


**Εύρεση & Διαχείριση
Πληροφορίας στον
Παγκόσμιο Ιστό**

Διδάσκων –
Δημήτριος Κατσαρός, Ph.D.

@ Τμ. Μηχανικών Η/Υ, Τηλεπικοινωνιών & Δικτύων
Πανεπιστήμιο Θεσσαλίας

Διάλεξη 14η: 30/05/2007


1



Η μέθοδος HITS

Η μέθοδος SALSA

2



Η μέθοδος HITS

Hypertext Induced Topic Search

3

Hypertext Induced Topic Search (HITS)

- Επινοήθηκε από τον Jon Kleinberg το 1998
- Διαφορές των δυο κύριων αλγορίθμων ranking:
 - Ο PageRank είναι query-independent
 - Ο HITS είναι query-dependent
 - Ο PageRank παράγει ένα μέτρο "σημαντικότητας" που χαρακτηρίζει κάθε ιστοσελίδα
 - Ο HITS παράγει δυο τέτοιους αριθμούς
 - Το authority score & το hub score
- Ο HITS αναλύει τις σελίδες ως authorities και hubs
 - Μια authority είναι μια σελίδα με πολλούς εισερχόμενους υπερσυνδέσμους
 - Ένα hub είναι μια σελίδα με πολλούς εξερχόμενους υπερσυνδέσμους
- Οι καλές authorities δείχνονται από καλά hubs, και τα καλά hubs δείχνουν σε καλές authorities

Τα Μηνεμάν Η.Υ. Τελεμακοσυνών & Δασίων, Πανεπιστήμιο Θεσσαλίας

4

Hypertext Induced Topic Search (HITS)

- Ο HITS ενσωματώθηκε στο έργο της IBM "CLEVER"
- Αποτελέσει τη βάση για τη μηχανή αναζήτησης Teoma (αγοράστηκε από την Ask Jeeves, τώρα Ask.com)
- Συνεπώς, κάθε σελίδα i έχει ένα **authority score** a_i και ένα **hub score** h_i
- Με e_{ij} συμβολίζουμε την ύπαρξη υπερσυνδέσμου από την ιστοσελίδα i στην j
- Υποθέτουμε ότι αρχικά έχουμε αναθέσει σε κάθε ιστοσελίδα ένα authority score x_i και ένα hub score y_i
- Ο HITS υπολογίζει επαναληπτικά τις ποσότητες:

$$x_i^k = \sum_{j: e_{ji} \in E} y_j^{(k-1)} \quad y_i^k = \sum_{j: e_{ij} \in E} x_j^{(k)} \quad k = 0, 1, 2, \dots$$

Τα Μηνεμάν Η.Υ. Τελεμακοσυνών & Δασίων, Πανεπιστήμιο Θεσσαλίας

5

Hypertext Induced Topic Search (HITS)

- Έστω ο πίνακας γειτνίασης L_{ij} με στοιχεία ίσα με 1, εάν υπάρχει υπερσυνδέσμος από την ιστοσελίδα i στην j , και ίσα με 0, στην άλλη περίπτωση
- Οι προηγούμενες επαναληπτικές εξισώσεις μπορούν να γραφούν με τη βοήθεια πινάκων ως εξής:

$$\mathbf{x}^{(k)} = \mathbf{L}^T \mathbf{y}^{(k-1)} \quad \text{και} \quad \mathbf{y}^{(k)} = \mathbf{L} \mathbf{x}^{(k)}$$

Ο αρχικός αλγόριθμος HITS

- Αρχικοποίηση του $\mathbf{y}^{(0)} = \mathbf{e}$ (και άλλες επιλογές αρχικοποίησης είναι πιθανές)
- Μέχρι να επέλθει σύγκλιση:
 - $\mathbf{x}^{(k)} = \mathbf{L}^T \mathbf{y}^{(k-1)}$
 - $\mathbf{y}^{(k)} = \mathbf{L} \mathbf{x}^{(k)}$
 - ++k;
 - Κανονικοποίηση των $\mathbf{x}^{(k)}$ και $\mathbf{y}^{(k)}$

Τα Μηνεμάν Η.Υ. Τελεμακοσυνών & Δασίων, Πανεπιστήμιο Θεσσαλίας

6

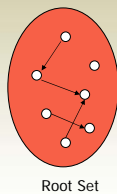
Hypertext Induced Topic Search (HITS)

- Οι προηγούμενες εξισώσεις μπορούν να απλοποιηθούν στις επόμενες:
$$\mathbf{x}^{(k)} = \mathbf{L}^T \mathbf{L} \mathbf{x}^{(k-1)} \quad \text{και} \quad \mathbf{y}^{(k)} = \mathbf{L} \mathbf{L}^T \mathbf{y}^{(k)}$$
- Άρα, ορίζουν την επαναληπτική power μέθοδο για τον υπολογισμό των κυρίαρχων ιδιοδιανυσμάτων των πινάκων $\mathbf{L}^T \mathbf{L}$ και $\mathbf{L} \mathbf{L}^T$
- Είναι παρόμοια περίπτωση με τον υπολογισμό του PageRank, αλλά χρησιμοποιείται διαφορετικός πίνακας συντελεστών
- Ο πίνακας $\mathbf{L}^T \mathbf{L}$ λέγεται **πίνακας authority**, αφού καθορίζει τα authority scores
- Ο πίνακας $\mathbf{L} \mathbf{L}^T$ λέγεται **πίνακας hub**, αφού καθορίζει τα hub scores
- Και οι δυο πίνακες είναι συμμετρικοί, θετικοί και semidefinite

Τη. Μιχαηλίδης Η.Υ. Τηλεπικοινωνιών & Δικτύων, Πανεπιστήμιο Θεσσαλίας

7

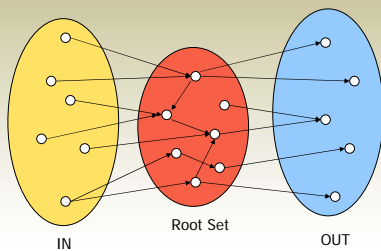
Υλοποίηση του HITS (1/5)



Τη. Μιχαηλίδης Η.Υ. Τηλεπικοινωνιών & Δικτύων, Πανεπιστήμιο Θεσσαλίας

8

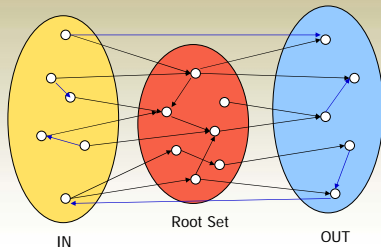
Υλοποίηση του HITS (2/5)



Τη. Μιχαηλίδης Η.Υ. Τηλεπικοινωνιών & Δικτύων, Πανεπιστήμιο Θεσσαλίας

9

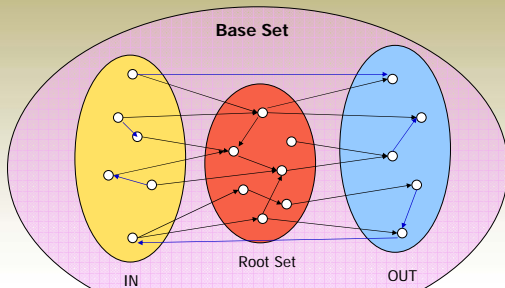
Υλοποίηση του HITS (3/5)



Τα Μηνεγιάν Η.Υ. Τελεμακοινωνιών & Διεθνών, Πανεπιστήμιο Θεσσαλίας

10

Υλοποίηση του HITS (4/5)



Τα Μηνεγιάν Η.Υ. Τελεμακοινωνιών & Διεθνών, Πανεπιστήμιο Θεσσαλίας

11

Υλοποίηση του HITS (5/5)

- Το Root Set βρίσκεται με τη βοήθεια μιας μηχανής αναζήτησης με keyword-search
- Το Base Set είναι ο Neighborhood Graph
- Ενσωματώνονται και τεχνικές σημασιολογικής συγγένειας στην επιλογή των σελίδων που θ' απαρτίζουν το Base Set.
- Για να μην μεγαλώσει σε τεράστιο μέγεθος το Base Set, ορίζουμε έναν μέγιστο αριθμό κόμβων (με εισερχόμενους/εξερχόμενους) υπερσυνδέσμους για τον κάθε κόμβο του Root Set τους οποίους ενσωματώνουμε στο Base Set
- Δεν χρειάζεται να υπολογίσουμε το κυρίαρχο ιδιοδιάνυσμα και για τον πίνακα $L^T L$ αλλά και για τον LL^T , αλλά μόνο για τον ένα πίνακα, αφού ισχύει: $y=Lx$

Τα Μηνεγιάν Η.Υ. Τελεμακοινωνιών & Διεθνών, Πανεπιστήμιο Θεσσαλίας

12

Σύγκλιση του HITS (4/6)

- Έχει δυο διακριτές ιδιοτιμές, τις 2 και 0, οι οποίες έχουν πολλαπλότητα 2
- Εάν ξεκινήσουμε με το $\mathbf{x}^{(0)} = \frac{1}{4} \mathbf{e}^T$ καταλήγουμε στο διάνυσμα authority $\mathbf{x}^{(\infty)} = (1/3 \ 1/3 \ 1/3 \ 0)^T$
- Εάν ξεκινήσουμε με το $\mathbf{x}^{(0)} = (1/4 \ 1/8 \ 1/8 \ 1/2)^T$ καταλήγουμε στο διάνυσμα authority $\mathbf{x}^{(\infty)} = (1/2 \ 1/4 \ 1/4 \ 0)^T$
- Αιτία του προβλήματος μοναδικότητας είναι η reducibility
- Θα λέμε ότι ένας πίνακας \mathbf{B} είναι reducible εάν υπάρχει πίνακας μετάθεσης \mathbf{Q} τέτοιος ώστε (οι \mathbf{X} και \mathbf{Z} είναι τετραγωνικοί):

$$\mathbf{Q}^T \mathbf{B} \mathbf{Q} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$$

Σύγκλιση του HITS (5/6)

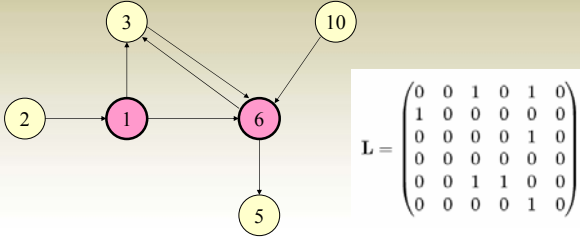
- Η reducibility ενός πίνακα σημαίνει ότι υπάρχουν καταστάσεις "καταβόθρες", ενώ η irreducibility σημαίνει ότι από οποιαδήποτε κατάσταση μπορώ να μεταβώ σε οποιαδήποτε άλλη κατάσταση
- Το θεώρημα Perron-Frobenius εγγυάται ότι ένας irreducible, μη-αρνητικός πίνακας έχει ένα μοναδικό, κανονικοποιημένο θετικό κυρίαρχο ιδιοδιάνυσμα, το λεγόμενο διάνυσμα Perron
- Συνεπώς, η reducibility του $\mathbf{L}^T \mathbf{L}$ είναι υπεύθυνη για η σύγκλιση σε περισσότερα του ενός διανύσματα-λύσεις
- Το ίδιο πρόβλημα αντιμετώπισε και ο πίνακας \mathbf{S} στο PageRank, αλλά με μια μετατροπή του πίνακα σε irreducible επιλύθηκε
- Το ίδιο μπορεί να γίνει και με τον HITS

Σύγκλιση του HITS (6/6)

- Αντί για τον αρχικό πίνακα authority, χρησιμοποιούμε τον πίνακα $\xi \mathbf{L}^T \mathbf{L} + (1-\xi)/n \mathbf{e} \mathbf{e}^T$
- Όμοια για τον πίνακα hub
- Αυτός ο τροποποιημένος HITS, λέγεται Exponential HITS
- Τέλος, ανεξάρτητα από το εάν η κυρίαρχη ιδιοτιμή του πίνακα επανάληψης \mathbf{B} είναι απλή ή πολλαπλή, η σύγκλιση σε ένα μη-μηδενικό διάνυσμα εξαρτάται από εάν το αρχικό διάνυσμα $\mathbf{x}^{(0)}$ δεν βρίσκεται στην εμβέλεια του $(\mathbf{B} - \lambda_1 \mathbf{I})$
- Εάν το $\mathbf{x}^{(0)}$ παράγεται τυχαία, τότε με (σχεδόν) βεβαιότητα δεν θα υφίσταται το πρόβλημα

Παράδειγμα εφαρμογής του HITS (1/4)

- Εστω ότι, σε απάντηση ενός ερωτήματος σε μια παραδοσιακή keyword-based μηχανή αναζήτησης, επιστρέφονται οι ιστοσελίδες που αντιστοιχούν στους κόμβους 1 και 6



$$L = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Παράδειγμα εφαρμογής του HITS (2/4)

$$L^T L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$L L^T = \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

- Τα κανονικοποιημένα κύρια ιδιοδιανύσματα με τα authority και hub scores είναι:

$$\mathbf{x}^T = (0 \quad 0 \quad 0.3660 \quad 0.1340 \quad 0.5 \quad 0)$$

$$\mathbf{y}^T = (0.3660 \quad 0 \quad 0.2113 \quad 0 \quad 0.2113 \quad 0.2113)$$

Παράδειγμα εφαρμογής του HITS (3/4)

- Μπορεί να συμβούν δυο τύπων ισοπαλίες
 - Στις τιμές 0
 - Μπορεί να αποφευχθούν με την primitivity τροποποίηση
 - Στις θετικές τιμές
 - αυτές είναι σπάνιες σε μεγάλους θετικά-ορισμένους πίνακες
 - μπορούν να επιλυθούν με FCFS
- Authority ranking = (6 3 5 1 2 10)
- Hub ranking = (1 3 6 10 2 5)
- Για λόγους σύγκρισης, υπολογίζουμε ξανά τα διανύσματα authority και hub, αλλά χρησιμοποιώντας τον irreducible πίνακα $\xi L^T L + (1-\xi)/n \mathbf{e}\mathbf{e}^T$ ως πίνακα authority και τον irreducible πίνακα $\xi L L^T + (1-\xi)/n \mathbf{e}\mathbf{e}^T$ ως πίνακα hub

Παράδειγμα εφαρμογής του HITS (4/4)

- Για $\xi=0.95$, έχουμε

$$\mathbf{x}^T = (0.0032 \quad 0.0023 \quad 0.3634 \quad 0.1351 \quad 0.4936 \quad 0.0023)$$
$$\mathbf{y}^T = (0.3628 \quad 0.0032 \quad 0.2106 \quad 0.0023 \quad 0.2106 \quad 0.2106)$$

- Στο παράδειγμα αυτό, η μετατροπή τους σε irreducible πίνακες δεν άλλαξε το ranking, ούτε το authority ούτε το hub ranking
- Όμως, απέφυγε τις ισοπαλίες στο 0

Πλεονεκτήματα/Μειονεκτήματα του HITS

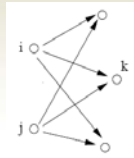
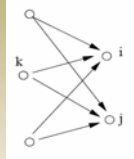
- **Παρουσιάζει δυο λίστες στο χρήστη**
 - Authoritative σελίδες, για εις βάθος αναζήτηση σε κάποιο αντικείμενο
 - Hub σελίδες, δηλ., portal σελίδες, για αναζήτηση σε εύρος
- **Λύνει ένα μικρό πρόβλημα, σε μέγεθος πινάκων**
- **Είναι query-dependent**
 - Σε run-time, δηλ., για κάθε ερώτημα του χρήστη, χτίσιμο του Base Set, και εύρεση ενός ιδιοδιανύσματος
- **Μπορεί να γίνει query-independent**, δηλ., να εκτελεστεί πάνω σε όλο το γράφημα του Web
- **Είναι πολύ επιρρεπής σε spamming**
 - Με προσθήκη συνδέσμων από/προς την ιστοσελίδα μας
 - Φυσικά, είναι ευκολότερο να αυξήσουμε το hub score της ιστοσελίδας μας, αλλά εξ' αιτίας της αλληλεξάρτησής τους μπορεί να αυξηθεί και το authority score ως αποτέλεσμα της αύξησης του hub score

Πλεονεκτήματα/Μειονεκτήματα του HITS

- **Παρουσιάζει το φαινόμενο topic drift**
 - Κατά το χτίσιμο του Base Set, είναι πιθανό ότι μια πολύ authoritative σελίδα, αλλά εκτός αντικειμένου της αναζήτησης, να εμφανιστεί στο Base Set επειδή έχει σύνδεσμο από/προς κάποια σελίδα του Root Set
 - Αυτή η σελίδα μπορεί να έχει τόσο βάρος, ώστε αυτή και όλες οι “γειτονικές” της να εμφανίζονται στην κορυφή της λίστας των αποτελεσμάτων, με συνέπεια η λίστα των αποτελεσμάτων να κυριαρχηθεί από σελίδες εκτός του ζητούμενου αντικειμένου
 - Το πρόβλημα μπορεί να αντιμετωπιστεί εάν “ζυγίσουμε” τα hub και authority scores των σελίδων με βάση τη σχετικότητα της κάθε σελίδας ως προς το αντικείμενο της αναζήτησης, δηλ., αντί για τον δυαδικό πίνακα L , να έχουμε κάτι ανάλογο με τον πίνακα του intelligent surfer

Σχέση του HITS με τη Βιβλιομετρία (1/2)

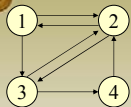
- **Co-citation:** Δυο σελίδες δέχονται υπερσύνδεσμο από την ίδια σελίδα
- Ο πίνακας authority $L^T L$ σχετίζεται με την έννοια του co-citation
- Ranking με βάση το inlink παρέχει αξιοπρεπή προσέγγιση του HITS authority
- **Co-reference:** Δυο σελίδες έχουν υπερσύνδεσμο προς την ίδια σελίδα
- Ο πίνακας hub LL^T σχετίζεται με την έννοια του co-reference
- Ranking με βάση το outlink παρέχει αξιοπρεπή προσέγγιση του HITS hub



Τη. Μιχαηλίδης Η.Υ. Τεχνολογιών & Διεθνών. Πανεπιστήμιο Θεσσαλίας

25

Σχέση του HITS με τη Βιβλιομετρία (2/2)



$$L = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$L^T L = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 3 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} = D_{in} + C_{cit}$$

$$LL^T = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} = D_{out} + C_{ref}$$

- D_{in} : διαγώνιος πίνακας με το indegree των κόμβων και C_{cit} ο πίνακας co-citation
 - Το στοιχείο (3,3) δείχνει ότι ο κόμβος 3 έχει indegree 2
 - Το στοιχείο (4,3) δείχνει ότι οι κόμβοι 4 και 3 δεν έχουν κοινό inlink
- D_{out} : διαγώνιος πίνακας με το outdegree των κόμβων και C_{ref} ο πίνακας co-reference
 - Το στοιχείο (1,2) δείχνει ότι οι κόμβοι 1 και 2 έχουν 1 κοινό σύνδεσμο (προς τον κόμβο 3)
 - Το στοιχείο (4,2) δείχνει ότι οι κόμβοι 4 και 2 δεν έχουν κοινό σύνδεσμο προς κάποιο κόμβο

Τη. Μιχαηλίδης Η.Υ. Τεχνολογιών & Διεθνών. Πανεπιστήμιο Θεσσαλίας

26


Query-independent HITS (1/4)

Ο query-independent αλγόριθμος HITS

- Αρχικοποίηση του $x^{(0)} = e/n$ (και άλλες επιλογές αρχικοποίησης είναι πιθανές)
- Μέχρι να επέλθει σύγκλιση:
 - $x^{(k)} = \xi L^T L x^{(k-1)} + (1-\xi)/n e$
 - $x^{(k)} = x^{(k)} / \|x^{(k)}\|_1$
 - $y^{(k)} = \xi L L^T y^{(k-1)} + (1-\xi)/n e$
 - $y^{(k)} = y^{(k)} / \|y^{(k)}\|_1$
 - ++k;
 - Θέτουμε $x = x^{(k)}$ και $y = y^{(k)}$
- Συγκλίνει στα μοναδικά θετικά διανύσματα hub και authority, ανεξάρτητα από τη reducibility του πίνακα
- Ο L είναι ο πίνακας γειτνίασης όλου του Web γραφήματος

Τη. Μιχαηλίδης Η.Υ. Τεχνολογιών & Διεθνών. Πανεπιστήμιο Θεσσαλίας


27



Επιτάχυνση του HITS

- Ο Kleinberg χρησιμοποίησε την power μέθοδο για τον υπολογισμό των κυρίαρχων δεξιών ιδιοδιανυσμάτων των πινάκων $L^T L$ και $L L^T$
- Οι πίνακες του HITS είναι πολύ μικροί σε σχέση με τους αντίστοιχους του PageRank, και κατά πάσα πιθανότητα, χρησιμοποιούνται τεχνικές που είναι memory-intensive για τον υπολογισμό των δυο πινάκων του HITS, π.χ., Lanczos
- Για τον query-dependent HITS δεν υπάρχει έρευνα σχετική με μεθόδους επιτάχυνσης
- Για τον query-independent HITS μπορούν να χρησιμοποιηθούν οι ίδιες τεχνικές που έχουμε συζητήσει για τον PageRank

31




Ευαισθησία του HITS

• **ΘΕΩΡΗΜΑ.** Έστω E ο πίνακας διαταραχής, ώστε $L^T L = L^T L + E$. Όταν η λ_1 είναι απλή, τότε

$$\sin \angle(\mathbf{x}, \mathbf{x}') \leq \|E\|_2 / (\lambda_1 - \lambda_2)$$

• Συνεπώς, εάν το χάσμα ιδιοδιανυσμάτων είναι μεγάλο, τότε ο πίνακας authority δεν είναι ευαίσθητος σε μικρές αλλαγές στο Web γράφημα

32



Η μέθοδος SALSA

Stochastic Approach for Link Structure Analysis

33

Ομοιότητες SALSA με HITS και PageRank

- Επινοήθηκε από τους Ronny Lempel και Shlomo Moran το 2000
- Συνδυασμός **HITS** και **PageRank**
- Ο SALSA χρησιμοποιεί **authority** και **hub score**
- Ο SALSA δημιουργεί ένα **neighborhood graph** χρησιμοποιώντας authority και hub ιστοσελίδες και υπερσυνδέσμους

34

Διαφορές SALSA με HITS και PageRank

- Η μέθοδος SALSA δημιουργεί ένα διμερές γράφημα (**bipartite graph**) των σελίδων authority και hub στο neighborhood γράφημα
- Το ένα σύνολο περιέχει τις hub σελίδες
- Το άλλο σύνολο περιέχει τις authority σελίδες
- Μια σελίδα μπορεί να περιέχεται και στα δυο σύνολα

35

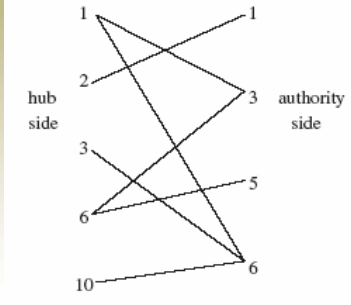
Neighborhood Graph N

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

36

Διμερές γράφημα G του Neighborhood Graph N

$V_h = \{1, 2, 3, 6, 10\}$,
 $V_a = \{1, 3, 5, 6\}$.



hub side authority side

Τη. Μιχαηλίδης Η.Υ. Τελεμαχωνιδών & Δασκίων, Πανεπιστήμιο Θεσσαλίας

37

Markov αλυσίδες

- Από το διμερές γράφημα **G** σχηματίζονται δυο πίνακες
 - Μια hub Markov chain με πίνακα **H**
 - Μια authority Markov chain με πίνακα **A**
- Οι πίνακες **H** και **A** μπορεί να παραχθούν από τον πίνακα γειτνίασης (adjacency matrix) **L** που έχουμε δει στον υπολογισμό του HITS και του PageRank
- Ο HITS χρησιμοποιεί τον unweighted matrix **L**
- Ο PageRank χρησιμοποιεί τη row-weighted έκδοση του πίνακα **L**
- SALSA χρησιμοποιεί row και column weighting

Τη. Μιχαηλίδης Η.Υ. Τελεμαχωνιδών & Δασκίων, Πανεπιστήμιο Θεσσαλίας

38

Πώς υπολογίζονται οι **H** και **A**;

- Έστω ότι L_r είναι ο **L** με κάθε μη-μηδενική γραμμή του να διαιρείται με το άθροισμά της
- Έστω ότι L_c είναι ο **L** με κάθε μη-μηδενική στήλη του να διαιρείται με το άθροισμά της

Τη. Μιχαηλίδης Η.Υ. Τελεμαχωνιδών & Δασκίων, Πανεπιστήμιο Θεσσαλίας

39

Παράδειγμα των L_r και L_c

$$L = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}, \quad L_r = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix},$$

and

$$L_c = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \end{matrix}.$$

Οι πίνακες H και A

- Ο πίνακας H , SALS hub matrix, αποτελείται από τις μη-μηδενικές γραμμές και στήλες του $L_r L_c^T$
- Ο πίνακας A , SALS authority matrix, αποτελείται από τις μη-μηδενικές γραμμές και στήλες του πίνακα $L_c^T L_r$

Οι πίνακες $L_r L_c^T$ και $L_c^T L_r$

$$L_r L_c^T = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & 0 & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 1 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix} \end{matrix}, \quad L_c^T L_r = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 & \frac{5}{8} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}.$$

Οι πίνακες \mathbf{H} και \mathbf{A}

$$\mathbf{H} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{5} & \frac{1}{5} & 0 & 0 & \frac{1}{5} \end{pmatrix} \end{matrix}.$$

$$\mathbf{A} = \begin{matrix} & \begin{matrix} 1 & 3 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 3 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{5}{6} \end{pmatrix} \end{matrix}.$$

Τα Μηχανών ΗΥ, Τηλεπικοινωνιών & Δικτύων, Πανεπιστήμιο Θεσσαλίας

43

Ιδιοδιανύσματα

- $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$
- $\mathbf{v}^T\mathbf{A} = \lambda\mathbf{v}^T$
- Αριθμητικός υπολογισμός: [Power μέθοδος](#)

Τα Μηχανών ΗΥ, Τηλεπικοινωνιών & Δικτύων, Πανεπιστήμιο Θεσσαλίας

44

Η power μέθοδος

- $\mathbf{X}_{k+1} = \mathbf{A}\mathbf{X}_k$
- $\mathbf{X}_{k+1}^T = \mathbf{X}_k^T\mathbf{A}$
- Συγκλίνει στο κυρίαρχο ιδιοδιάνυσμα (**dominant eigenvector**), δηλ., σ' αυτό που αντιστοιχεί στην κυρίαρχη ιδιοτιμή ($\lambda = 1$).

Τα Μηχανών ΗΥ, Τηλεπικοινωνιών & Δικτύων, Πανεπιστήμιο Θεσσαλίας

45



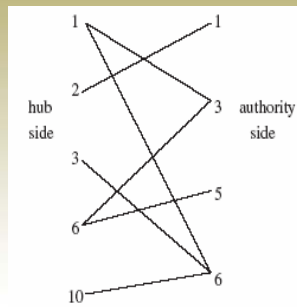
Η power μέθοδος

- Οι πίνακες **H** και **A** πρέπει να είναι **irreducible**, ώστε να συγκλίνει η power μέθοδος σε ένα **unique eigenvector**, ξεκινώντας από κάποια αρχική τιμή
- Εάν το neighborhood γράφημα **G** είναι συνδεδεμένο (**connected**), τότε ο **H** και ο **A** είναι irreducible
- Εάν το **G** δεν είναι συνδεδεμένο, τότε η εκτέλεση της power μεθόδου στον **H** και **A** δεν θα έχει ως αποτέλεσμα τη σύγκλιση σε ένα μοναδικό κυρίαρχο ιδιοδιάνυσμα



Στο παράδειγμα το G δεν είναι συνδεδεμένο

- Είναι προφανές ότι το γράφημα δεν είναι συνδεδεμένο, αφού η σελίδα 2 στο σύνολο hub συνδέεται μόνο με τη σελίδα 1 στο σύνολο authority, και αντίστροφα
- Οι **H** και **A** είναι reducible και επομένως περιέχουν **multiple irreducible connected components**





Connected Components

- Ο πίνακας **H** περιέχει δυο connected components, $C = \{2\}$ και $D = \{1, 3, 6, 10\}$
- Ο πίνακας **A** περιέχει δυο connected components, $E = \{1\}$ και $F = \{3, 5, 6\}$



Cutting και Pasting. Μέρος Ι

- Εκτελούμε την **power method** σε κάθε συνιστώσα των **H** και **A**

$$\pi_h^T(C) = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \pi_h^T(D) = \begin{pmatrix} 1 & 3 & 6 & 10 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \end{pmatrix},$$

$$\pi_a^T(E) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \pi_a^T(F) = \begin{pmatrix} 3 & 5 & 6 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{pmatrix}.$$



Cutting και Pasting. Μέρος ΙΙ

- Ενόνομε τις δυο συνιστώσες για κάθε πίνακα
- Πρέπει να πολλαπλασιάσουμε κάθε στοιχείο του διανύσματος με το κατάλληλο **weight**

$$H: \pi_h^T = \begin{pmatrix} 1 & 2 & 3 & 6 & 10 \\ \frac{1}{6} \cdot \frac{1}{3} & \frac{1}{3} \cdot 1 & \frac{1}{6} \cdot \frac{1}{6} & \frac{1}{3} \cdot \frac{1}{3} & \frac{1}{6} \cdot \frac{1}{6} \end{pmatrix} \\ = \begin{pmatrix} .2667 & .2 & .1333 & .2667 & .1333 \end{pmatrix}.$$

$$A: \pi_a^T = \begin{pmatrix} 1 & 3 & 5 & 6 \\ \frac{1}{4} \cdot 1 & \frac{3}{4} \cdot \frac{1}{3} & \frac{5}{4} \cdot \frac{1}{6} & \frac{6}{4} \cdot \frac{1}{2} \end{pmatrix} \\ = \begin{pmatrix} .25 & .25 & .125 & .375 \end{pmatrix}.$$

SALSA hub ranking: 1/6 2 3/10
 HITS hub ranking: 1 3/6/10 2 5
 SALSA authority ranking: 6 1/3 5
 HITS hub ranking: 6 3 5 1 2/10



Πλεονεκτήματα/Μειονεκτήματα του

- Δεν εμφανίζει το φαινόμενο **topic drift**, όπως ο HITS
- Παρέχει **authority** και **hub scores**
- **Χειρίζεται το spamming** καλύτερα από ότι ο HITS, αλλά όχι τόσο καλά όσο ο PageRank
- Είναι **query-dependent**
