Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

TÀI LIỆU LÝ THUYẾT KHAI THÁC WEB

Chủ đề 4

KHAI THÁC CẤU TRÚC WEB (PHẦN 2)

Giảng viên: ThS. Lê Ngọc Thành
Email: lnthanh@fit.hcmus.edu.vn

HKII/2011-2012

NỘI DUNG

- Phân tích mạng xã hội
  - Khái niệm
  - Tính trung tâm
  - Tính uy tín
- Đồng trích dẫn và mối nối danh mục
- Mô hình xếp hạng trang Web
  - PageRank
  - HITS

Phân tích mạng xã hội

- Phân tích mạng xã hội (social network analysis) là nghiên cứu các đối tượng xã hội (như con người trong một tổ chức,…và được gọi là các tác nhân) và các mối quan hệ hay tương tác giữa chúng.

Thể hiện mạng xã hội

- Các mối quan hệ và tương tác có thể được thể hiện bằng *một mạng lưới hay đồ thị* trong đó mỗi đỉnh (hay node) thể hiện một tác nhân và mỗi liên kết thể hiện một quan hệ.
- Từ mạng lưới này, chúng ta có thể nghiên cứu các thuộc tính của cấu trúc và vai trò cũng như uy tín của mỗi tác nhân xã hội.

Phân tích mạng XH trên Web

- Do Web về cơ bản là một *mạng xã hội ảo* ở đó mỗi trang có thể được xem như một tác nhân xã hội và mỗi liên kết như là một quan hệ.
- Rất nhiều kết quả từ mạng xã hội có thể được áp dụng và mở rộng cho việc sử dụng trong ngữ cảnh Web.

Một Số Khái Niệm Trong Phân Tích Mạng Xã Hội

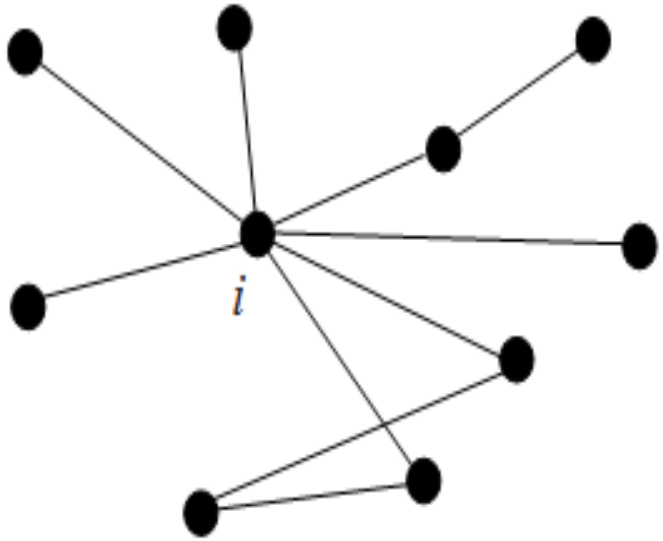
-

Tính Trung Tâm

Tính trung tâm

- *Tính trung tâm* (centrality): Một tác nhân mang tính trung tâm nghĩa là tác nhân đó liên quan đến rất nhiều các quan hệ.
- Ví dụ: trong một tổ chức, một người với các quan hệ rộng hay các giao tiếp với rất nhiều người khác trong tổ chức thì được xem như có nhiều quan trọng hơn một người với ít quan hệ hơn.

Ví dụ tính trung tâm



- Tác nhân i là tác nhân trung tâm nhất bởi vì nó có thể giao tiếp với nhiều tác nhân khác nhất.

Các loại tính trung tâm

- Có 3 loại phổ biến:
  - *Tính trung tâm bậc* (degree centrality)
  - *Tính trung tâm gần* (closeness centrality)
  - *Tính trung tâm trung gian* (betweeness centrality)
- Mỗi loại đều được xem xét trên cả hai đồ thị có hướng và đồ thị vô hướng

Tính trung tâm bậc

- Đối với đồ thị vô hướng, *trung tâm bậc của tác nhân i* (kí hiệu là $C_D(i)$) là bậc của node tác nhân i, kí hiệu là d(i) và được chuẩn hóa với bậc cực đại, n-1.

$$C_D(i) = \frac{d(i)}{n-1}$$

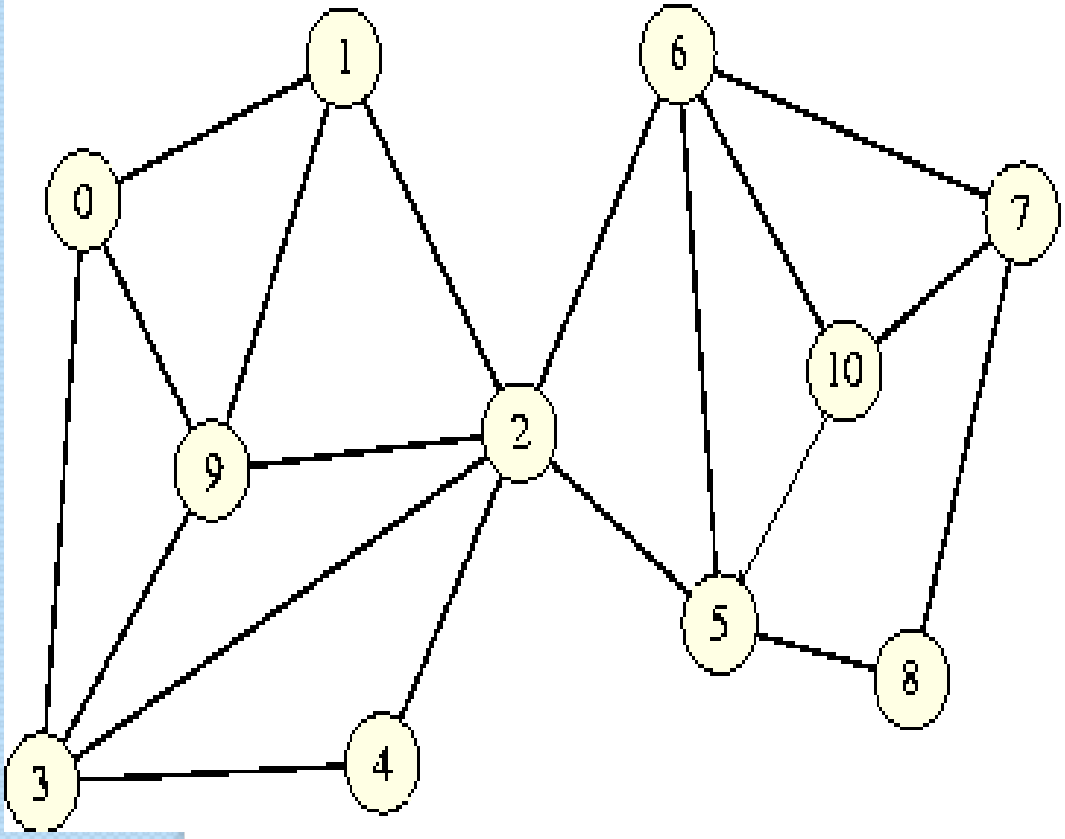
trong đó n là tổng số các tác nhân trong mạng.

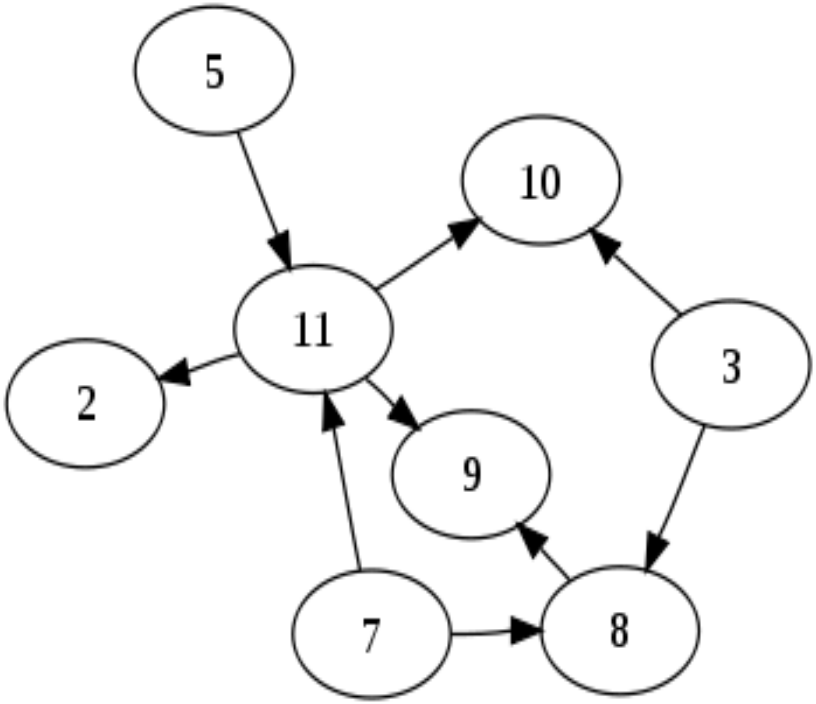
Tính trung tâm bậc (tt)

- Đối với đồ thị có hướng, chúng ta cần phân biệt liên kết đến và liên kết đi của tác nhân i .
- Trung tâm bậc được định nghĩa chỉ *dựa trên bậc ngoài* (số liên kết đi), $d_o(i)$:

$$C_D(i) = \frac{d_o(i)}{n-1}$$

Bài tập tính trung tâm bậc





- Xác định tính trung tâm bậc của mọi đỉnh trong hai đồ thị trên.

Tính trung tâm gần

- *Tính trung tâm gần* (closeness centrality): một tác nhân i là trung tâm gần nếu nó có thể tương tác ***dễ dàng*** với tất cả các tác nhân khác.
- Hay, khoảng cách của i đến tất cả tác nhân khác đều ngắn.

Tính trung tâm gần (tt)

- *Khoảng cách ngắn nhất* từ tác nhân i đến tác nhân j (kí hiệu d(i,j)) được đo bằng số liên kết trên đường đi ngắn nhất.
- Tính trung tâm gần của tác nhân i được kí hiệu là $C_c(i)$ và được chuẩn hóa với n-1 là tổng các khoảng cách ngắn nhất từ i đến tất cả các tác nhân khác.

Tính trung tâm gần (tt)

- Đối với đồ thị vô hướng: trung tâm gần $C_c(i)$ của tác nhân i được định nghĩa như:

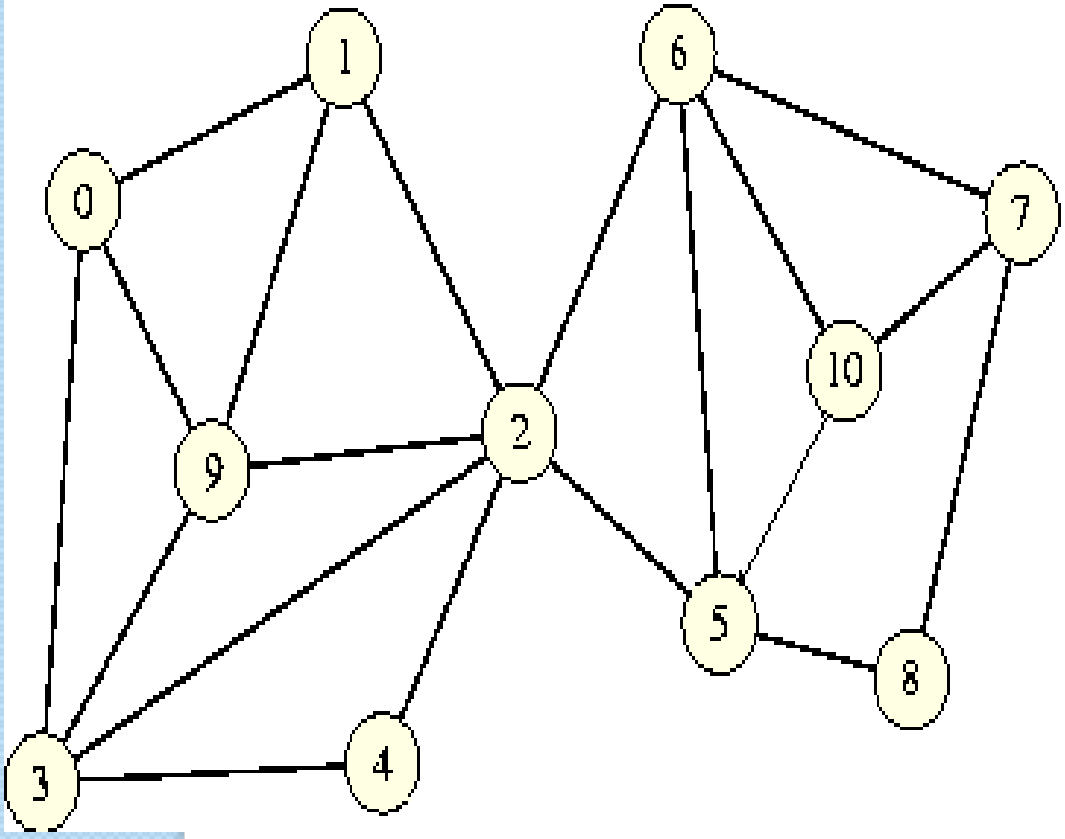
$$C_c(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

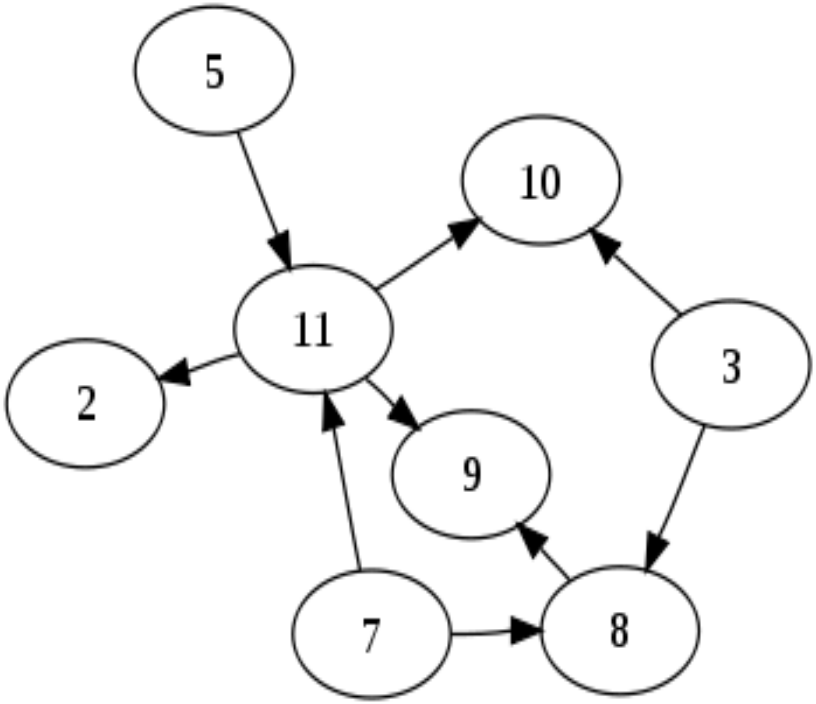
Lưu ý: biểu thức này chỉ thực hiện được trong trường hợp đồ thị liên thông

Tính trung tâm gần (tt)

- Đối với đồ thị có hướng, biểu thức tương tự có thể được sử dụng.
- Tính toán khoảng cách cần xem xét các hướng của liên kết.

Bài tập tính trung tâm gần





- Xác định tính trung tâm gần cho mọi định trong hai đồ thị trên.

Tính trung tâm trung gian

- Tính *trung tâm trung gian* (betweenness centrality):Nếu hai tác nhân j và k không kề nhau muốn tương tác và tác nhân i nằm giữa j và k thì i có thể có một số kiểm soát lên các tương tác của chúng.
- Nếu i ở trên đường đi của rất nhiều các tương tác loại này thì i là một tác nhân quan trọng.

Tính trung tâm trung gian (tt)

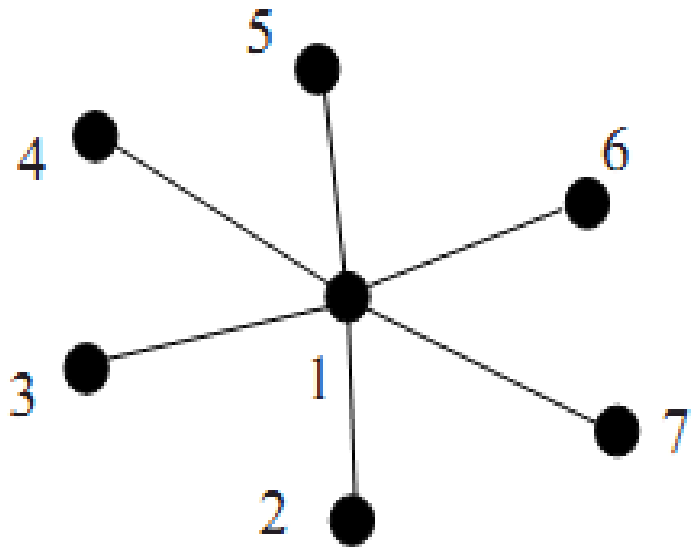
- Đối với đồ thị vô hướng, tính trung gian của một tác nhân i được định nghĩa bằng số lượng đường đi ngắn nhất qua i (kí hiệu $p_{jk}(i)$, $j \neq i$ và $k \neq i$) và được chuẩn hóa bởi tổng số lượng đường đi ngắn nhất của tất cả các cặp tác nhân ngoại trừ i:

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$

Tính trung tâm trung gian (tt)

- $C_B(i)$ có giá trị nhỏ nhất là 0, đạt được khi i nằm trên đường đi không ngắn nhất.
- Cực đại của nó là (n-1)(n-2)/2, số các cặp tác nhân ngoại trừ i.

Tính trung tâm trung gian (tt)



- Với hình trên, tác nhân 1 là tác nhân trung tâm nhất.
- Do nó nằm trên tất cả 15 đường đi ngắn nhất liên kết với 6 tác nhân khác.
- $C_B(1)$ có giá trị cực đại là 15 và $C_B(2) = C_B(3) = C_B(4) = C_B(5) = C_B(6) = C_B(7) = 0$

Tính trung tâm trung gian (tt)

- Để đảm bảo giá trị nằm giữa 0 và 1, $C_B(i)$ được chuẩn hóa với (n-1)(n-2)/2, đó là giá trị cực đại của $C_B(i)$:

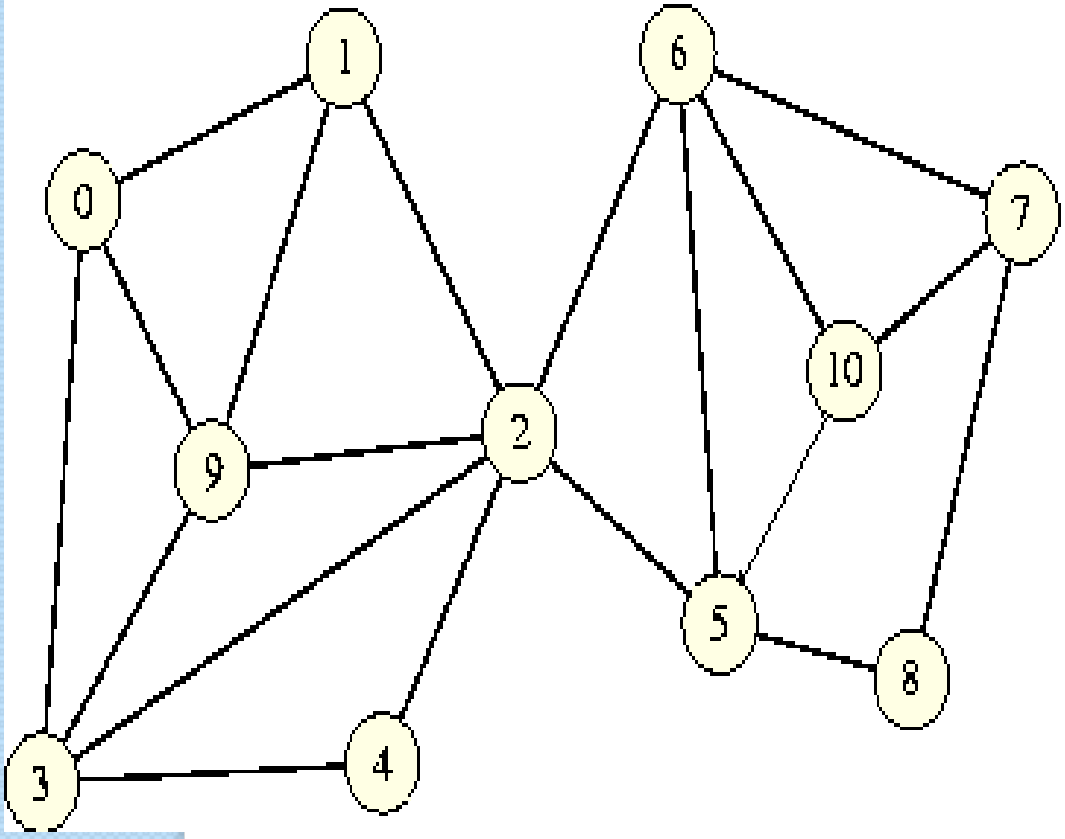
$$C_B(i) = \frac{2\sum_{j<k}\frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

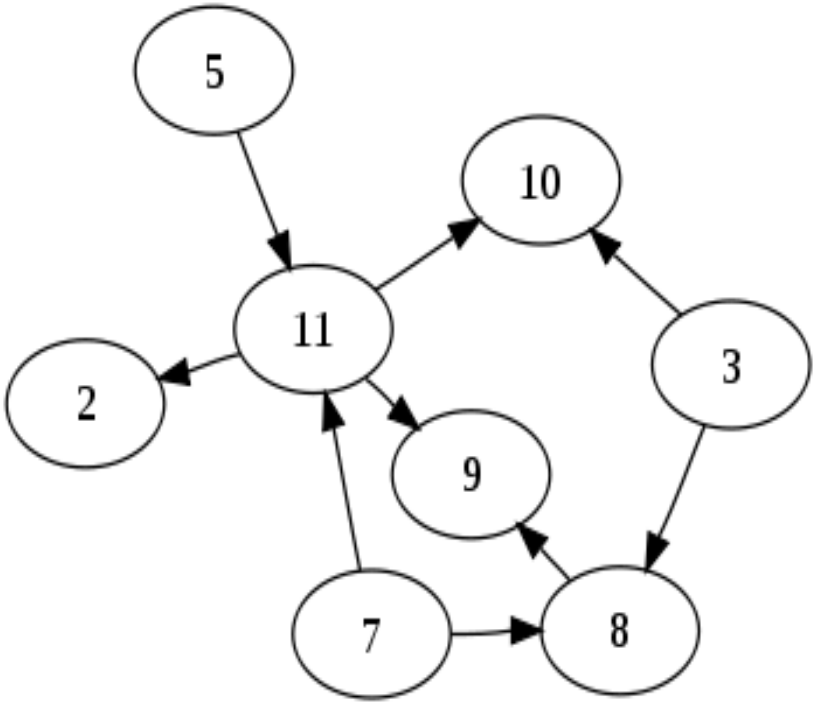
- Không giống như độ đo tính gần, tính trung gian có thể được tính thậm chí nếu đồ thị không liên thông.

Tính trung tâm trung gian (tt)

- Đối với đồ thị có hướng, biểu thước tương tự có thể được sử dụng.
- Nhưng phải nhân với 2 bởi vì bây giờ có (n-1)(n-2) cặp do một đường đi từ j đến k thì khác với một đường đi từ k đến j.
- Tương tự, p_{jk} phải xem xét các đường đi từ cả hai hướng.

Bài tập tính trung tâm trung gian





- Xác định tính trung tâm trung gian cho mọi định trong hai đồ thị trên.

Một Số Khái Niệm Trong Phân Tích Mạng Xã Hội

-

Tính Uy Tín

Tính uy tín

- *Tính uy tín* (prestige): là một độ đo xác định tính nổi bật của một tác nhân.
- Một tác nhân uy tín được xem như là đối tượng nhận các quan hệ mở rộng.
- Khi xét độ uy tín của một tác nhân, chúng ta chỉ nhìn vào các quan hệ hướng đến tác nhân (liên kết trong).
- Tính uy tín không thể được tính toán nếu như các quan hệ không có hướng hay trong đồ thị vô hướng.

Tính uy tín (tt)

- Sự khác nhau chính giữa các khái niệm tính trung tâm và tính uy tín là ở chỗ tính trung tâm tập trung các liên kết ngoài trong khi tính uy tín tập trung trên liên kết trong.

Các loại tính uy tín

- Có 3 loại phổ biến:
  - Tính uy tín bậc (degree prestige)
  - Tính uy tín xấp xĩ (proximity prestige)
  - Tính uy tín hạng (rank prestige)
- Tính uy tín hạng hình thành nên cơ sở của thuật toán phân tích liên kết trang Web bao gồm cả PageRank và HITS.

Tính uy tín bậc

- *Tính uy tín bậc* (degree prestige): độ đo đơn giản nhất của tính uy tín của một tác nhân i (kí hiệu $P_D(i)$) là bậc trong của nó,

$$P_D(i) = \frac{d_I(i)}{n-1}$$

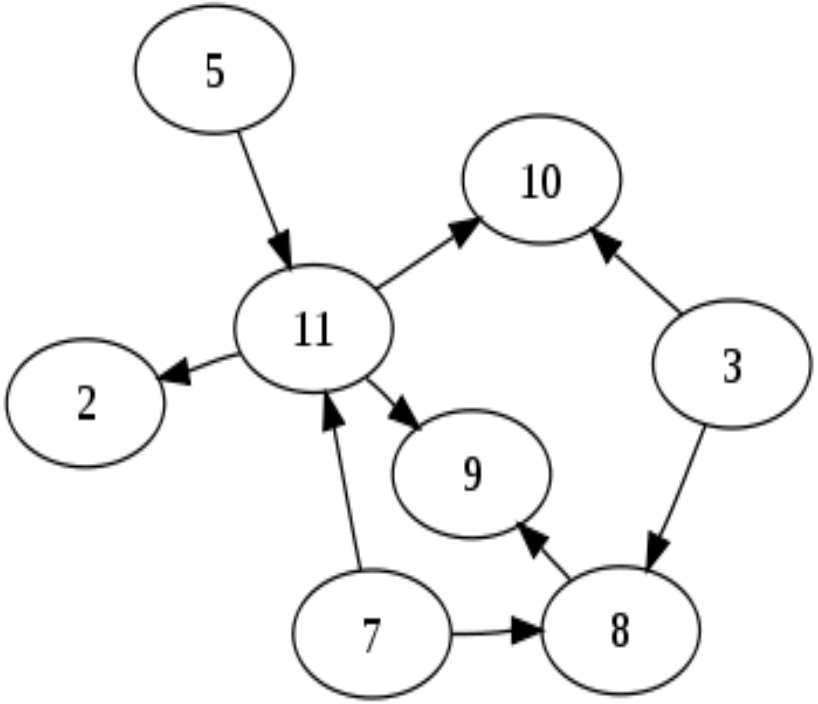
với $d_I(i)$ là bậc trong của i (số lượng liên kết trong của i)

n là tổng số tác nhân trong mạng.

Tính uy tín bậc (tt)

- Việc chia cho n-1 sẽ chuẩn hóa giá trị uy tín vào khoảng 0 đến 1.
- Giá trị uy tín lớn nhất là 1 khi mỗi tác nhân khác liên kết đến hay chọn tác nhân i

Bài tập tính uy tín bậc



- Xác định tính uy tín bậc cho mọi định trong đồ thị trên.

Tính uy tín xấp xĩ

- *Uy tín xấp xĩ* (proximity prestige): uy tín được xem xét cả hai tác nhân trực tiếp và gián tiếp liên kết tác tác nhân i.
- Nghĩa là, chúng ta xem mọi tác nhân j có thể đến i.

Tính uy tín xấp xĩ (tt)

- Gọi I_i là tập của các tác nhân mà có thể đến tác nhân i và cũng được gọi là miền ảnh hưởng của tác nhân i.
- Gọi d(j,i) kí hiệu cho khoảng cách đường đi ngắn nhất từ tác nhân j đến tác nhân i.
- Độ uy tín xấp xĩ được tính:

$$\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}$$

với $|I_i|$ là kích thước của tập I_i.

Tính uy tín xấp xĩ (tt)

- Để chuẩn hóa về miền [0,1], ta tính khoảng cách xấp xĩ của các tác nhân đến i:

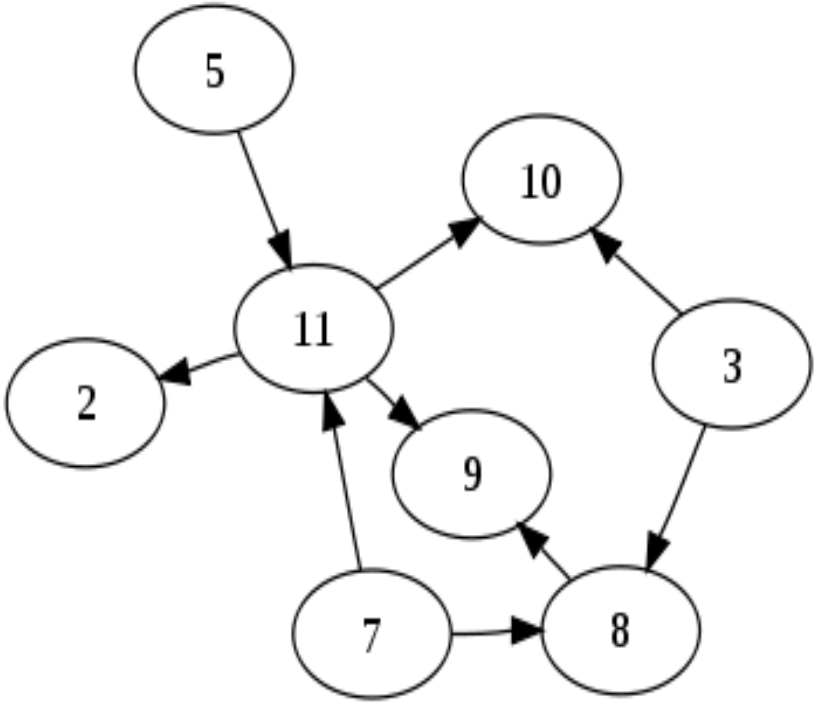
$$P_P(i) = \frac{|I_i|/(n-1)}{(\sum_{j=1}^{n} d(i,j))/|I_i|}$$

với $|I_i|/(n-1)$ là tỉ lệ các tác nhân có thể đến được tác nhân i.

Tính uy tín xấp xĩ (tt)

- Trong trường hợp mọi tác nhân đều có thể đến i, $|I_i|/(n-1) = 1$.
- Tỉ lệ là 1 nếu mọi tác nhân đều kề với i, $P_P(i) = 1$.
- Trong trường hợp không có tác nhân nào có thể đến i thì $|I_i| = 0$ và $P_P(i) = 0$.

Bài tập tính uy tín xấp xĩ



- Xác định tính uy tín xấp xĩ cho mọi định trong đồ thị trên.

Uy tín hạng

- *Uy tín hạng* (rank prestige): một người i được chọn bởi một người quan trọng sẽ có nhiều uy tín hơn được chọn bởi một người ít quan trọng.
- Ví dụ, một CEO công ty bình chọn cho một người thì có nhiều độ quan trọng hơn một công nhân bình chọn cho người này.
- Vì vậy uy tín của một người chịu ảnh hưởng bởi các xếp hạng hay trạng thái của các tác nhân liên quan.

Uy tín hạng (tt)

- Uy tín hạng $P_R(i)$ được định nghĩa như một liên kết tuyến tính của các liên kết trỏ đến i:

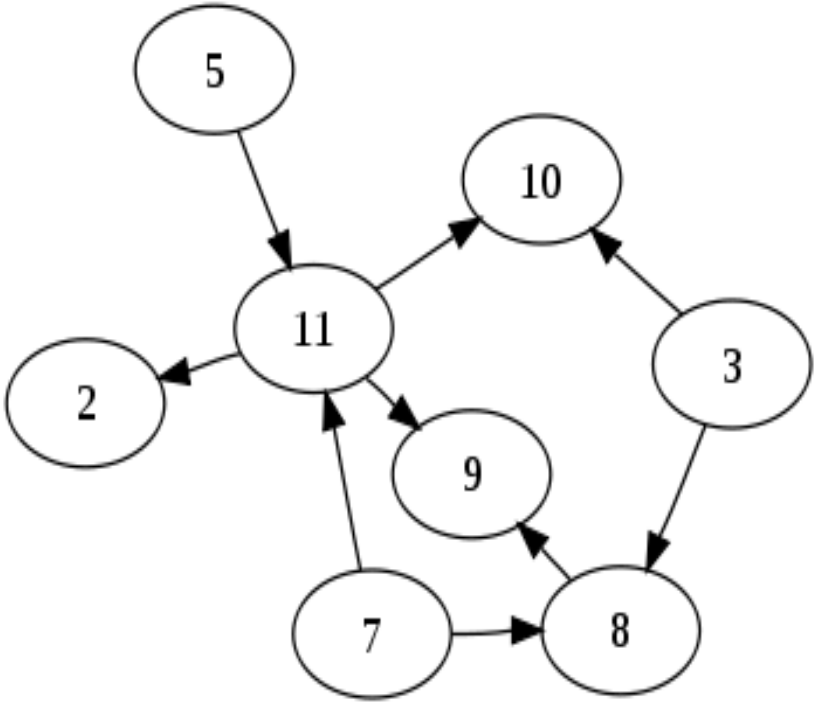
$$P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + \ldots + A_{ni}P_R(n)$$

- Trong đó $A_{ji} = 1$ nếu j trỏ đến i và 0 nếu ngược lại.

Uy tín hạng (tt)

Biểu thức trên chỉ ra rằng uy tín hạng của một tác nhân là một hàm của những hạng của các tác nhân, người mà bình chọn hay chọn tác nhân này.

Bài tập tính uy tín hạng



- Xác định tính uy tín hạng cho mọi đỉnh trong đồ thị trên. Giả sử khởi tạo ban đầu uy tín hạng của mọi đỉnh đều là 0.1

Uy tín hạng (tt)

- Chúng ta có n biểu thức của n tác nhân, nên có thể viết chung trong một ma trân đơn vị.
- P được thể hiện như vector chứa tất cả giá trị uy tín hạng:

$$P = (P_R(1), P_R(2), \ldots, P_R(n))^T.$$

- Ma trận A (trong đó $A_{ij} = 1$ nếu i trỏ đến j và 0 nếu ngược lại) để thể hiện ma trận kề của đồ thị:

$$P = A^T P$$

NỘI DUNG

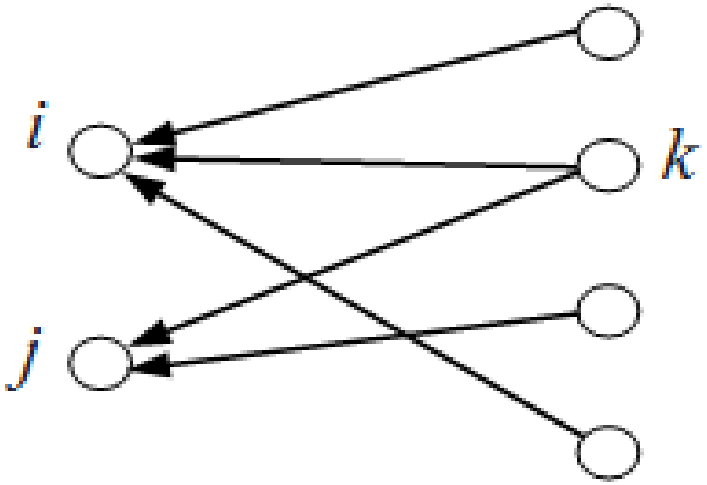
- Phân tích mạng xã hội
  - Khái niệm
  - Tính trung tâm
  - Tính uy tín
- Đồng trích dẫn và mối nối danh mục
- Mô hình xếp hạng trang Web
  - PageRank
  - HITS

Trích dẫn

- *Trích dẫn* (citation): Một cộng đồng học thuật luôn luôn trích dẫn đến công việc liên quan trước đây để thừa nhận nguồn gốc của một số ý tưởng trong cộng đồng và để so sánh phương pháp mới với công việc hiện tại.

Đồng trích dẫn

- Đồng trích dẫn (co-citation): nếu bài báo k đều cùng trích dẫn cả hai bài báo i và j thì hai bài báo i và j có thể liên quan với nhau trong một số ngữ cảnh.



Đồng trích dẫn (tt)

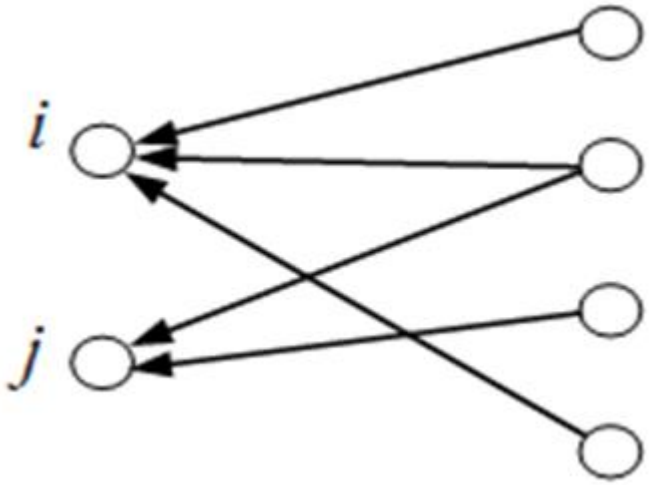
- Gọi L là ma trận trích dẫn, với L_{ki} = 1 nếu bài báo k trích dẫn bài báo i, và 0 nếu ngược lại.
- Đồng trích dẫn (C_{ij}) là số lượng các bài báo đồng trích dẫn i và j:

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$$

trong đó n là tổng số bài báo.

- Từ C_{ij} người ta hình thành ma trận đồng trích dẫn C của các cặp (i,j) khác nhau.

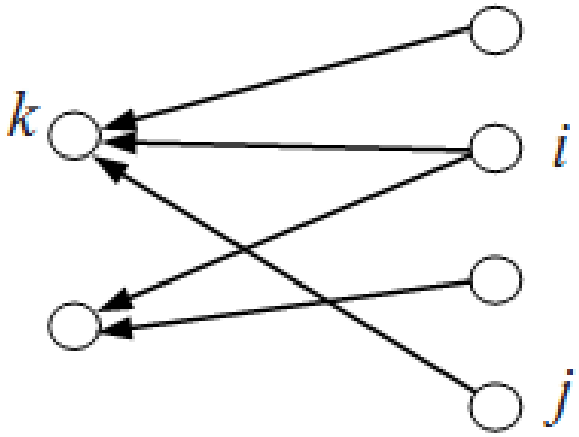
Xác định đồng trích dẫn



$C_{ij} = ?$

Mối nối danh mục

- Mối nối danh mục (Bibliographic Coupling): nếu cả hai bài báo i và j trích dẫn bài báo k thì chúng có thể liên quan với nhau.



- Ngược lại với đồng trích dẫn

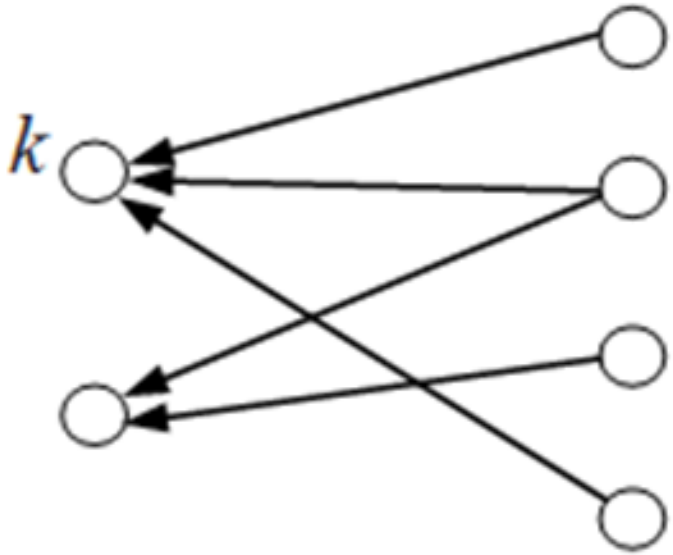
Mối nối danh mục

- Mối nối danh mục (B_{ij}) là số lượng bài báo được trích dẫn bởi cả hai bài báo i và j:

$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk}$$

- Từ B_{ij} hình thành ma trận mối nối danh mục B của các cặp (i,j) khác nhau.

Xác định mối nối thư mục



$B_{ij} = ?$

NỘI DUNG

- Phân tích mạng xã hội
  - Khái niệm
  - Tính trung tâm
  - Tính uy tín
- Đồng trích dẫn và mối nối danh mục
- Mô hình xếp hạng trang Web
  - PageRank
  - HITS

Giới thiệu PageRank

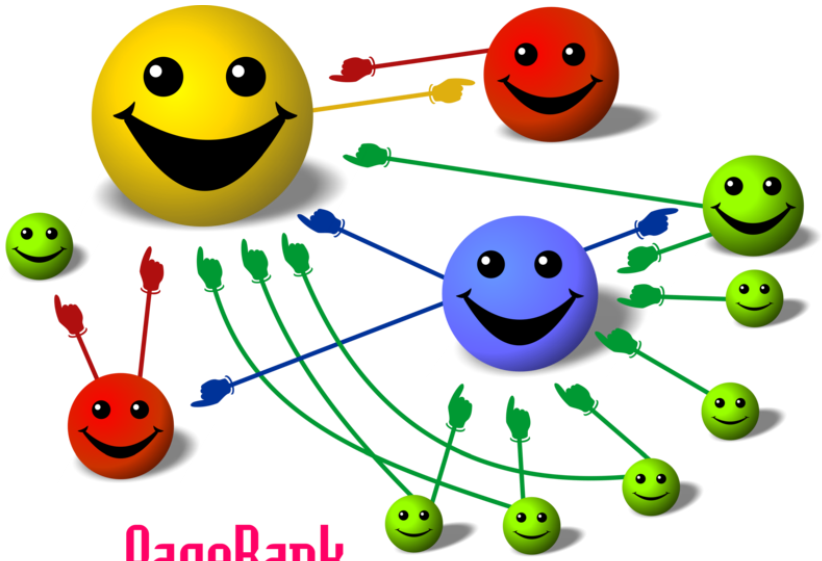
- Tác giả: Sergey Brin và Larry Page
- Đăng tại hội nghị WWW lần 7 vào tháng 7, 1998.
- Google
  dựa trên đó.

Thuật toán PageRank

- Là thuật toán xếp hạng tĩnh
- Độ đo hạng của một trang offline gọi là giá trị PageRank và giá trị này không phụ thuộc vào câu truy vấn
- Ý tưởng: dựa trên tính uy tín (prestige)

Nhắc lại tính uy tín – 1/2

- Một siêu kiên kết từ một trang đến trang khác là sự truyền đạt ngầm tính uy tín đến trang web này.
- Vì vậy, càng nhiều liên kết trong mà một trang i nhận được, thì trang i càng uy tín.



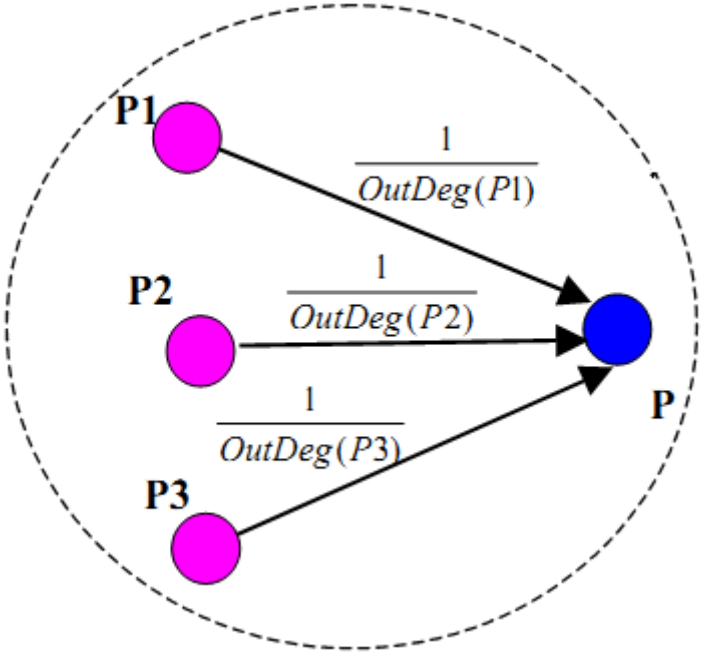
Nhắc lại tính uy tín – 2/2

- Các trang trỏ đến trang i cũng có độ uy tín của riêng nó.
- Một trang với một độ uy tín cao hơn trỏ đến i thì có nhiều quan trọng hơn một trang với độ uy tín thấp hơn trỏ đến i.
- Vì vậy một trang là quan trọng nếu nó được trỏ bởi nhiều trang quan trọng khác.

Thuật toán PageRank (tt)

- Độ quan trọng của trang i (hay điểm PageRank của i) được xác định bởi tổng các điểm PageRank của tất cả trang mà trỏ đến i.
- Do một trang có thể trỏ đến nhiều trang khác, điểm uy tín của nó sẽ được chia ra giữa tất cả các trang mà nó trỏ đến.

Thuật toán PageRank (tt)

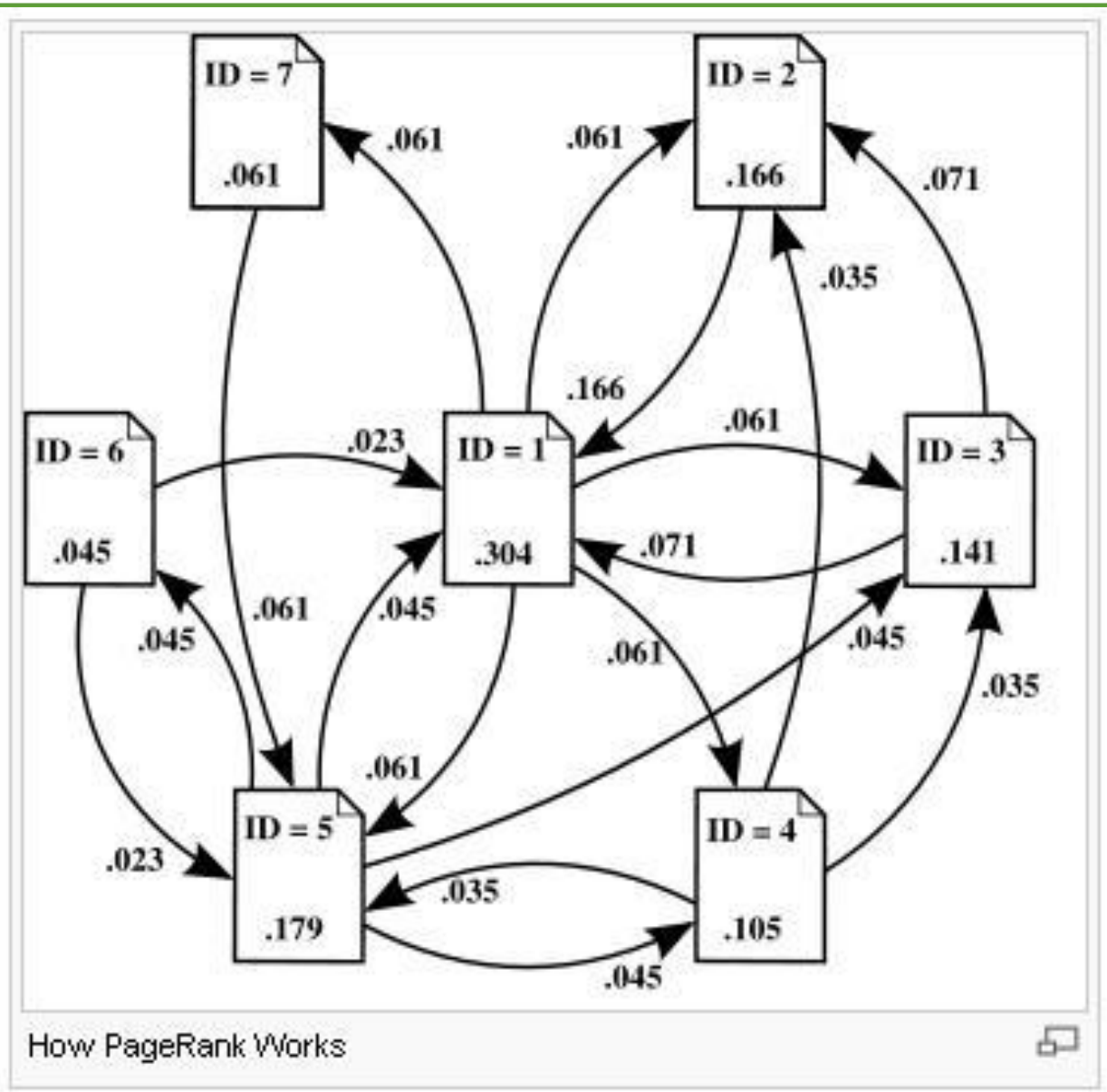


Hạng của một trang web phụ thuộc trên hạng của các trang trỏ đến nó

$$PR(P) = \frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}$$

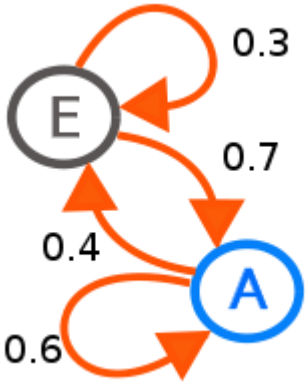
OutDeg (Pi) là số liên kết ngoài của trang Pi

Ví dụ điểm PageRank



Nhận xét

- Điểm PageRank cũng thể hiển sự chuyển đổi của người dùng từ một trang này đến một trang khác.
- Mỗi siêu liên kết đóng vai trò như xác suất chuyển đổi.
- Giống với mô hình chuỗi Markov (Markov Chain)



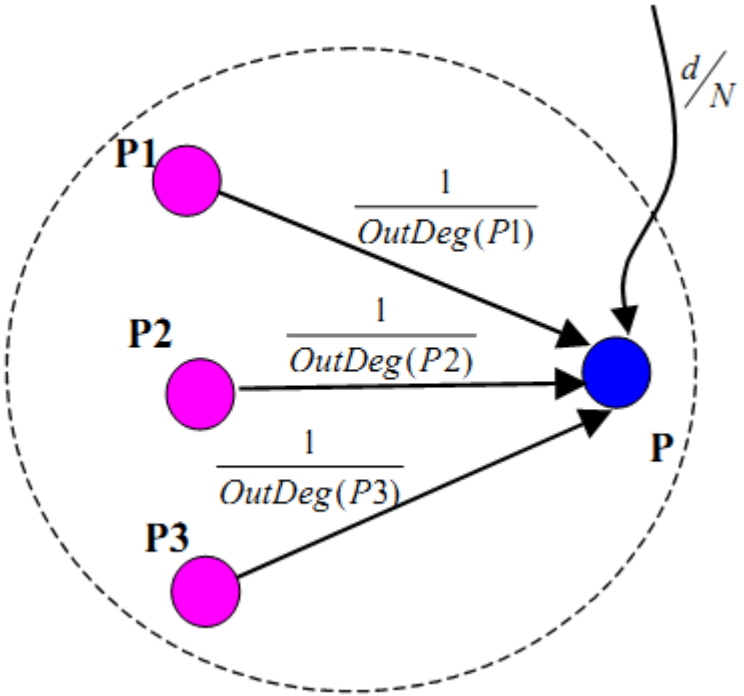
Chuỗi Markov và PageRank

- Việc lướt web có thể được mô hình hóa bởi chuỗi Markov.
- Bài toán lướt web trở thành bài toán xác suất.
- Người lướt Web sẽ click trên các siêu liên kết trong trang i hoàn toàn ngẫu nhiên với xác suất là 1/OutDeg(i)

Chuỗi Markov và PageRank

- Ngoài việc click trên các liên kết, người lướt web có thể nhảy đến một trang mà không cần đến siêu liên kết.
- Vì vậy, trong mô hình này, người lướt web có hai lựa chọn:
  - Với xác suất d, nhảy đến một trang ngẫu nhiên mà không cần đến siêu liên kết.
  - Với xác suất (1-d), chọn ngẫu nhiên một siêu liết kết để đến trang khác

Mô hình cải tiến



Mô hình với 2 khả năng chọn lựa của người dùng:
- Bấm click
- Nhập URL

$$PR(P) = d/N + (1-d)\left(\frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}\right)$$

N là số node trong đồ thị hay tổng số trang trên web
d gọi là hệ số tắt dần ∈ [0,1]

Công thức tổng quát

- Gọi A là ma trận với:

$$A_{ij} = \begin{cases} \frac{1}{OutDeg_i} & nếu\ có\ liên\ kết\ i\ đến\ j \\ 0 & nếu\ ngược\ lại \end{cases}$$

- Gọi D là xác suất nhảy ngẫu nhiên của các trang web $[1/N, 1/N, \ldots, 1/N]^T$
- **Công thức tổng quát:**

$$\mathbf{PR = d*D + (1-d)*A^T*PR}$$

Thuật toán PageRank

Đặt **PR** ← $[r_1, r_2,\ldots,r_N]$ trong đó r_i là hạng khởi tạo ban đầu của trang i, N là tổng số trang trên Web.

d ← 0.85; **D** ← $[1/N, 1/N, \ldots,1/N]^T$

A là ma trận kề xác suất chuyển

do

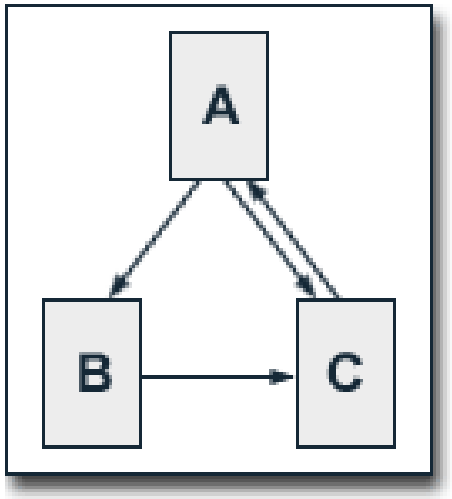
$$\mathbf{PR}_{i+1} \leftarrow d*\mathbf{D} + (1-d)*\mathbf{A}^T*\mathbf{PR}_i$$

$$\delta \leftarrow |\mathbf{PR}_{i+1} - \mathbf{PR}_i|$$

while $\delta > \varepsilon$ (với ε là ngưỡng hội tụ)

return PR

Ví dụ PageRank



Chọn d = 0.5

PR(A) = 0.5/3 + 0.5 PR(C)

PR(B) = 0.5/3 + 0.5 (PR(A) / 2)

PR(C) = 0.5/3 + 0.5 (PR(A) / 2 + PR(B))

Bài tập PageRank

- Tính giá trị PageRank của các trang trong hình slide trước qua 12 vòng lặp

Vòng lặp thứ	PR(A)	PR(B)	PR(C)
0	1	1	1
1	…	…	…
2	…	…	…
…	…	…	…
…	…	…	…
12	…	…	…

Điểm mạnh và yếu của PageRank

- Điểm mạnh:
  - Khả năng chống lại spam: do không dễ gì để thêm liên kết đến trang mình vào các trang quan trọng khác
  - Điểm PageRank là điểm toàn cục và không phụ thuộc vào câu truy vấn: giá trị PageRank của tất cả các trang Web được tính toán và lưu trữ offline chứ không phải tại thời điểm truy vấn

Điểm mạnh và yếu của PageRank

- Điểm yếu:
  - Độc lập với truy vấn: PageRank không phân biệt giữa trang uy tín chung và những trang uy tín trên chủ đề truy vấn
  - Không xem xét tính thời gian: Web là môi trường động và thay đổi liên tục. Những trang chất lượng trong quá khứ có thể không là các trang chất lượng bây giờ hay trong tương lai.

Điểm mạnh và yếu của PageRank

- Lưu ý: hiện nay, Google không chỉ xếp hạng trang dựa trên liên kết cũng như việc chọn tham số, cách khắc phục vấn đề trên là một điều bí mật.

NỘI DUNG

- Phân tích mạng xã hội
  - Khái niệm
  - Tính trung tâm
  - Tính uy tín
- Đồng trích dẫn và mối nối danh mục
- Mô hình xếp hạng trang Web
  - PageRank
  - HITS

Giới thiệu HITS

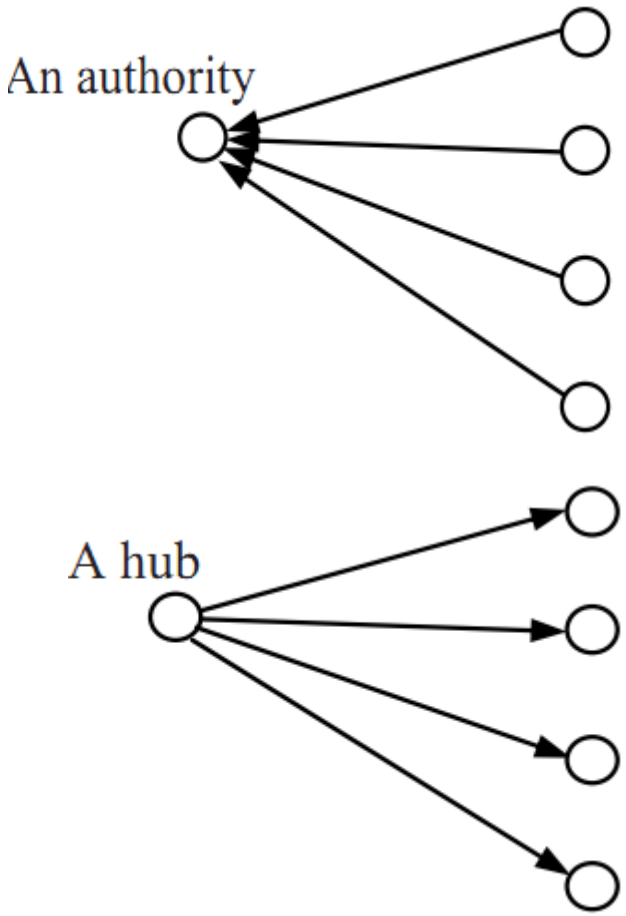
- Tác giả Jon Kleinberg
- Đăng tại hội nghị ACM-SIAM, tháng 1 năm 1998

Thuật toán HITS

- HITS (Hypetext Induced Topic Search): là thuật toán xếp hạng phụ thuộc vào câu truy vấn tìm kiếm.
- Với một câu truy vấn, HITS tìm kiếm các trang liên quan và sau đó sinh ra hai cách xếp hạng:
  - Xếp hạng authority
  - Xếp hạng hub

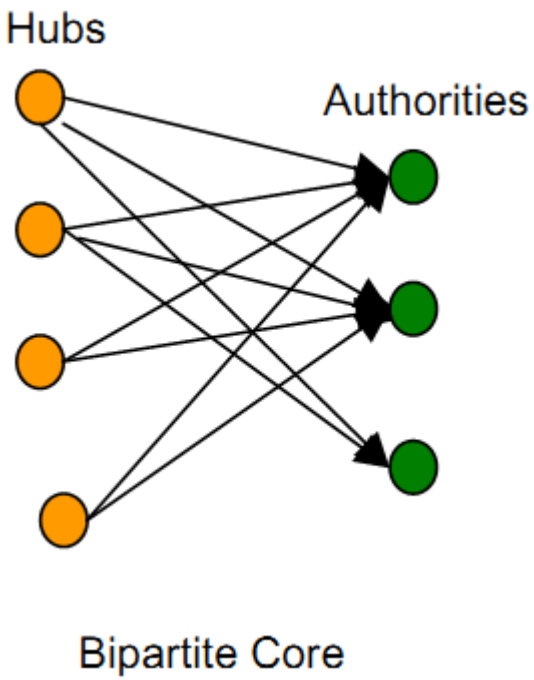
Nhắc lại Authority và Hub

- Một authority là một trang có nhiều liên kết đến.
  - Ý tưởng: trang có nhiều nội dung uy tín về một chủ đề nên rất nhiều người tin tưởng nó và liên kết đến nó.
- Một hub là một trang có nhiều liên kết ngoài.
  - Ý tưởng: khi người dùng đến một trang hub này, sẽ tìm thấy nhiều liên kết hữu ích mà sẽ đưa họ đến trang nội dung tốt trên chủ đề.

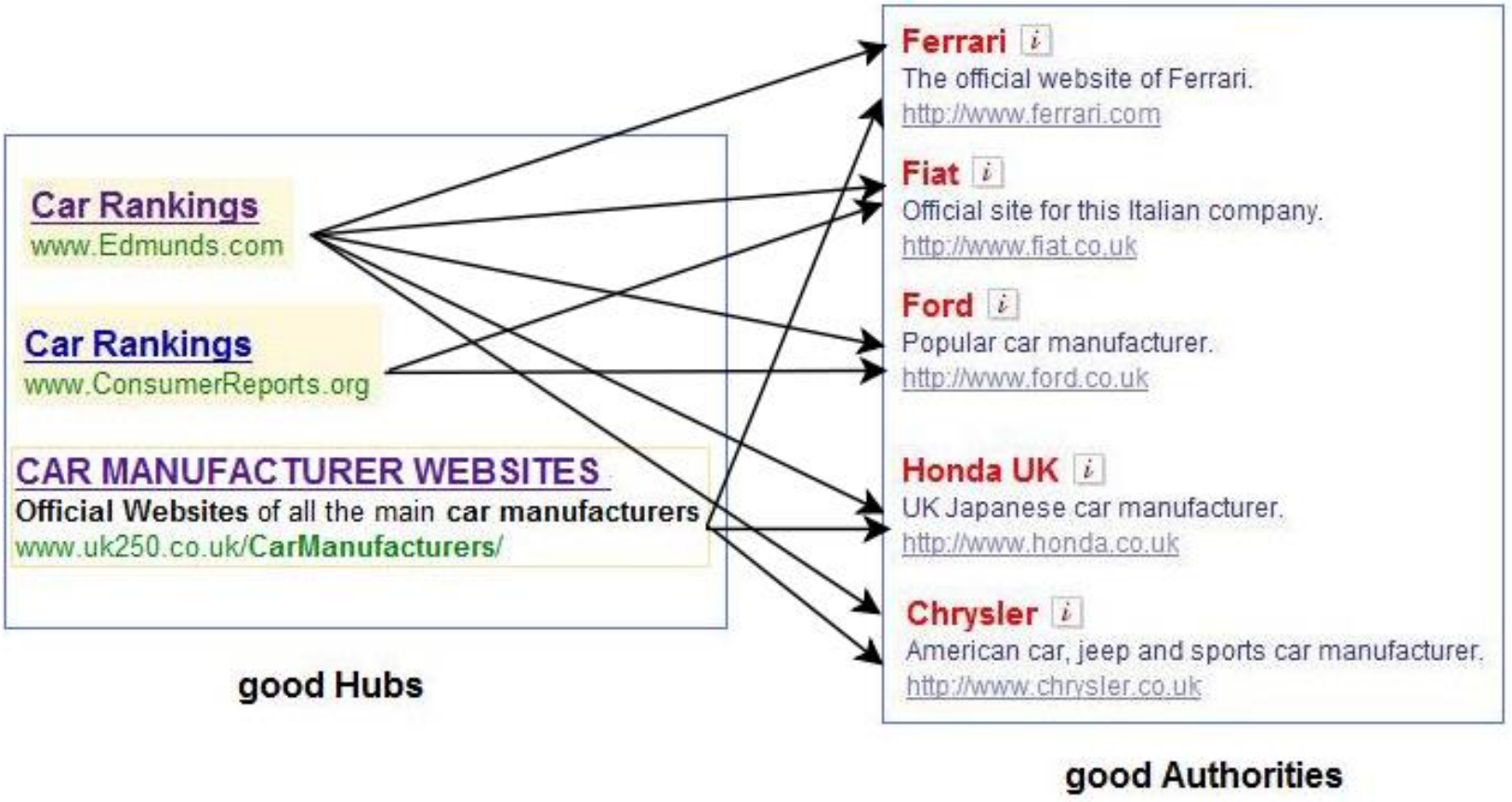


Ý tưởng HITS

- Ý tưởng chính của HITS là một điểm hub tốt sẽ chỉ đến nhiều authority tốt và một authority tốt được trỏ bởi nhiều hub tốt.
- Vì vậy, authority và hub có một mối quan hệ hỗ tương lẫn nhau.



Ví dụ HITS

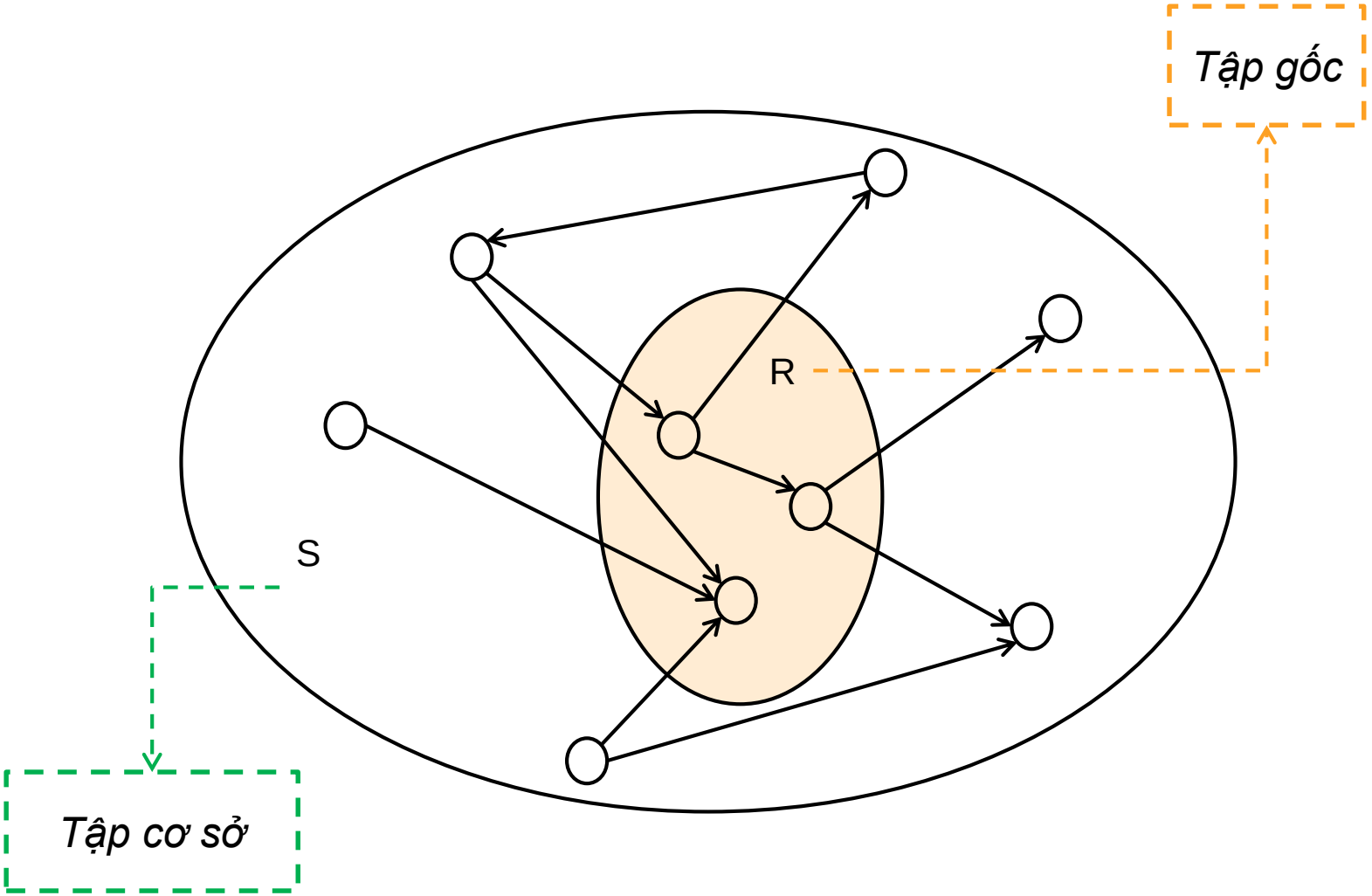


Query: Top automobile makers

Các bước thuật toán HITS

- B1: HITS gửi câu truy vấn q đến hệ thống tìm kiếm. Sau đó tập hợp t trang được xếp hạng cao nhất. Gọi tập này là tập gốc R.
- B2: Phát triển R bằng cách thêm vào bất kì trang được trỏ bởi R và trỏ đến R. Ta được tập lớn hơn gọi là tập cơ sở S.

Các bước thuật toán HITS (tt)



Các bước thuật toán HITS (tt)

- B3: Gán mỗi trang i $\in$ S một trọng số authority a(i) và trọng số hub h(i). Ban đầu gán khởi tạo tất cả trọng số đều bằng nhau.
- B4: Tính toán trọng số mới:

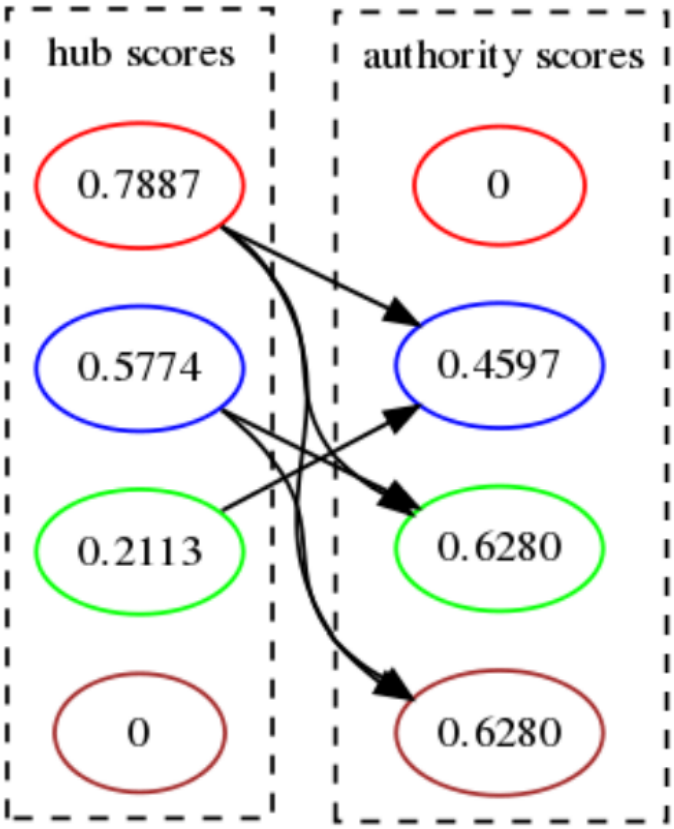
$$\mathrm{a}(i) = \sum_{(j,i)\in E} h(j)$$

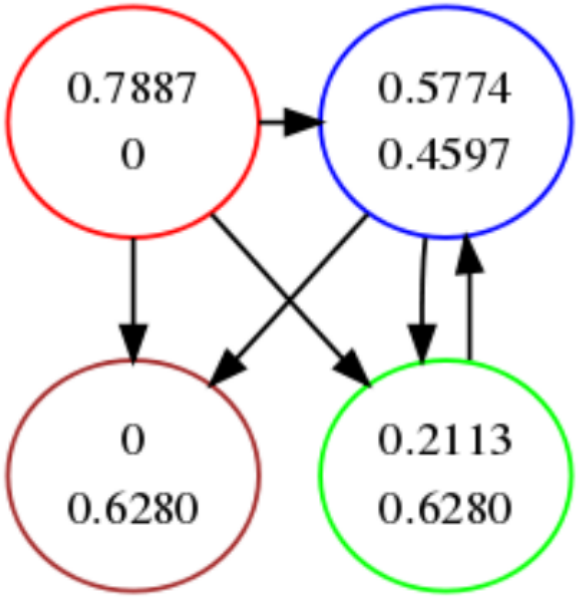
$$h(i) = \sum_{(i,j)\in E} a(j)$$

- B5: Chuẩn hóa trọng số mới để tổng là 1
- B6: Lặp lại B4 cho đến khi đạt đến một điểm cố định (hay thỏa ngưỡng ε_a và ε_h)

Các bước thuật toán HITS (tt)

- HITS sẽ chọn một vài trang có hạng authority và hub cao, sau đó trả chúng về người sử dụng.





Mã giả thuật toán HITS

HITS-Iterate(G)

$a_0 \leftarrow h_0 \leftarrow (1, 1, \ldots, 1)$;

$k \leftarrow 1$

Repeat

$a_k \leftarrow L^T L a_{k-1}$;

$h_k \leftarrow L L^T h_{k-1}$;

$a_k \leftarrow a_k / ||a_k||_1$; // normalization

$h_k \leftarrow h_k / ||h_k||_1$; // normalization

$k \leftarrow k + 1$;

until $||a_k - a_{k-1}||_1 < \varepsilon_a$ and $||h_k - h_{k-1}||_1 < \varepsilon_h$;

return a_k and h_k

$$L_{ij} = \begin{cases} 1 \text{ nếu } \exists(i,j) \\ 0 \text{ ngược lại} \end{cases}$$

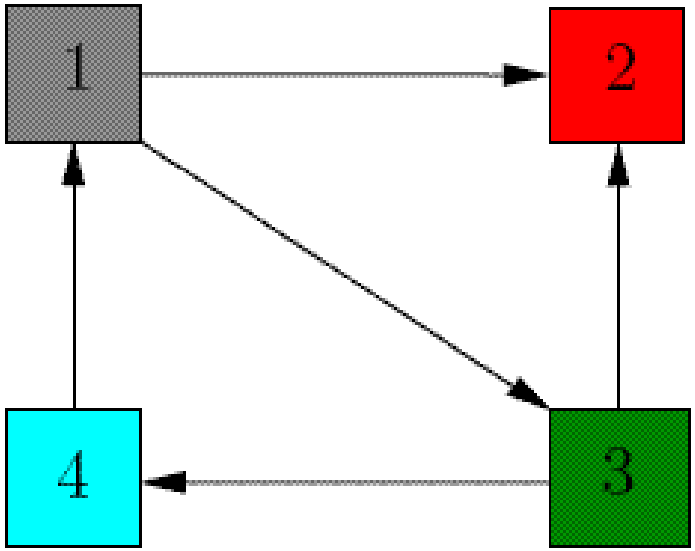
$\mathbf{a} = \mathbf{L}^T\mathbf{h}$

$\mathbf{h} = \mathbf{La}$

$\mathbf{a}_k = \mathbf{L}^T\mathbf{L}\mathbf{a}_{k-1}$

$\mathbf{h}_k = \mathbf{L}\mathbf{L}^T\mathbf{h}_{k-1}$

Bài tập HITS



- Tính toán trọng số hub và authority của đồ thị trên qua 3 vòng lặp.

HITS với Đồng trích dẫn

- Ma trận đồng trích dẫn C:

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj} = (L^T L)_{ij}$$

Ta thấy rằng ma trận authority ($L^T L$) của HITS là ma trận đồng trích dẫn C trong ngữ cảnh Web.

HITS với Mối nối danh mục

- Ma trận mối nối danh mục B:

$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk} = (LL^T)_{ij}$$

Điều này cũng chứng tỏ rằng ma trận hub (LL^T) của HITS là ma trận mối nối danh mục B trong ngữ cảnh Web.

Điểm mạnh và yếu của HITS

- HITS có các điểm mạnh:
  - Điểm mạnh chính của HITS là khả năng xếp hạng các trang theo chủ đề truy vấn, nó có thể cung cấp nhiều trang authority và hub liên quan.
  - Việc xếp hạng có thể cũng được kết hợp với truy vấn thông tin (information retrieval).

Điểm mạnh và yếu của HITS

- HITS có một vài điểm yếu:
  - HITS không có khả năng chống spam. Dễ dàng ảnh hưởng HITS bằng cách thêm liên kết ngoài từ một trang để trỏ đến rất nhiều authority tốt. Điều này đẩy điểm hub của trang lên rất mạnh. Bởi vì độ hub và authority phụ thuộc lẫn nhau, đến lượt nó cũng tăng độ authority của trang.

Điểm mạnh và yếu của HITS

- HITS có một vài điểm yếu (tt):
  - Vấn đề khác của HITS là trôi dạt chủ đề. Trong việc mở rộng tập gốc, nó có thể dễ dàng tập hợp rất nhiều trang mà không liên quan chủ đề tìm kiếm. Bởi vì liên kết ngoài của một trang có thể không trỏ đến trang mà nó liên quan và liên kết trong đến trang có thể cũng không liên quan bởi vì mọi người đặt siêu liên kết với rất nhiều lí do, kể cả spam.

Điểm mạnh và yếu của HITS

- HITS có một vài điểm yếu (tt):
  - Thời gian truy vấn cũng là một trở ngại chính. Có tập gốc, mở rộng nó và sau đó thực hiện tính toán vector riêng là những tiến trình tiêu tốn khá nhiều thời gian.

TÀI LIỆU THAM KHẢO

- Chapter 7. B. Liu, *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*, Springer Series on Data-Centric Systems and Applications, 2007.
- http://ignatius.atw.hu/mining.pdf
- http://www.cis.upenn.edu/~cis455/slides/18-PageRank.pptx

KẾT THÚC PHẦN 2