

The image shows the HUST logo and course information on a white background with a pattern of blue dots. The logo consists of a red square with a white star and the text "ĐẠI HỌC BÁCH KHOA" in white, followed by "ĐẠI HỌC BÁCH KHOA HÀ NỘI" and "HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY" in black. Below the logo is the course title "Nhập môn Khoa học dữ liệu (IT4142)" in red, followed by the lecturers "PGS.TS Thân Quang Khoát & PGS.TS Phạm Văn Hải" and "Team lecturers" in red. At the bottom left is the slogan "ONE LOVE. ONE FUTURE." in red. The background is white with a pattern of blue dots of varying sizes arranged in a circular, pixelated-like pattern.

 **ĐẠI HỌC
BÁCH KHOA HÀ NỘI**
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

**Nhập môn
Khoa học dữ liệu
(IT4142)**

PGS.TS Thân Quang Khoát & PGS.TS Phạm Văn Hải
Team lecturers

ONE LOVE. ONE FUTURE.

2

Contents

- Lecture 1: Tổng quan về Khoa học dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Làm sạch và tích hợp dữ liệu
- Lecture 4: Phân tích và khám phá dữ liệu
- Lecture 5: Trực quan hoá dữ liệu
- Lecture 6: Trực quan hoá dữ liệu đa biến
- Lecture 7: Học máy
- Lecture 8: Phân tích dữ liệu lớn
- Lecture 9: Báo cáo tiến độ bài tập lớn và hướng dẫn
- **Lecture 10+11: Phân tích một số kiểu dữ liệu**
- Lecture 12: Đánh giá kết quả phân tích



Các bài toán chính trong phân tích liên kết

- Xếp hạng đồ thị: Phân tích vai trò của các đỉnh trong đồ thị
- Nhận diện cộng đồng: Phát hiện các cộng đồng bao gồm các thành viên có tính chất tương tự
- Dự đoán liên kết: Dự đoán sự tiến hóa của đồ thị theo thời gian
- Phân loại đồ thị: Phân loại các đỉnh và các cạnh của đồ thị vào các lớp cho trước



Nội dung

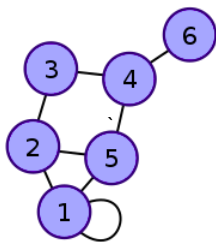
1. Xếp hạng đồ thị
2. Nhận diện cộng đồng
3. Học biểu diễn đồ thị



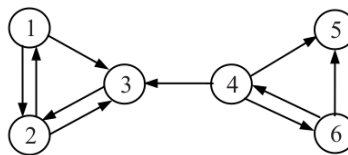
5

1. Xếp hạng đồ thị

- 1.1 Các khái niệm cơ bản của đồ thị



a) Đồ thị vô hướng



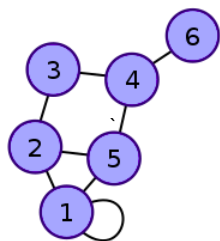
b) Đồ thị có hướng



6

Ma trận kề

$$a[i, j] \begin{cases} = 1 \text{ nếu tồn tại cạnh } (i, j) \\ = 0 \text{ nếu ngược lại} \\ = 2 \text{ nếu tồn tại cạnh từ một đỉnh đến chính nó} \end{cases}$$



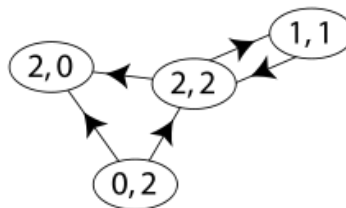
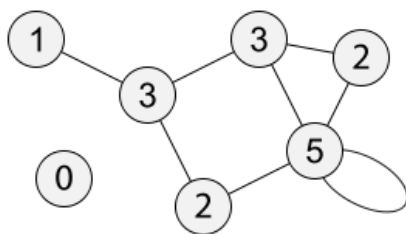
$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$



7

Bậc của đỉnh

- $d_i(i)$ = số nút trở tới i
- $d_o(i)$ = số nút i trở tới



8

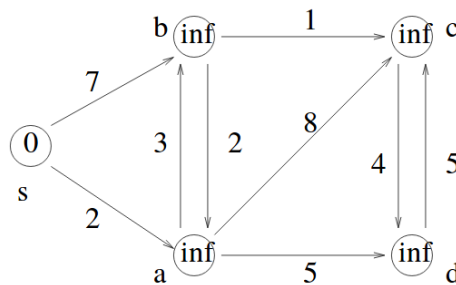
1.2 Thuật toán Dijkstra

- Tìm đường đi ngắn nhất từ một đỉnh s tới các đỉnh còn lại của đồ thị
- $d(v)$: Khoảng cách từ đỉnh v tới đỉnh s
 - **B1**: Khởi tạo $d(s) = 0$; $d(v) = \infty$
 - **B2**: Sắp xếp các đỉnh v theo một trật tự xác định trên hàng đợi Q
 - **B3**: Lấy một đỉnh u thuộc hàng đợi Q và cập nhật khoảng cách $d(v)$ (nếu cần) với mỗi đỉnh v liền kề với u
 - Quay lại **B2** cho đến khi xử lý hết các đỉnh



9

VD



10

VD (tiếp)

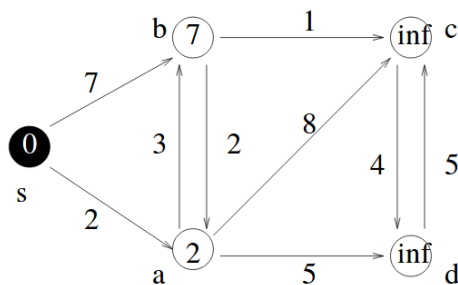
v	s	a	b	c	d
$d[v]$	0	∞	∞	∞	∞
$pred[v]$	nil	nil	nil	nil	nil
$color[v]$	W	W	W	W	W

v	s	a	b	c	d
$d[v]$	0	∞	∞	∞	∞



11

VD (tiếp)



12

VD (tiếp)

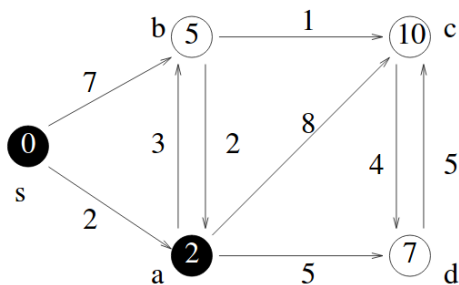
v	s	a	b	c	d
$d[v]$	0	2	7	∞	∞
$pred[v]$	nil	s	s	nil	nil
$color[v]$	B	W	W	W	W

v	a	b	c	d
$d[v]$	2	7	∞	∞



13

VD (tiếp)



14

VD (tiếp)

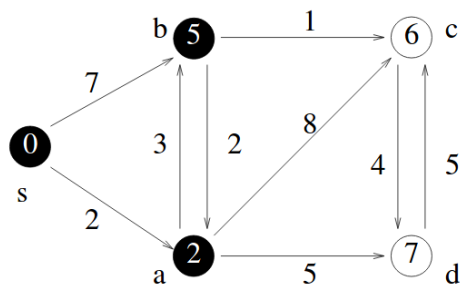
v	s	a	b	c	d
$d[v]$	0	2	5	10	7
$pred[v]$	nil	s	a	a	a
$color[v]$	B	B	W	W	W

v	b	c	d
$d[v]$	5	10	7



15

VD (tiếp)



16

VD (tiếp)

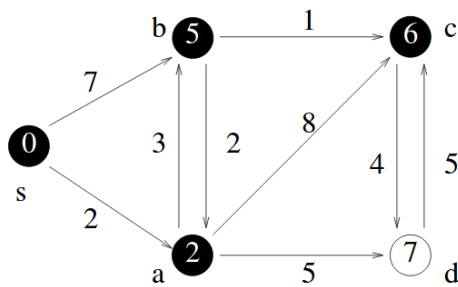
v	s	a	b	c	d
$d[v]$	0	2	5	6	7
$pred[v]$	nil	s	a	b	a
$color[v]$	B	B	B	W	W

v	c	d
$d[v]$	6	7



17

VD (tiếp)



18

VD (tiếp)

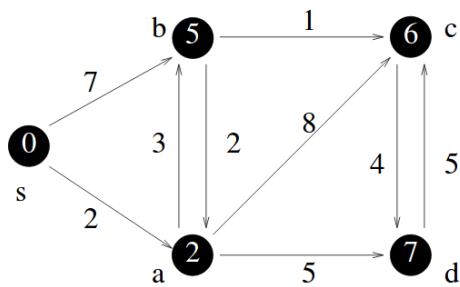
v	s	a	b	c	d
$d[v]$	0	2	5	6	7
$pred[v]$	nil	s	a	b	a
$color[v]$	B	B	B	B	W

v	d
$d[v]$	7



19

VD (tiếp)



20

VD (tiếp)

v	s	a	b	c	d
$d[v]$	0	2	5	6	7
$pred[v]$	nil	s	a	b	a
$color[v]$	B	B	B	B	B

$$Q = \emptyset.$$



21

1.3 Độ trung tâm: Độ trung tâm lân cận

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

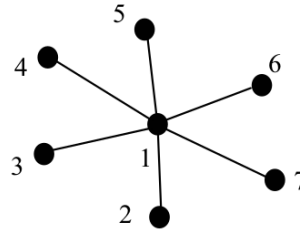
$d(i, j)$: Khoảng cách ngắn nhất từ nút i tới nút j



22

Độ trung tâm trung gian

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



$p_{jk}(i)$: Số lượng đường đi ngắn nhất từ j tới k mà đi qua i

$$C_B(1) = 15, C_B(2) = C_B(3) = C_B(4) = C_B(5) = C_B(6) = C_B(7) = 0$$



23

1.4 Độ quan trọng: Độ quan trọng theo bậc

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

$d_I(i)$: Số nút trở tới i



24

Độ quan trọng lân cận

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

I_i : Các nút có thể đi tới i



25

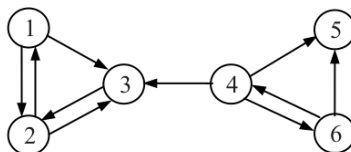
1.5 Thuật toán Pagerank

- Xếp hạng đồ thị dựa trên cấu trúc tổng quát
- Đối với các đồ thị lớn, thứ hạng được tính xấp xỉ bằng thuật toán lặp dựa trên 'random walk'
- Có ứng dụng quan trọng trong máy tìm kiếm web
- Nhược điểm: Không phụ thuộc vào câu truy vấn



26

Ma trận chuyển tiếp



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



27

Ma trận chuyển tiếp (tiếp)

- Chuẩn hóa:

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



28

Công thức xếp hạng

$$R(A) = (1 - d) / N + d * \sum_{B: (B,A) \in E} R(B) / d_o(B)$$

$R(A)$: Thứ hạng của đỉnh A

d : damping factor

N : số đỉnh của đồ thị

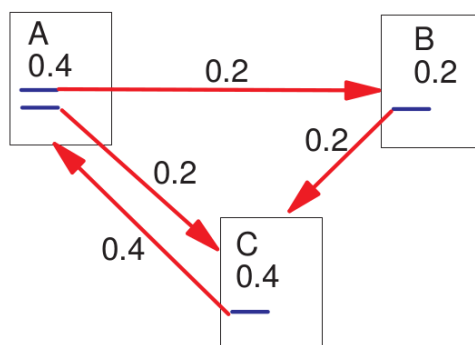
(B,A) cạnh của đồ thị

$d_o(B)$ bậc ra của đỉnh B



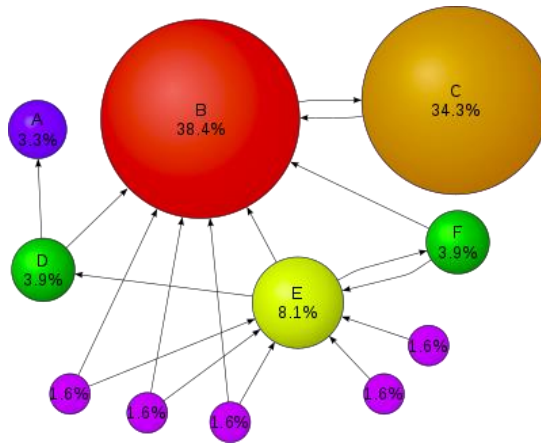
29

VD ($d = 1$)



30

VD ($d = 0.85$)



31

Thuật toán lặp

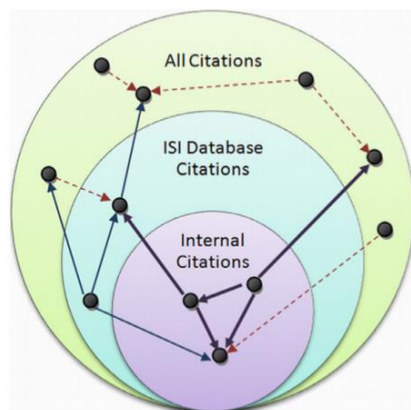
Algorithm PageRank(d, E)

1. Khởi tạo thứ hạng các trang $R^{(0)}$;
2. $i = 1$;
3. **repeat**
4. **for** mỗi trang A **do**
5. $R^{(i)}(A) = (1 - d) / N + d * \sum_{B: (B,A) \in E} R^{(i-1)}(B) / d_o(B)$;
6. **endfor**
7. $i++$;
8. **until** hội tụ



32

Ứng dụng: Phân tích trích dẫn

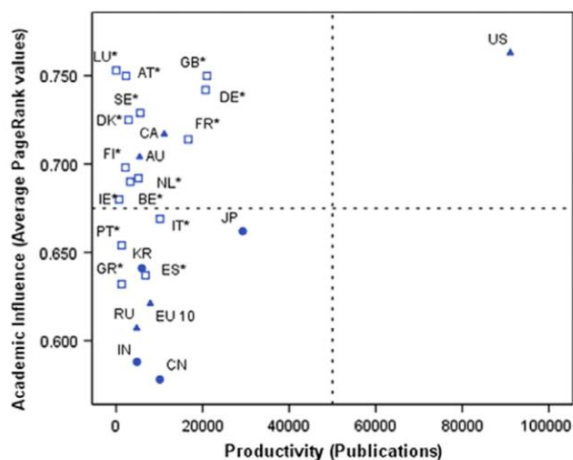


Guan et al. 2008. "Bringing Page-Rank to the Citation Analysis"



35





Ứng dụng: Phân tích trích dẫn (tiếp)



36

1.6 Thuật toán HITS

- Hypertext Induced Topic Search
- J. Kleinberg. "Authoritative Sources in a Hyperlinked Environment." In Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms (SODA'98), pp. 668–677, 1998.

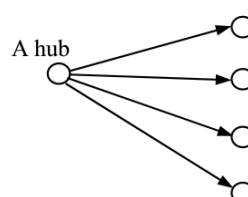
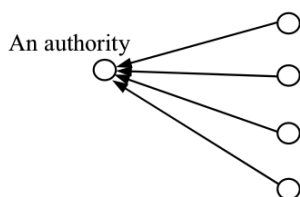
	Spam filtering	Query relevance	Execution
HIST			Online
PageRank			Offline



37

Authority/Hub

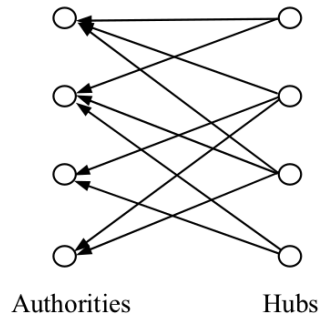
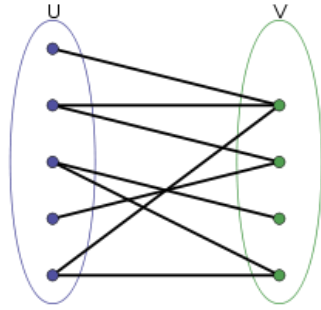
- Authority: Trang được trỏ tới nhiều
- Hub: Trang trỏ tới nhiều trang khác
- Authority và hub có mối quan hệ tương hỗ



38

Bigraph

- Các nút chia thành hai tập không giao nhau
- Mỗi cạnh đều nối hai nút thuộc hai tập



39

Thuật toán

- Đầu vào: Câu truy vấn q
- Đầu ra: Điểm authority và hub của các trang liên quan đến q
- Thuật toán:
 - 1 - Truy hồi thông tin
 - 2 - Mở rộng đồ thị
 - 3 - Tính ranking



40

1-Truy hồi thông tin

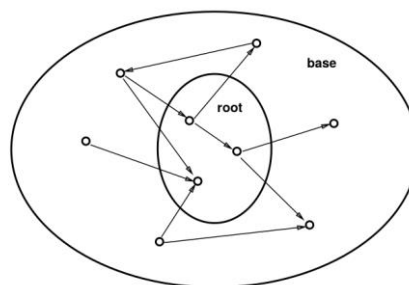
- Y/c một máy tìm kiếm có chứa các văn bản liên quan đến câu truy vấn q (vd Google, Coccoc)
 - Đưa q vào máy tìm kiếm và lấy về tập $root$ W gồm k trang liên quan nhất đến q (vd $k = 200$)



41

2- Mở rộng đồ thị

- Từ tập $root$ W , mở rộng ra tập $base$ S
- Với mỗi trang p trong W
 - Bổ sung các trang mà p trỏ tới
 - Bổ sung các trang trỏ tới p



42

3- Tính thứ hạng

Authority score (a)

Hub score (h)

$$G = (V, E)$$

$$L_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

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

$$\sum_{i=1}^n a(i) = 1$$

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

$$\sum_{i=1}^n h(i) = 1$$



43

3- Tính thứ hạng (tiếp)

$$a = L^T h$$

$$h = La$$

HITS-Iterate(G)

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

$k \leftarrow 1$

Repeat

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

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

$a_k \leftarrow a_k / \|a_k\|_1; \quad // \text{normalization}$

$h_k \leftarrow h_k / \|h_k\|_1; \quad // \text{normalization}$

$k \leftarrow k + 1;$

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

return a_k and h_k



44



HUST

 hust.edu.vn  fb.com/dhbkhn

THANK YOU !

45