα offline και ύστερα κοιτάζει να την ταιριάξει με το ερώτημα του χρήστη
- Αρνητικά
  - Δεν εξαρτάται από το query !!
    - Μια σελίδα μπορεί να έχει κύρος γενικά αλλά όχι για ένα συγκεκριμένο query topic.
    - Δεν λαμβάνει υπόψη το χρόνο

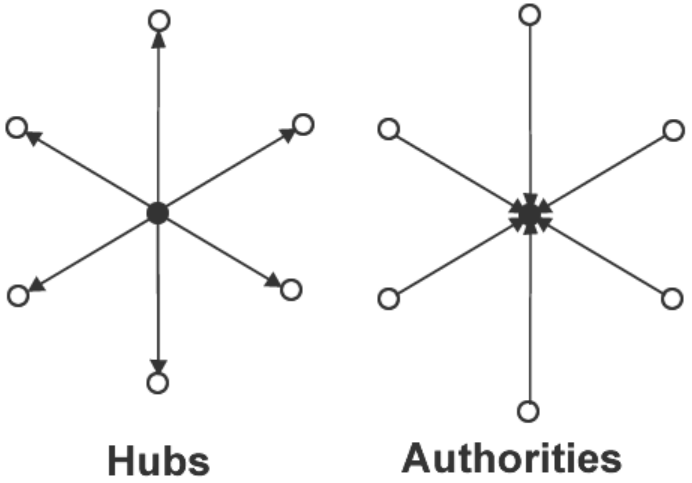
Timed PageRank

- Οι παλαιότερες σελίδες είναι πιο πιθανόν να περιέχουν περισσότερα inlinks από τις νεότερες άρα ευνοούνται
- Παράλληλα οι χρήστες κατά κανόνα προτιμούν νεότερες σελίδες
- Ιδέα
  - Αντί για damping factor d μια συνάρτηση f(t) (0…1)
    - t=η διαφορά μεταξύ τρέχουσας ώρας και update ώρας της σελίδας
  - Μια παλιά σελίδα θα έχει μεγάλο f(t) άρα αν όπου d=f(t) ο χρήστης θα έχει μεγάλη πιθανότητα να μεταπηδήσει σε μια άλλη
  - Μια νέα σελίδα θα έχει μικρό f(t) άρα αν όπου d=1-f(t) ο χρήστης θα έχει μεγάλη πιθανότητα να ακολουθήσει ένα outlink της σελίδας.

HITS

- Hypertext Induced Topic Search
- Είναι εξαρτημένος του query
  - <> του PageRank
- Όταν ένας χρήστης κάνει ένα query, ο HITS επιστρέφει τις σχετικές σελίδες και μετά κάνει 2 αξιολογήσεις (rankings)
  - Authority ranking
  - Hub ranking

HITS



- Authority
  - Μια σελίδα με πολλά inlinks
- Hub
  - Μια σελίδα με πολλά outlinks
    - Είναι οργανωτής, υπό την έννοια πως εάν ένας χρήστης μπει σε ένα hub, μπορεί από εκεί να πάει σε πολλές σελίδες
- Ιδέα
  - Good Hub→Good Authorities &
  - Good Authority←Good Hubs

HITS

- Συλλογή δεδομένων
  - Ο χρήστης υποβάλλει ένα query q
    - Ο HITS συλλέγει top-k σχετικές σελίδες (k=200)
      - Λέγεται root set W
  - Αυξάνει το W βάζοντας κάθε σελίδα (μέχρι 50) που δείχνει μια σελίδα του W και κάθε σελίδα που δείχνεται από μια σελίδα του W
    - Λέγεται base set S

HITS

- Στη συνέχεια κάθε σελίδα του S λαμβάνει
  - Authority score
    - $a(i) = \sum_{(j,i)\in E} h(j)$
  - Hub score
    - $\mathrm{h(i)} = \sum_{(j,i)\in E} a(j)$
  - Η λύση είναι παρόμοια με το PageRank
    - Επανάληψη με έναρξη (a(0)=h(0)=1,1,…,1)