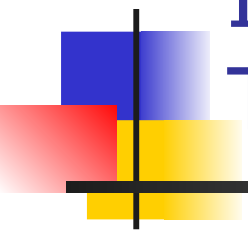
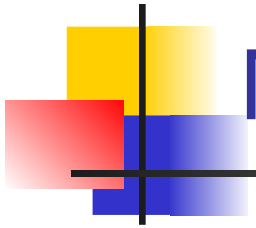


IT4853
Tìm kiếm và trình diễn thông tin

Bài 20. Phân tích liên kết, HITS

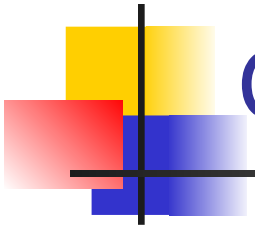
IIR.C21. Link analysis

Bộ môn Hệ thống thông tin
Viện CNTT & TT



Nội dung chính

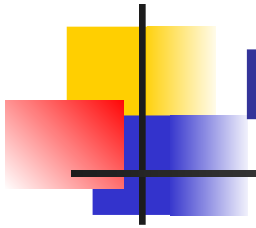
- Giải thuật HITS
- Tính hội tụ của giải thuật HITS



Giải thuật HITS

- Giải thuật HITS chia kết quả phù hợp trên Web thành hai nhóm:
  - **Nhóm 1.** *Hubs*. Nhóm các trang chứa liên kết đáp ứng tốt nhu cầu thông tin;
    - Ví dụ, cho truy vấn [ĐHBK Hà Nội]: Trang chứa danh sách tài liệu nói về trường ĐHBK Hà Nội là một hub.
  - **Nhóm 2.** *Authorities*. Nhóm các trang trực tiếp đáp ứng tốt nhu cầu thông tin.
    - Trang chủ của trường ĐHBK Hà Nội đối với truy vấn đã cho;

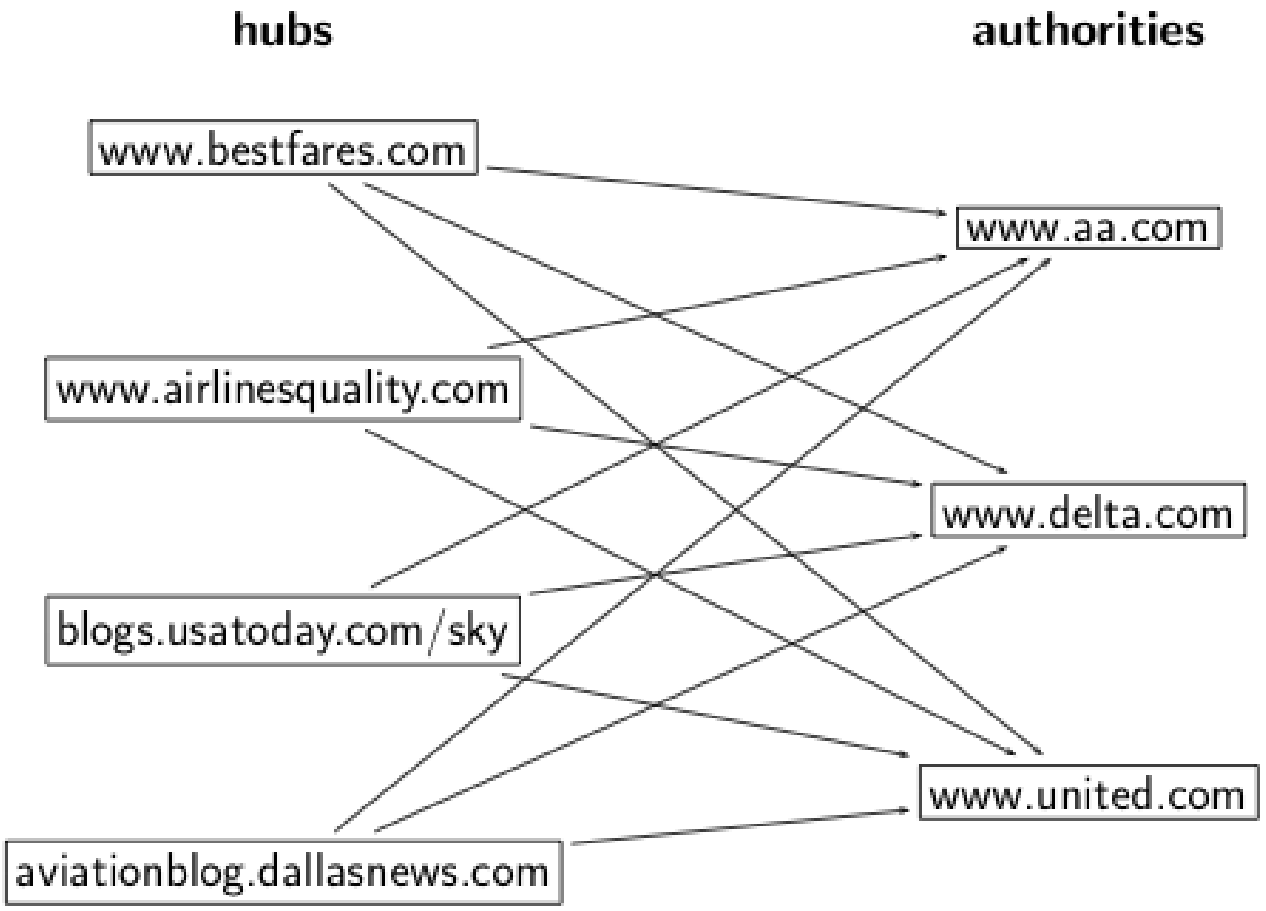
Hyperlink-Induced Topic Search (HITS), Klei98
Hầu hết các hệ thống tìm kiếm hiện nay không phân biệt hai dạng kết quả này.

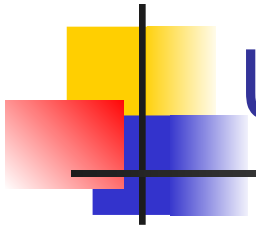


Khái niệm Hub và authority

- Một trang Hub về một chủ đề chứa nhiều liên kết tới những trang authorities thuộc chủ đề đó;
- Một trang Authority tốt được trích dẫn bởi nhiều trang Hub thuộc về chủ đề đó.
- Hub và Authority là hai khái niệm tương hỗ. Chúng ta sẽ tính Hub và Authority bằng vòng lặp.

Khái niệm Hub và authority (2)

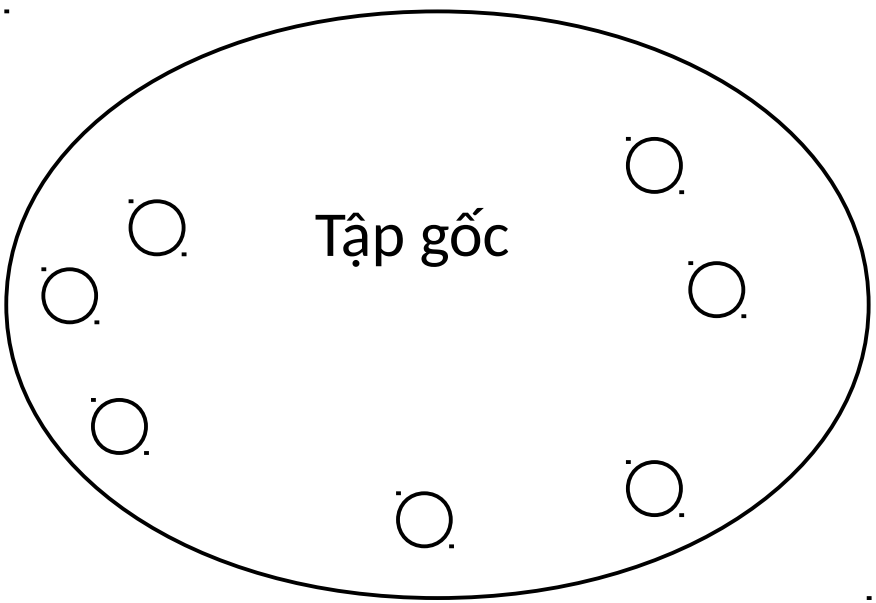




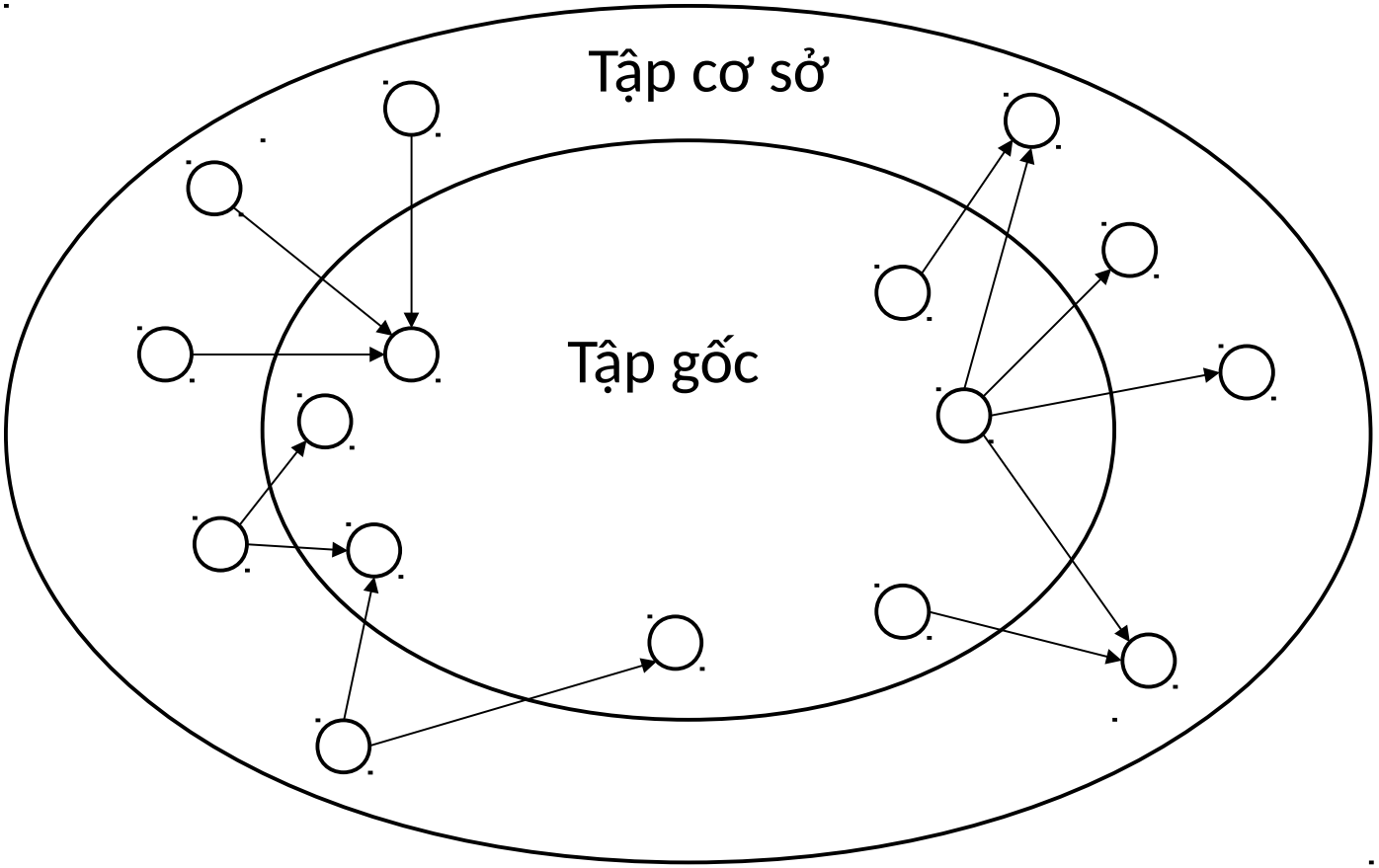
Ứng dụng HITS trong tìm kiếm

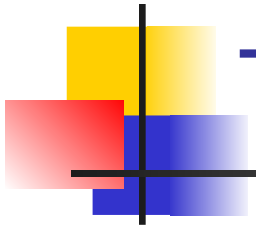
- Thực hiện tìm kiếm thông thường
  - Gọi kết quả tìm kiếm thu được là tập gốc
- Sau đó thêm vào tập gốc tất những trang liên quan đến trang bất kỳ trong tập gốc (theo in-link hoặc out-link)
  - Gọi tập thu được là tập cơ sở
- Cuối cùng, tính hubs và authorities cho tập cơ sở
- Xếp hạng các kết quả theo hub và authority
  - Hai danh sách kết quả tách biệt cho những trang có hub cao nhất và có authority cao nhất.

Tập gốc



Tập cơ sở





Tập gốc và tập cơ sở

- Tập gốc thường có 200-1000 trang.
- Tập cơ sở có thể chứa tới 5000 trang.
  - Phù hợp để xử lý trong quá trình thực hiện truy vấn.

[Klei98]

Ví dụ kết quả tìm kiếm
Truy vấn: japan elementary schools

Hubs

- schools
- LINK Page-13
- ˜ú–{,ÌŠw□Z
- □a‰„□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www...iglobe.ne.jp/~IKESAN
- ,I,f,j□¬Šw□Z,U˜N,P‘g•¨Œê
- □ÒŠ—'¬—§□ÒŠ—˜Œ□¬Šw□Z
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- –y˜i□¬Šw□Z,Ìƒz□[ƒ□ƒy□[ƒW
- UNIVERSITY
- ‰J—³□¬Šw□Z DRAGON97-TOP
- □Â‰⁸□¬Šw□Z,T˜N,P‘gƒz□[ƒ□ƒy□[ƒW
- ¶µ°é¼ÂÁ© ¥á¥Ë¥å¡¼ ¥á¥Ë¥å¡¼

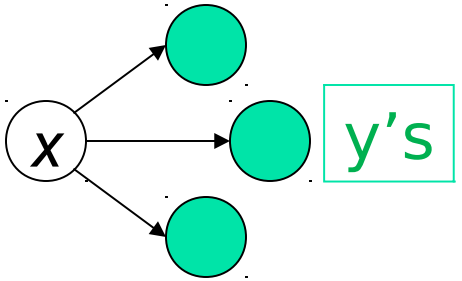
Authorities

- The American School in Japan
- The Link Page
- ‰⁸□è□s—§ˆä˜c□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
- Kids' Space
- ˆÀ□é□s—§ˆÀ□é□¼•”□¬Šw□Z
- ‹{□é‹³ç‘åŠw•□'®□¬Šw□Z
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- □_˜Þ□ìŒ§□E‰¡•I□s—§'†□ì□¼□¬Šw□Z,Ìƒy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

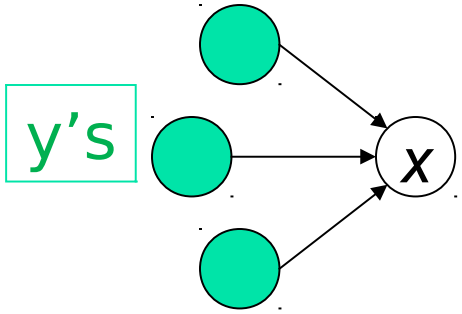
Tính hubs và authorities

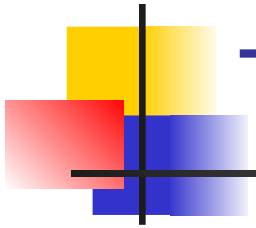
- Khởi tạo: với mọi trang *x*, *h(x)*←*1*; *a(x)* ←*1*;
- Lặp cập nhật *h(x)*, *a(x)*

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



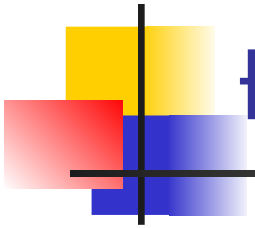
$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$





Tính hubs và authorities (2)

- Để ngăn a() và h() trở nên quá lớn, chúng ta có thể chia a() và h() cho hằng số sau mỗi bước;
  - Không ảnh hưởng đến kết quả tìm kiếm;
    - Chỉ quan trọng thứ tự, không quan trọng các giá trị cụ thể.

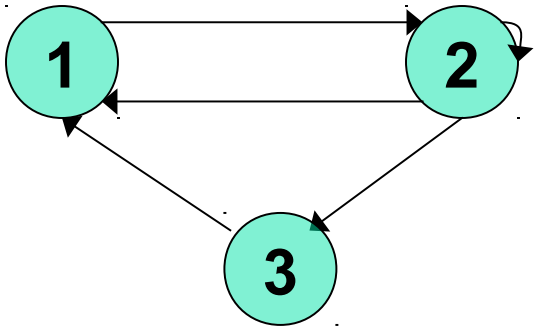


Đặc điểm của giải thuật HITS

- HITS có thể gom một vài trang web chất lượng tốt không phụ thuộc vào nội dung trang web;
- Sau khi thiết lập tập cơ sở, chúng ta chỉ thực hiện phân tích liên kết, không sử dụng nội dung;
- Trang web trong tập cơ sở có thể không chứa bất kỳ từ khóa truy vấn nào;
- Theo lý thuyết, đối với một truy vấn tiếng anh có thể trả về một trang tiếng nhật
  - Nếu tồn tại liên kết giữa những trang tiếng anh và tiếng nhật;
  - Cảnh báo: topic drift- các trang tìm được theo liên kết có thể hoàn toàn không liên quan đến câu truy vấn.

Biểu diễn luật cập nhật bằng các phép toán ma trận

- Đặt A là ma trận kề kích thước NxN:
  - N là kích thước tập cơ sở.
  - $A_{ij} = 1$ nếu tồn tại liên kết i→j và = 0 trong trường hợp ngược lại.



A=

	1	2	3
1	0	1	0
2	1	1	1
3	1	0	0

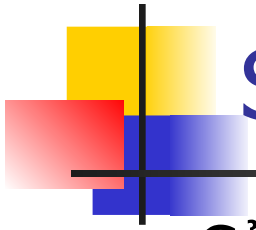
Biểu diễn luật cập nhật bằng các phép toán ma trận (2)

- Gọi $\vec{h}$ và $\vec{a}$ là các vec-tơ hub và authority.
- Có thể biểu diễn luật cập nhật như sau:

  $\vec{h}=A\vec{a}$; $\vec{a}=A^t\vec{h}$
- ➔ $\vec{h}=AA^t\vec{h}$ và $\vec{a}=A^tA\vec{a}$.
- Như vậy, $\vec{h}$ là vec-tơ riêng của AA^t và $\vec{a}$ là vec-tơ riêng của A^tA.
- Giải thuật HITS:
  - Tính $\vec{h} = A\vec{a}$
  - Tính $\vec{a} = A^T\vec{h}$
  - Lặp cho tới khi hội tụ

Tính hội tụ của giải thuật HITS

- Chúng ta có $\vec{h} = AA^T\vec{h}$ and $\vec{a} = A^TA\vec{a}$
- Như vậy $\vec{h}$ là vec-tơ riêng của AA^T và $\vec{a}$ là vec-tơ riêng của A^TA.
  - hubs và authorities là các vec-tơ riêng, có thể tính bằng phương pháp lũy thừa.

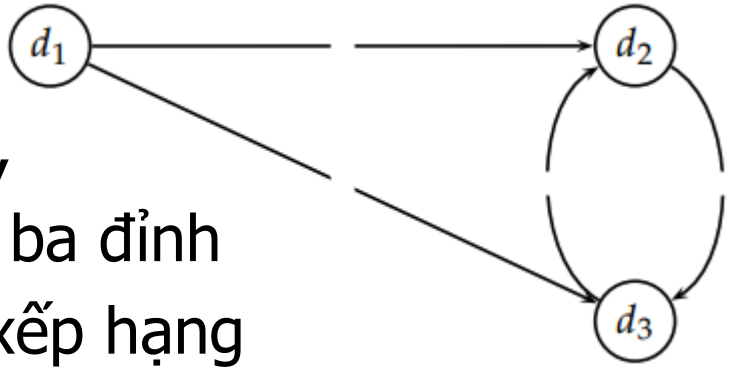


So sánh PageRank và HITS

- Cả HITS và PageRank đều được hình thức hóa bằng bài toán tìm vec-tơ riêng của ma trận.
- PageRank có thể tính trước, HITS phải được tính trong quá trình thực hiện truy vấn
  - Hạn chế tiềm năng ứng dụng thực tế vì khối lượng tính toán lớn.
- … tuy nhiên, có thể hoán đổi vị trí, áp dụng HITS cho toàn bộ Web và PageRank cho tập kết quả!
- Tuy nhiên: trên Web một trang hub thường đồng thời là một trang authority!
  - Như vậy khác biệt giữa xếp hạng theo HITS và theo PageRank có thể không quá lớn.

Bài tập 25.1

Cho đồ thị:



Hãy tính PageRank,

Hub, Authority cho ba đỉnh

của đồ thị này, và xếp hạng

các đỉnh theo các tiêu trí tính được, ghi chú các đỉnh đồng hạng.

PageRank: Theo mô hình duyệt web ngẫu nhiên với bước nhảy. Xác suất nhảy bằng 0.1

Hub/Authority: Chuẩn hóa các giá trị sao cho giá trị lớn nhất bằng 1.