

# Tìm kiếm thông tin

## Chương 11. Phân tích liên kết

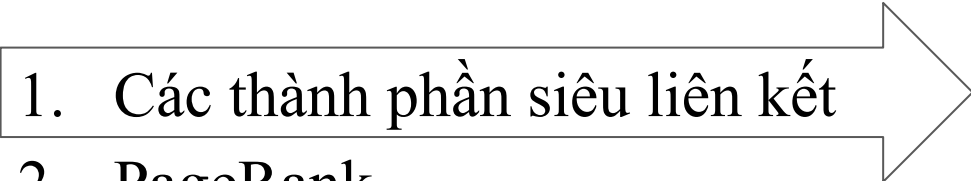
*Soạn bởi: TS. Nguyễn Bá Ngọc*

2021

# Nội dung

1. Các thành phần siêu liên kết
2. PageRank
3. Hub, Authority và giải thuật HITS

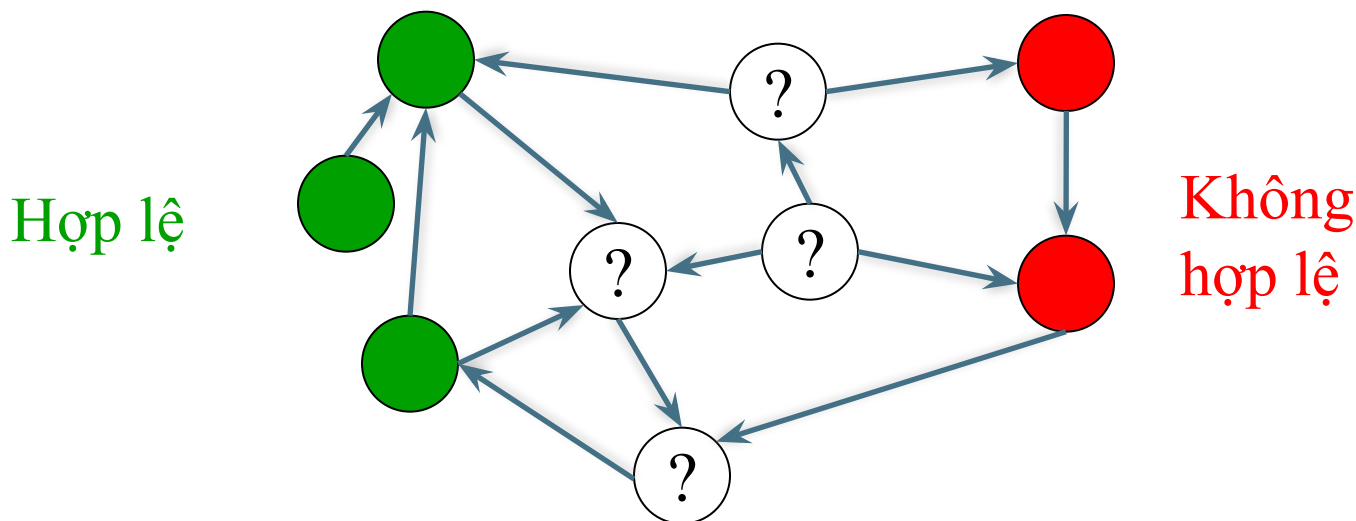
# Nội dung

- 
1. Các thành phần siêu liên kết
  2. PageRank
  3. Hub, Authority và giải thuật HITS

# Siêu liên kết trong môi trường Web

- Số lượng siêu liên kết trong môi trường Web rất lớn
- Là nguồn dữ liệu đáng tin cậy để xác thực và xếp hạng
  - Phân loại email - Chặn email rác & địa chỉ được sử dụng để gửi email rác.
  - Phân loại trang Web - Loại bỏ các trang chứa nội dung độc/không hợp lệ?

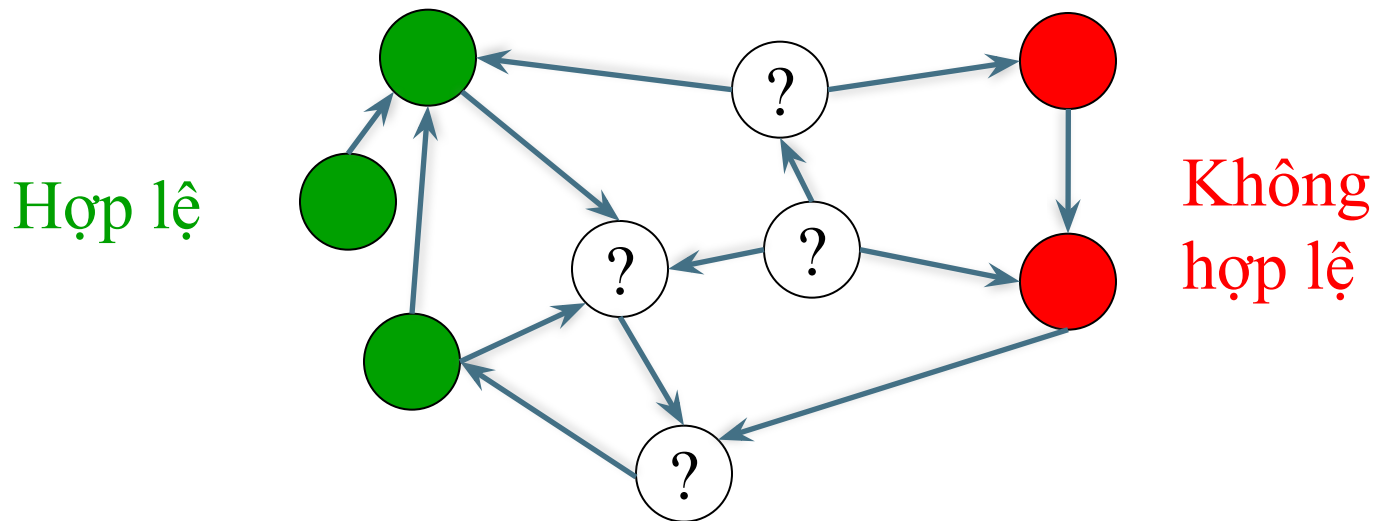
Nguồn hợp lệ, nguồn không hợp lệ và chưa biết



# Ví dụ 11.1. Phân loại nguồn tin

*Nguồn hợp lệ, không hợp lệ và chưa xác định*

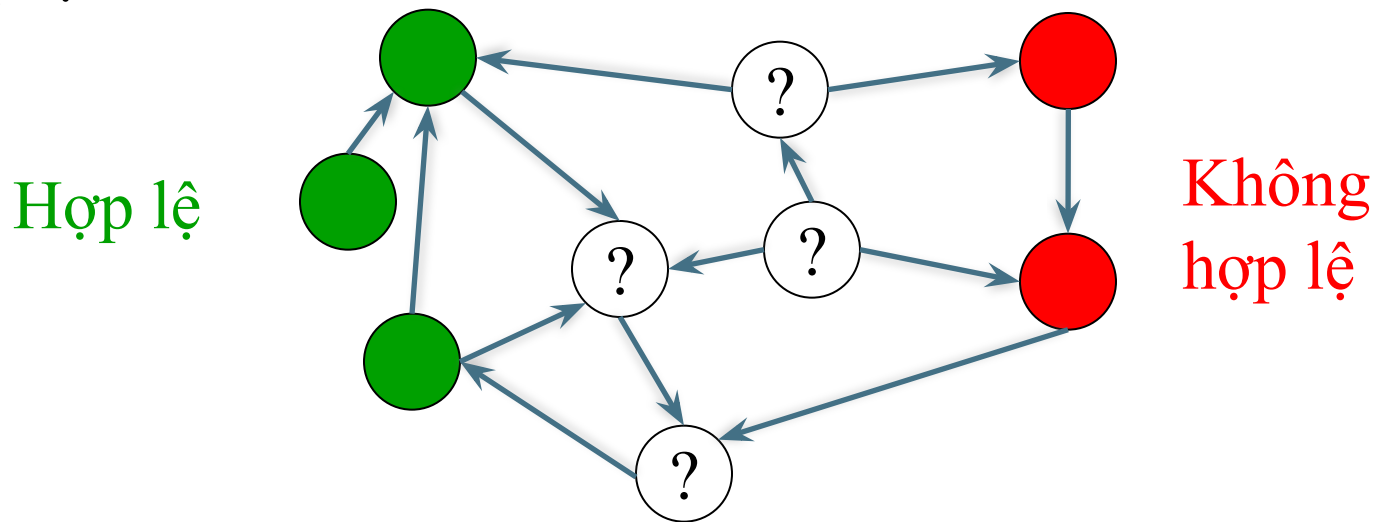
- Trang hợp lệ không chứa liên kết đến trang không hợp lệ
- Các tổ hợp còn lại đều có thể



# Ví dụ 11.1. Phân loại nguồn tin<sup>(2)</sup>

Lô-gic đơn giản: Các nút hợp lệ không chứa liên kết tới các nút không hợp lệ

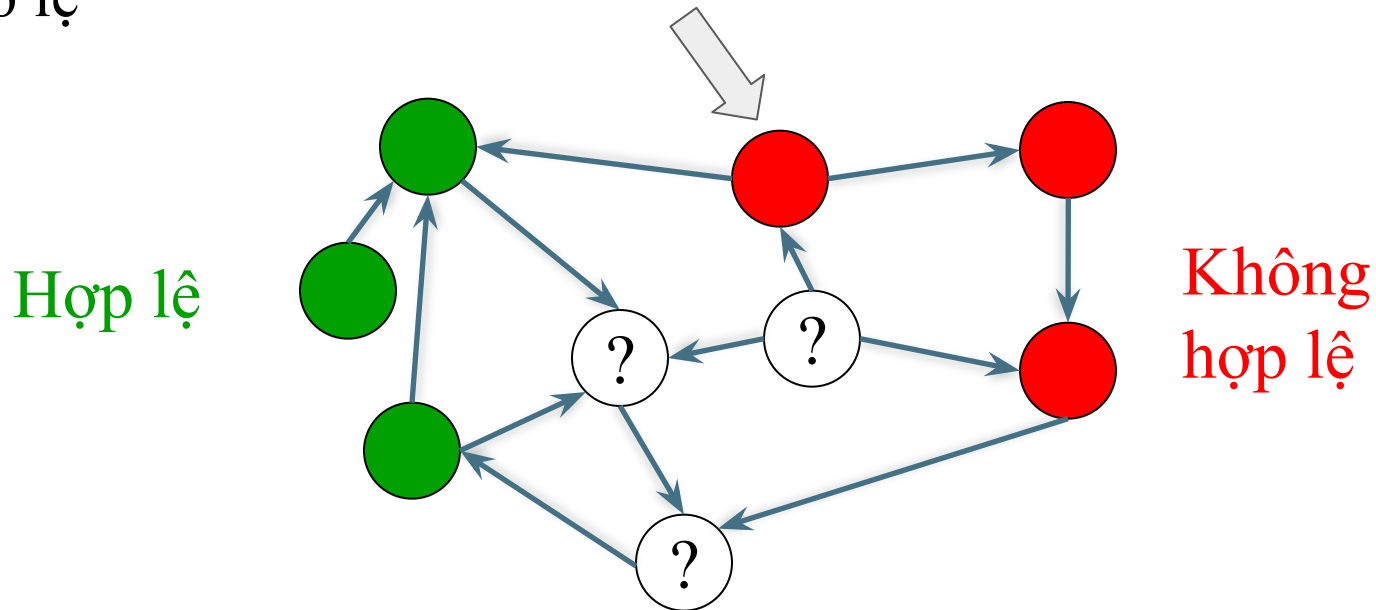
- Nếu 1 trang chứa liên kết tới trang không hợp lệ, thì đó là trang không hợp lệ
- Nếu 1 trang hợp lệ chứa liên kết tới trang khác, thì đó là trang hợp lệ



# Ví dụ 11.1. Phân loại nguồn tin<sup>(3)</sup>

Lô-gic đơn giản: Các nút hợp lệ không chứa liên kết tới các nút không hợp lệ

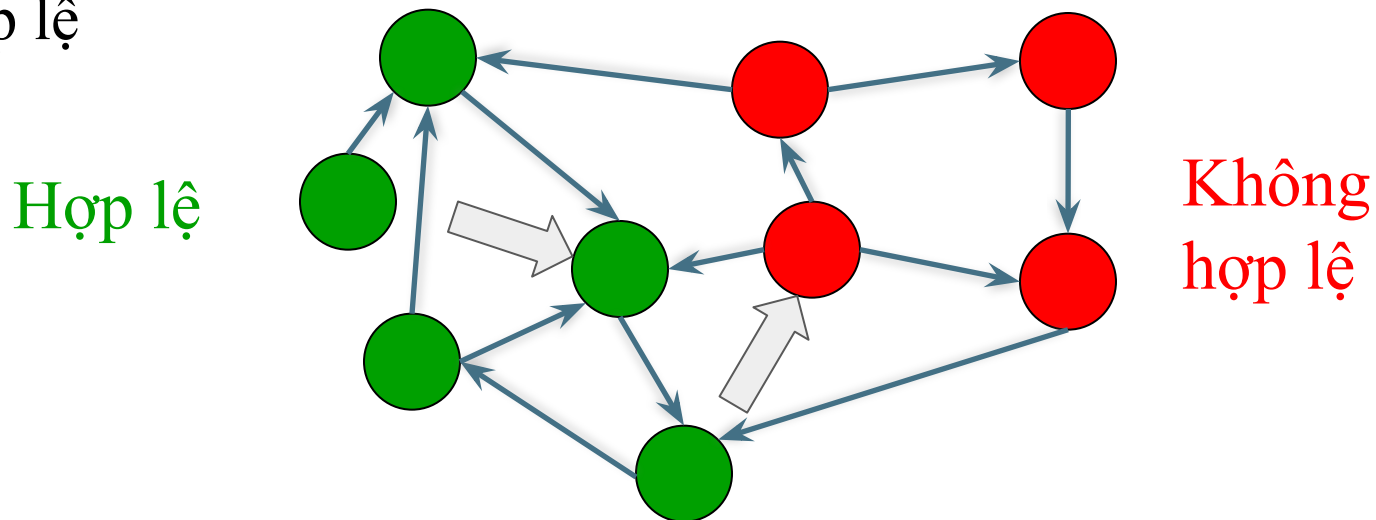
- Nếu 1 trang chứa liên kết tới trang không hợp lệ, thì đó là trang không hợp lệ
- Nếu 1 trang hợp lệ chứa liên kết tới trang khác, thì đó là trang hợp lệ



# Ví dụ 11.1. Phân loại nguồn tin<sup>(4)</sup>

Lô-gic đơn giản: Các nút hợp lệ không chứa liên kết tới các nút không hợp lệ

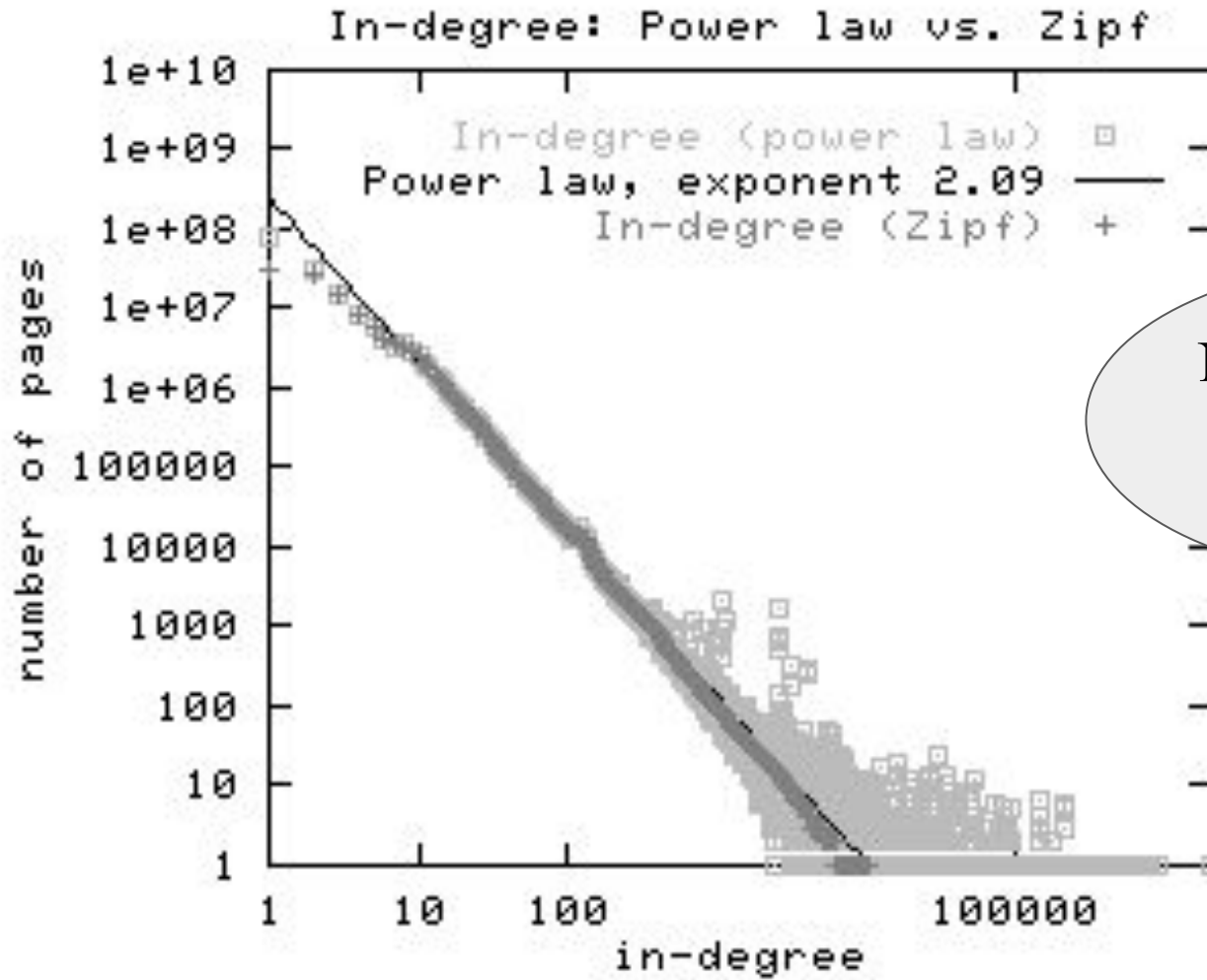
- Nếu 1 trang chứa liên kết tới trang không hợp lệ, thì đó là trang không hợp lệ
- Nếu 1 trang hợp lệ chứa liên kết tới trang khác, thì đó là trang hợp lệ



*Bên cạnh đó cũng có thể sử dụng phương pháp phân lớp dựa trên xác suất [Chương 8]*



# Ví dụ 11.2. Bậc vào bất thường



Nội dung rác gây  
sai lệch luật lũy  
thừa

# Các ví dụ khác về phân tích liên kết

- Mạng xã hội là một nguồn giàu thông tin về hành vi nhóm
  - Ví dụ: Sự tương đồng của khách hàng - Goel+Goldstein 2010
  - Các khách hàng có bạn mua sắm nhiều cũng thường mua sắm

<http://www.cs.cornell.edu/home/kleinber/networks-book/>

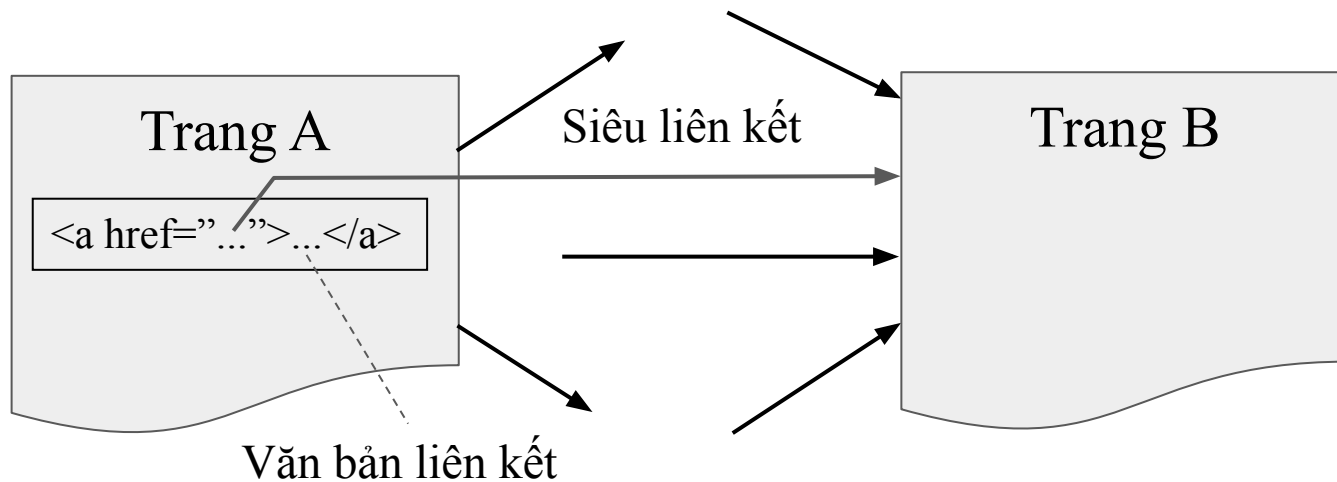
# Mối quan tâm chính trong TKTT

- Phân tích liên kết để tìm kiếm
  - Tính điểm và xếp hạng
  - Phân cụm dựa trên liên kết - cấu trúc chủ đề từ các liên kết
  - Các liên kết như các đặc trưng trong phân lớp - các văn bản có liên kết với nhau có xu hướng thuộc cùng 1 chủ đề.
- Thu thập dữ liệu
  - Dựa trên các liên kết đã thấy, tiếp theo chúng ta sẽ thu thập những gì?

# Các giả thuyết liên kết

**Giả thuyết 1:** Siêu liên kết thể hiện sự đánh giá tích cực của trang chứa liên kết đối với trang được trỏ tới (tín hiệu chất lượng của trang được trỏ tới).

**Giả thuyết 2:** Văn bản liên kết (nằm trong thẻ `<a>`) có nội dung mô tả liên kết.



# Ví dụ 11.3. Siêu liên kết

## Previous Editions

- [The C++ Programming Language \(Special Edition\)](#)
- [Book list](#)
- [My book covers](#)

## Reviews

*Giả thuyết 1*

I don't have to agree with a review or blog post to list it, but it helps if I think at least some parts make sense.

- June 17, 2012: Verity Stob: [Software >Bjarne Again: Hallelujah for C++](#) in The Register.
- June 17, 2013: The Meglomaniac Bore [Blog entry](#). ``you can learn something new nearly every other page - even as a seasoned C++ developer I was still updating my knowledge."
- June 8, 2013: Peter Lee: [The C++ Programming Language \(4th edition\): Bjarne Stroustrup, Thou Art The Man!](#).
- [Amazon](#). Please ignore the stars and focus on the comments that convey information. Note that several comments were made before [hardcover](#) and [electronic](#) versions were available.

## Interviews

*Giả thuyết 2*

- [An interview by Danny Kalev for InformIT.](#)

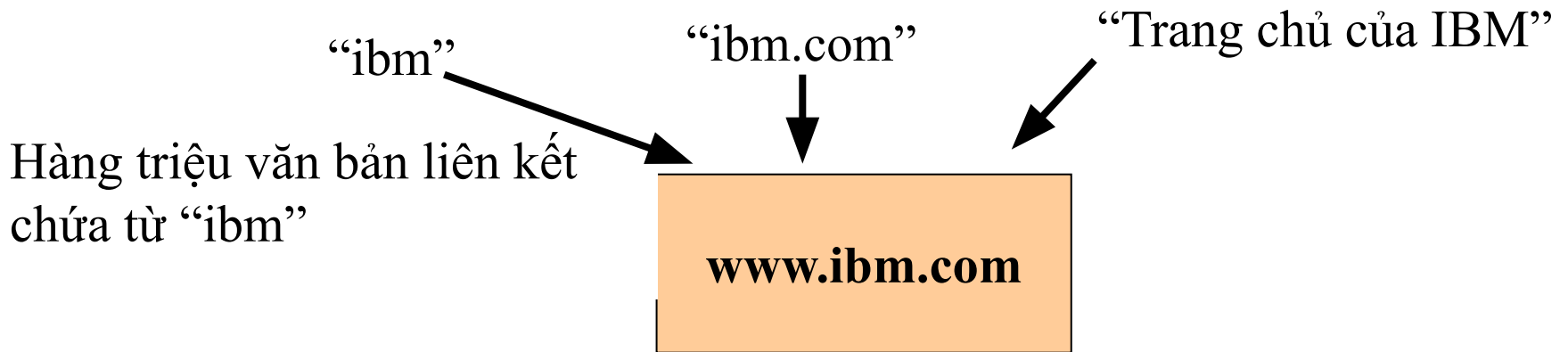
## Translations

I have no idea what the rate of progress is on translations:

- [amazon](#). Naturally, a variety of people comment. The comments vary in their level of sophistication, degree of insight, and level of professionalism. For the many insightful and polite comments: Thanks! It is not possible for an author to comment on every mistake, misunderstanding, and misrepresentation in these ``reviews." Instead, here are a few responses to comments that I happen to disagree with or find potentially misleading:
  - **The book is too big.** I agree, but *some* book must cover essentially all of C++, and this is that book. If you can make do with less, please do, but I feel obliged to aim for completeness. This book is close to complete from a programmer's point of view. Language lawyers need the standard, but this book is a good place to start even for those. Like the standard, this book covers the C++ language and the ISO C++ standard library.

# Văn bản liên kết

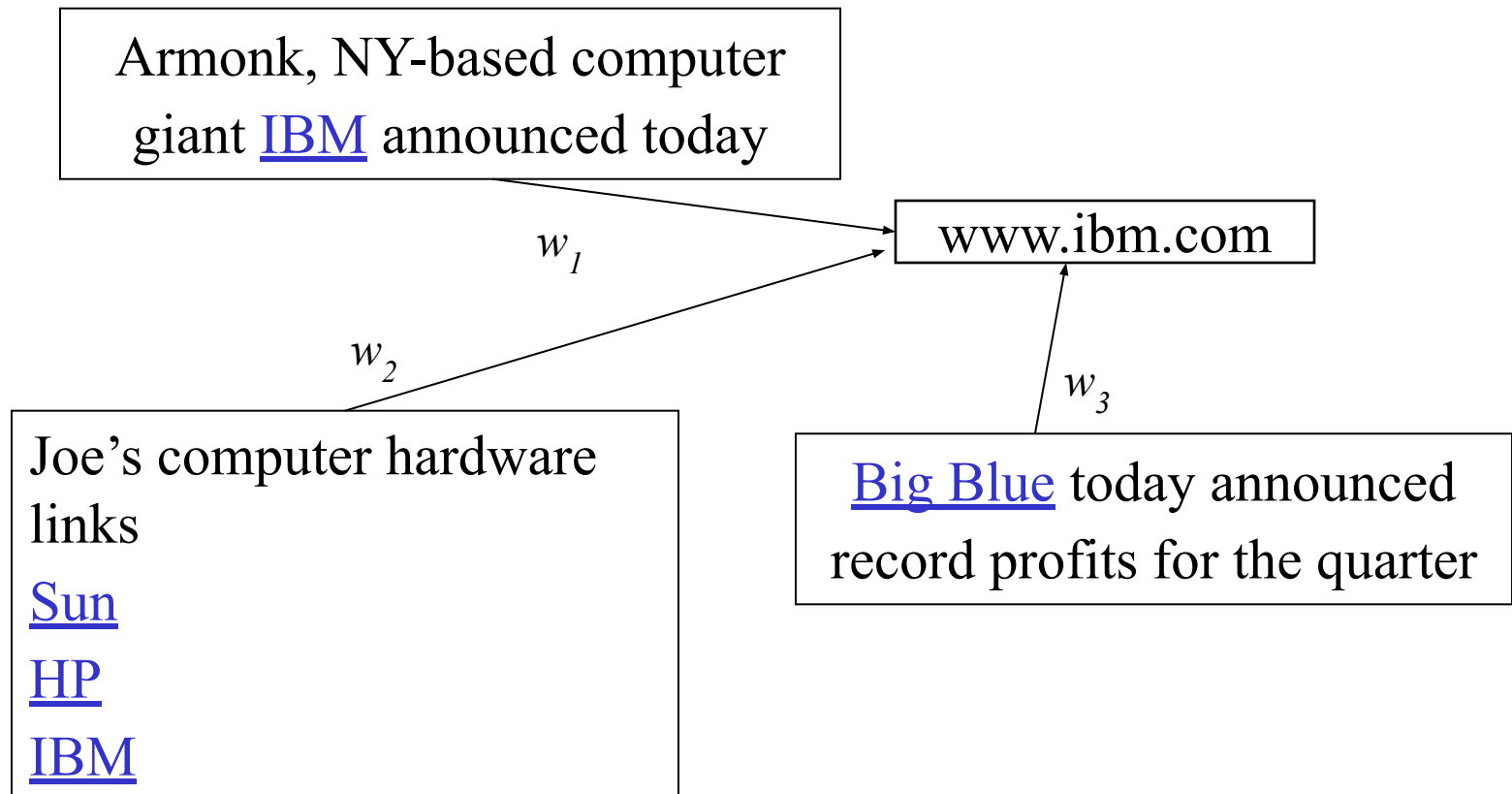
- Trang [www.ibm.com](http://www.ibm.com) có nội dung đa phần là hình ảnh, rất ít từ ibm.
- Với từ ibm làm sao để phân biệt giữa:
  - Trang chủ của IBM
  - Trang thông tin bản quyền của IBM (nhiều từ ibm)
  - Trang tin rác của đối thủ (tần suất từ cao tùy ý)



Tìm kiếm trên [nội dung] + [văn bản liên kết] sẽ hiệu quả hơn nếu chỉ tìm kiếm trên [nội dung]

# Đánh chỉ mục trang Web

- Các nội dung được đánh chỉ mục cho 1 trang Web  $d$  bao gồm nội dung có trong  $d$  và các nội dung văn bản mô tả liên kết (có trọng số) trỏ tới  $d$ .




# Đánh chỉ mục trang Web<sub>(2)</sub>

- Bổ xung các văn bản liên kết đôi khi có thể gây ra các hiệu ứng không mong muốn, ví dụ, sai lệch do nội dung rác, v.v..
- Có thể tính điểm văn bản liên kết với trọng số dựa trên uy tín của trang chứa liên kết
  - Ví dụ, nếu chúng ta cho rằng liên kết từ wikipedia.org là uy tín, thì có thể đặt trọng số cao hơn (tin tưởng hơn) cho các văn bản liên kết từ những nguồn đó.



# Nội dung

1. Các thành phần siêu liên kết
  2. PageRank
  3. Hub, Authority và giải thuật HITS
- 

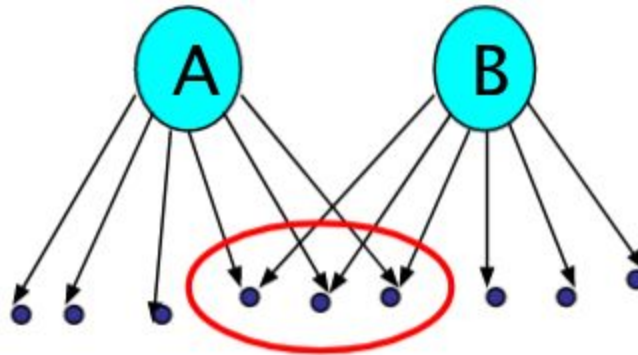
# Cơ sở của PageRank

- Các trích dẫn giữa những ấn phẩm in được phân tích trước khi có Web.
- Trong các ấn phẩm in: sách, báo, tạp chí v.v.
  - Một tài liệu có thể trích dẫn các tài liệu khác (tài liệu tham khảo)
    - Thể hiện sự gắn kết về nội dung
- Một số ứng dụng của dữ liệu trích dẫn:
  - Xác định độ tương đồng giữa các tài liệu
  - Xếp hạng tạp chí
  - v.v.

*Siêu liên kết trong môi trường Web một phần có ý nghĩa tương tự trích dẫn trong ấn phẩm in*

# Độ tương đồng dựa trên trích dẫn

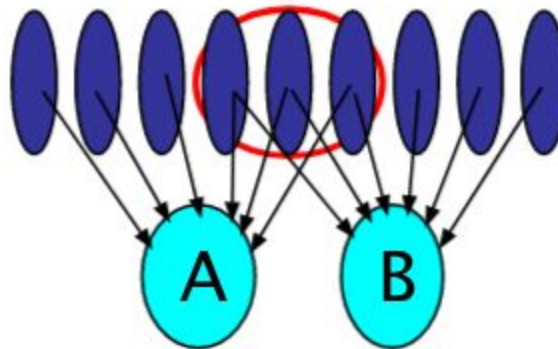
- Kessler (1963), đánh giá độ tương đồng giữa 2 tài liệu dựa trên số lượng tài liệu tham khảo mà cả 2 tài liệu cùng trích dẫn - Bibliographic coupling



*Có nên chuẩn hóa theo tổng #tài liệu tham khảo trong A + #tài liệu tham khảo trong B hay không?*

# Độ tương đồng dựa trên dữ liệu trích dẫn<sub>(2)</sub>

- Small (1973), đánh giá độ tương đồng giữa 2 tài liệu thông qua số lượng tài liệu cùng trích dẫn cả 2 tài liệu đó - Cocitation.



*Có nên chuẩn hóa theo tổng #tài liệu trích dẫn A + #tài liệu trích dẫn B hay không?*

# Chỉ số tầm ảnh hưởng

- Impact factor, Garfield, 1972.
- Được tính hàng năm bởi ISI (*Institute for Scientific Information*)
- Được sử dụng để xếp hạng các tạp chí.
- Chỉ số ảnh hưởng của 1 tạp chí J trong năm Y bằng số lượng trích dẫn trung bình từ các bài viết trong năm Y tới các bài viết của tạp chí J trong năm Y - 1 hoặc Y - 2.
  - Không phân biệt chất lượng bài viết.

# Xếp hạng dựa trên trích dẫn

- Pinski và Narin, 1976, xếp hạng các bài báo dựa trên phân tích trích dẫn

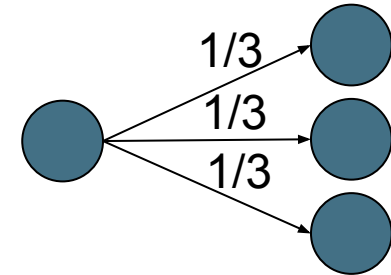
*PageRank kế thừa phương pháp này.*

# Liên kết Web vs. trích dẫn tài liệu

- Môi trường Web:
  - Hàng tỉ người tham gia với các sở thích riêng;
  - *Bất kỳ ai cũng có thể đưa nội dung lên Web;*
  - *(Nội dung có thể chưa được kiểm duyệt)*
  - Nhiều nội dung rác.
- Từ khi máy tìm kiếm bắt đầu sử dụng các siêu liên kết để xếp hạng (khoảng 1998), các *cộng đồng liên kết* cũng bắt đầu xuất hiện nhằm thu lợi từ xếp hạng của máy tìm kiếm
  - Một nhóm trang Web chứa nhiều liên kết chéo nhau nhằm chiếm thứ hạng cao trong xếp hạng của máy tìm kiếm Web.
  - *(Một người có thể gia nhập cộng đồng liên kết để tăng số lượng liên kết tới trang của mình...)*

# Khái niệm Pagerank

Mô hình duyệt Web ngẫu nhiên

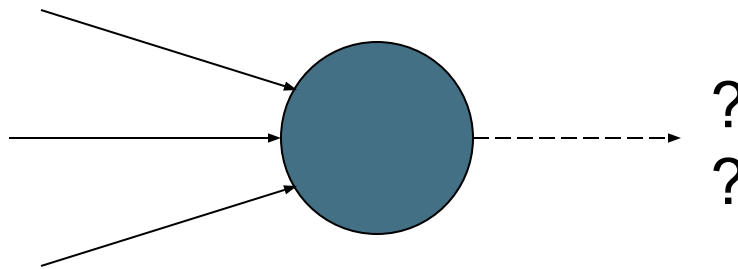


- Giả sử 1 người dùng thực hiện di chuyển ngẫu nhiên trên Web:
  - Bắt đầu với 1 trang ngẫu nhiên
  - Sau mỗi bước, với xác suất đồng nhất chọn ngẫu nhiên 1 liên kết có trong trang hiện tại.
- Sau 1 số bước đủ lớn mỗi trang có 1 tỉ lệ được mở ổn định, không phụ thuộc vào điểm bắt đầu, đại lượng này là PageRank của trang Web và được sử dụng để xếp hạng các trang Web.



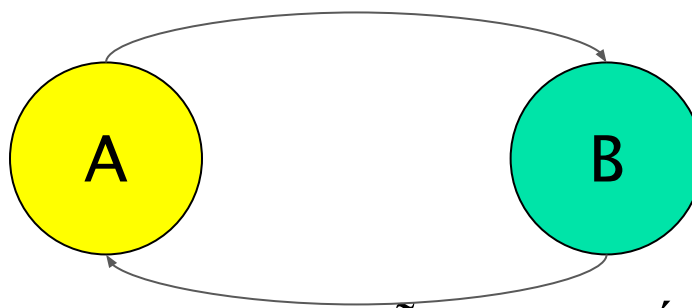
# Hạn chế của duyệt Web ngẫu nhiên

- Web có rất nhiều trang không có liên kết đi ra - trang đóng:
  - Tiến trình duyệt Web ngẫu nhiên có thể dừng ở trang đóng
  - => Không có tỉ lệ mở ổn định với số bước lớn.



# Hạn chế của duyệt Web ngẫu nhiên<sub>(2)</sub>

- Tiến trình duyệt Web ngẫu nhiên có thể diễn ra theo chu trình tuần tự khép kín
  - Xét đồ thị Web đơn giản như sau:



- Giả sử tiến trình duyệt Web ngẫu nhiên bắt đầu từ A. Sau 1 số chẵn bước di chuyển thì tỉ lệ ghé thăm A và B là cân bằng. Sau 1 số lẻ bước thì tỉ lệ ghé thăm A cao hơn B.
- $\Rightarrow$  Không có tỉ lệ ổn định trong trường hợp này.

# Bổ xung bước nhảy

Quy tắc di chuyển theo bước nhảy:

- Nếu đang ở trang đóng thì chuyển tới 1 trang bất kỳ được lựa chọn ngẫu nhiên;
- Nếu ngược lại (không ở trang đóng), thì với xác suất  $\alpha$  (ví dụ  $\alpha = 10\%$ ) nhảy tới một trang ngẫu nhiên,
  - và với xác suất còn lại ( $1 - \alpha = 90\%$ ), di chuyển theo 1 liên kết được chọn ngẫu nhiên
  - Tham số  $\alpha$  được gọi là tỉ lệ nhảy ngẫu nhiên.

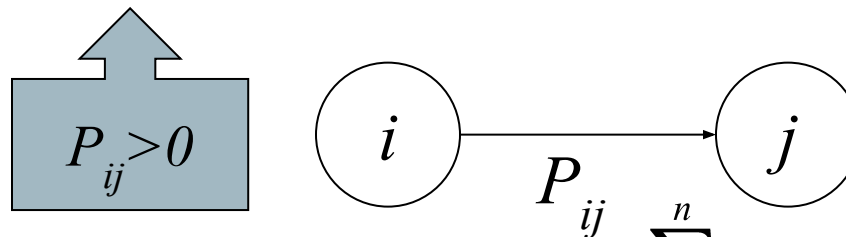
# Tác dụng của bước nhảy

- Tiến trình di chuyển ngẫu nhiên với bước nhảy không thể bị mắc lại ở bất kỳ trạng nào;
- và không thể diễn ra theo chu trình tuần tự khép kín;
- $\Rightarrow$  Vì vậy luôn tồn tại tỉ lệ được mở xem ổn định với số bước lớn.

*Tiếp theo chúng ta sẽ phân tích chi tiết hơn lý do vì sao?*

# Chuỗi Markov

- Một chuỗi Markov bao gồm  $n$  trạng thái, cùng với 1 ma trận xác suất chuyển trạng thái kích thước  $n \times n$  (ma trận  $P$ )
- Ở 1 bước bất kỳ chuỗi Markov luôn có 1 trạng thái xác định, và sau mỗi bước chuỗi Markov có thể chuyển sang 1 trạng thái khác.
- Với  $1 \leq i, j \leq n$ , phần tử  $P_{ij}$  của ma trận cho biết xác suất trạng thái  $j$  là trạng thái tiếp theo, nếu  $i$  đang là trạng thái hiện tại.



- Dễ dàng nhận thấy, với  $i$  bất kỳ, thì  $\sum_{j=1}^n P_{ij} = 1$ .
- Có thể biểu diễn tiến trình duyệt Web bằng chuỗi Markov

# Chuỗi Markov Ergodic

- Nếu chuỗi Markov thỏa mãn điều kiện ergodic, thì luôn tồn tại 1 phân bố xác suất duy nhất cho các trạng thái sau số bước chuyển đủ lớn:
  - Sau 1 số bước chuyển đủ lớn, xác suất 1 trạng thái đang là trạng thái của chuỗi bằng tỉ lệ số lần chuỗi đã chuyển tới trạng thái đó.
  - Các giá trị xác suất không phụ thuộc vào trạng thái bắt đầu mà chỉ phụ thuộc vào  $P$ .

# Điều kiện tồn tại PageRank

- Tồn tại PageRank cho tiến trình duyệt trên đồ thị Web nếu thỏa mãn 2 điều kiện sau:
  - **Điều kiện 1:** Có thể di chuyển từ 1 trang bất kỳ tới bất kỳ trang nào trong đồ thị.
  - **Điều kiện 2:** Không thể chia tập đỉnh thành nhiều tập con sao cho tiến trình duyệt Web ngẫu nhiên diễn ra theo trình tự tuần tự và khép kín trong các tập con đó.
  - *(Nếu các điều kiện được đáp ứng thì chuỗi Markov tương ứng với đồ thị Web cũng thỏa mãn điều kiện Ergodic).*

*Tính PageRank bằng cách nào?*

# Tính ma trận xác suất chuyển trạng thái

- Đặt  $A$  là ma trận kề của đồ thị, phần tử ở dòng  $i$  cột  $j$ ,  $A_{ij} = 1$  nếu có cạnh từ  $i$  tới  $j$ ;  $A_{ij} = 0$  nếu ngược lại; đặt  $\alpha$  là xác suất nhảy ngẫu nhiên.
- Có thể tính ma trận  $P$  từ ma trận  $A$  qua 4 bước sau:
  - 1) Với những hàng toàn 0 trong  $A$ : Thay 0 bằng  $1/N$ , trong đó  $N$  là số lượng trang Web/kích thước ma trận vuông  $A$ .
  - 2) Với những hàng còn lại: Thay 1 bằng  $1/\text{số } 1 \text{ trong hàng đó}$ .
  - 3) Nhân ma trận kết quả thu được sau các bước 1 và 2 với  $1 - \alpha$ .
  - 4) Cộng  $\alpha/N$  vào từng phần tử của ma trận thu được sau bước 3.



# Các vec-tơ xác suất

- Đặt  $x = (x_1, \dots, x_n)$  (định dạng dòng) là vec-tơ phân bố xác suất trạng thái của tiến trình ngẫu nhiên.
- Ví dụ,  $(000\dots 1\dots 000)$  nghĩa là tiến trình đang ở trạng thái  $i$   
 $1 \dots i \dots n$
- Các giá trị  $x_i$  cho biết xác suất tiến trình ngẫu nhiên đang ở trạng thái  $i$ .

$$\sum_{i=1}^n x_i = 1.$$

# Cập nhật các vec-tơ xác suất

- Nếu vec-tơ xác suất hiện đang là  $x = (x_1, \dots, x_n)$ , thì phân bố xác suất ở bước tiếp theo là gì?
- Như đã biết dòng  $i$  của ma trận  $P$  cho chúng ta biết xác suất chuyển tới các trạng thái từ  $i$ .
- Như vậy với  $x$  là phân bố xác suất hiện tại thì phân bố xác suất ở bước tiếp theo là  $xP$ 
  - Kế tiếp là  $xP^2$ , và sau đó là  $xP^3$ , v.v..
  - Dãy sẽ hội tụ? Nếu có thì giá trị ở trạng thái hội tụ là gì?

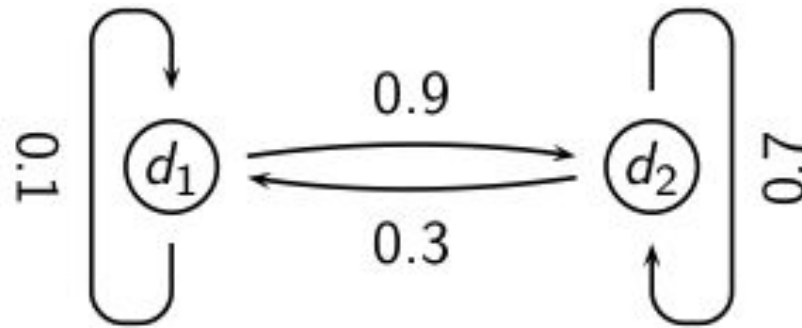
*Phương pháp tìm PageRank bằng cách lặp nhân  $x$  với  $P$  cho tới khi  $x$  không thay đổi được gọi là phương pháp lũy thừa.*

# Ý nghĩa đại số của PageRank

- Đặt  $a = (a_1, \dots, a_n)$  là vec-tơ định dạng dòng của phân bố xác suất ổn định.
- Với  $a$  là phân bố ổn định, chúng ta có  $a = aP$ 
  - Như vậy  $a$  là nghiệm của hệ phương trình.
  - Đồng thời  $a$  là vec-tơ riêng (trái) của  $P$ .
    - (tương ứng với vec-tơ riêng chính của  $P$  với giá trị riêng lớn nhất)
    - Ma trận xác suất chuyển trạng thái luôn có giá trị riêng lớn nhất  $= 1$ .

## Ví dụ 11.4. Phương pháp lũy thừa

Cho 2 văn bản  $d_1$ ,  $d_2$  và các xác suất di chuyển như sau:



**Yêu cầu:** Tính các giá trị PageRank.

## Ví dụ 11.4. Phương pháp lũy thừa<sub>(2)</sub>

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= xP$
$t_1$	0.3	0.7	0.24	0.76	$= xP^2$
$t_2$	0.24	0.76	0.252	0.748	$= xP^3$
$t_3$	0.252	0.748	0.2496	0.7504	$= xP^4$
			...		...
$t_\infty$	0.25	0.75	0.25	0.75	$= xP^\infty$

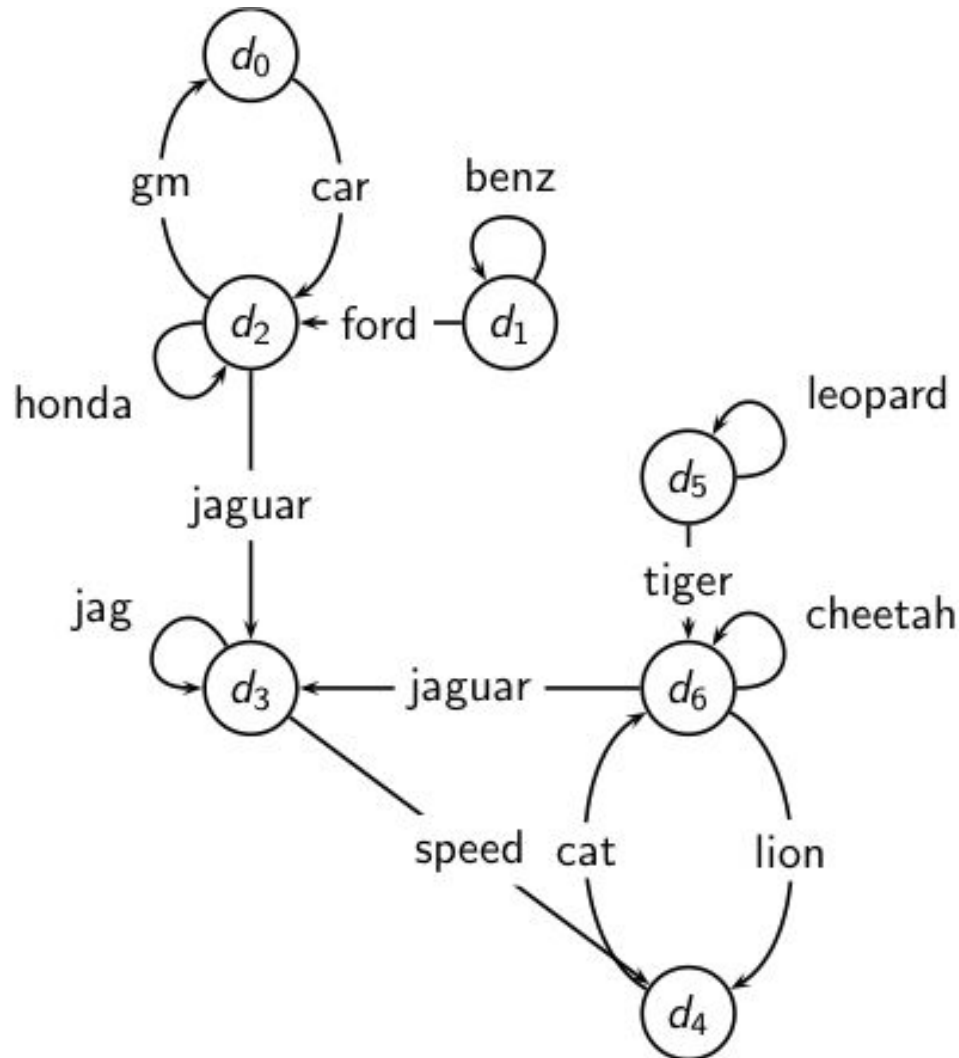
$$\vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Ví dụ 11.5. Tính P và PageRank

Cho đồ thị Web



# Ví dụ 11.5. Tính P và PageRank<sub>(2)</sub>

Với mô hình duyệt Web ngẫu nhiên

	$d_0$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
$d_0$	0.00	0.00	1.00	0.00	0.00	0.00	0.00
$d_1$	0.00	0.50	0.50	0.00	0.00	0.00	0.00
$d_2$	0.33	0.00	0.33	0.33	0.00	0.00	0.00
$d_3$	0.00	0.00	0.00	0.50	0.50	0.00	0.00
$d_4$	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$d_5$	0.00	0.00	0.00	0.00	0.00	0.50	0.50
$d_6$	0.00	0.00	0.00	0.33	0.33	0.00	0.33

# Ví dụ 11.5. Tính P và PageRank<sub>(3)</sub>

*Nếu sử dụng bước nhảy*

	$d_0$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
$d_0$	0.02	0.02	0.88	0.02	0.02	0.02	0.02
$d_1$	0.02	0.45	0.45	0.02	0.02	0.02	0.02
$d_2$	0.31	0.02	0.31	0.31	0.02	0.02	0.02
$d_3$	0.02	0.02	0.02	0.45	0.45	0.02	0.02
$d_4$	0.02	0.02	0.02	0.02	0.02	0.02	0.88
$d_5$	0.02	0.02	0.02	0.02	0.02	0.45	0.45
$d_6$	0.02	0.02	0.02	0.31	0.31	0.02	0.31



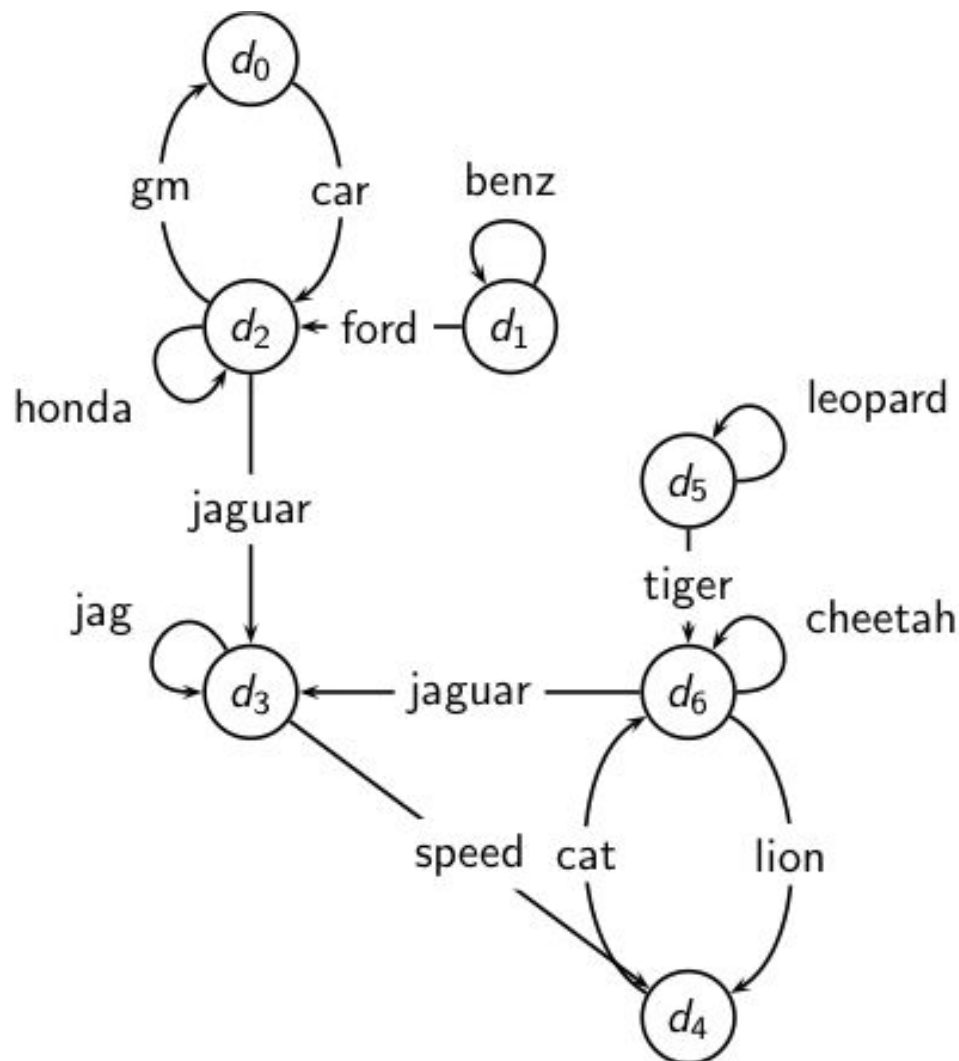
# Ví dụ 11.5. Tính P và PageRank<sub>(4)</sub>

*Các lũy thừa*

	$\vec{X}$	$\vec{X}P^1$	$\vec{X}P^2$	$\vec{X}P^3$	$\vec{X}P^4$	$\vec{X}P^5$	$\vec{X}P^6$	$\vec{X}P^7$	$\vec{X}P^8$	$\vec{X}P^9$	$\vec{X}P^{10}$	$\vec{X}P^{11}$	$\vec{X}P^{12}$	$\vec{X}P^{13}$
$d_0$	0.14	0.06	0.09	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
$d_1$	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
$d_2$	0.14	0.25	0.18	0.17	0.15	0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11
$d_3$	0.14	0.16	0.23	0.24	0.24	0.24	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25
$d_4$	0.14	0.12	0.16	0.19	0.19	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
$d_5$	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
$d_6$	0.14	0.25	0.23	0.25	0.27	0.28	0.29	0.29	0.30	0.30	0.30	0.30	0.31	0.31

# Ví dụ 11.5. Tính P và PageRank<sub>(5)</sub>

*Các giá trị PageRank*




	PageRank
$d_0$	0.05
$d_1$	0.04
$d_2$	0.11
$d_3$	0.25
$d_4$	0.21
$d_5$	0.04
$d_6$	0.31

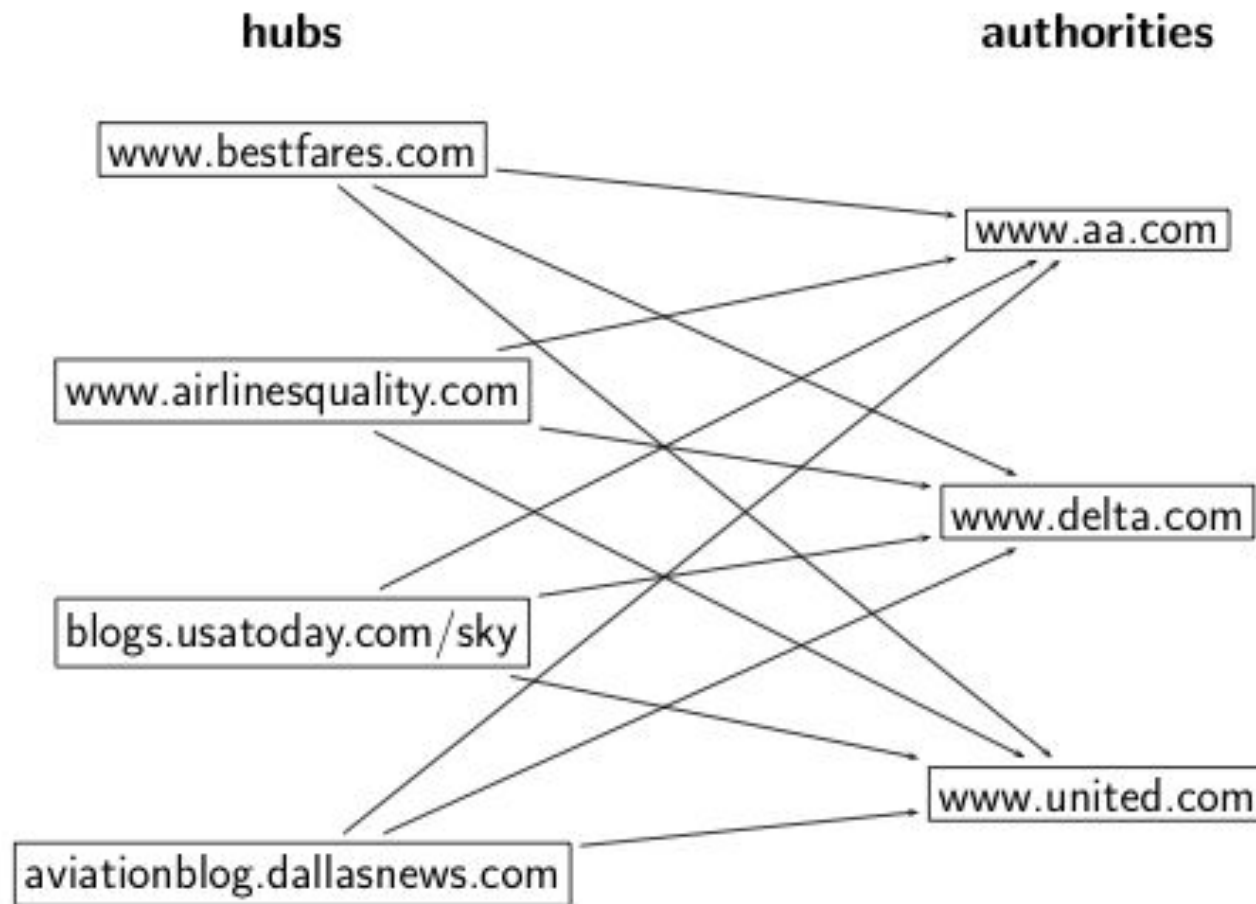
# Duyệt Web trong thực tế

- Trong thực tế người dùng không duyệt Web theo cách ngẫu nhiên:
  - Người dùng có thể sử dụng nút Back, danh sách trang yêu thích của trình duyệt, công cụ tìm kiếm, v.v..
  - Mô hình chuỗi Markov không mô phỏng hết được các tình huống thực tế
- Nếu chỉ sử dụng PageRank thì kết quả tìm kiếm có thể không đủ tốt.

# Nội dung

1. Các thành phần siêu liên kết
  2. PageRank
  3. Hub, Authority và giải thuật HITS
- 

# Khái niệm Hub và Authority



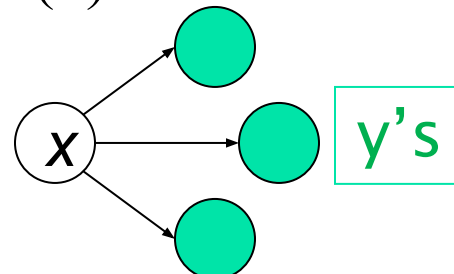
*Hub - chứa nhiều liên kết,  
tính chất bao quát*

*Authority - được nhiều trang  
trở tới, tính chất phổ biến.*

# Tính Hubs và Authorities

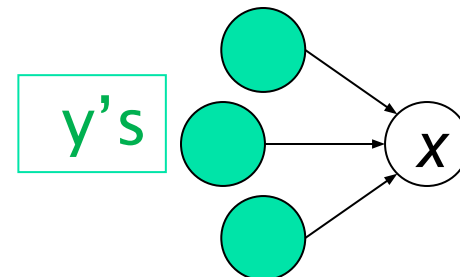
- Khởi tạo: Với mỗi trang  $x$ , đặt  $h(x) = 1$  và  $a(x) = 1$ .
- Cập nhật:

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



- Điểm bao quát của một trang bằng tổng điểm phổ biến của các liên kết có trong nó.

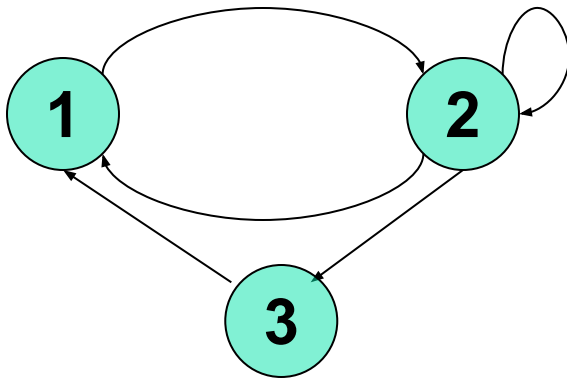
$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$



- Điểm phổ biến của một trang bằng tổng điểm bao quát của các trang trỏ tới nó.
- Có thể chuẩn hóa vec-tơ sau mỗi bước lặp để các giá trị không trở nên quá lớn
  - Không ảnh hưởng tới giá trị tương đối của các phần tử.

# Tính Hubs và Authorities<sub>(2)</sub>

- Đặt  $A$  là ma trận kề kích thước  $N \times N$  (biểu diễn dạng ma trận kề của đồ thị Web):
  - $N$  là số lượng trang đang được phân tích (số lượng đỉnh của đồ thị)
  - $A_{ij} = 1$  nếu tồn tại liên kết  $i \Rightarrow j$  và  $A_{ij} = 0$  nếu ngược lại.



$A =$

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

- Đặt  $\vec{h}$  và  $\vec{a}$  là các vec-tơ Hub và Authority.
- Chúng ta có  $\vec{h} = A * \vec{a}$  và  $\vec{a} = A^t * \vec{h}$

# Tính Hubs và Authorities<sub>(3)</sub>

- Chúng ta có  $\vec{h} = A^* A^t \vec{h}$  và  $\vec{a} = A^t A^* \vec{a}$
- Như vậy  $\vec{h}$  là vec-tơ riêng phải của  $A^* A^t$ , và  $\vec{a}$  là vec-tơ riêng phải của  $A^t A^*$ .
  - Có thể tính các vec-tơ  $\vec{h}$  và  $\vec{a}$  bằng phương pháp lũy thừa.



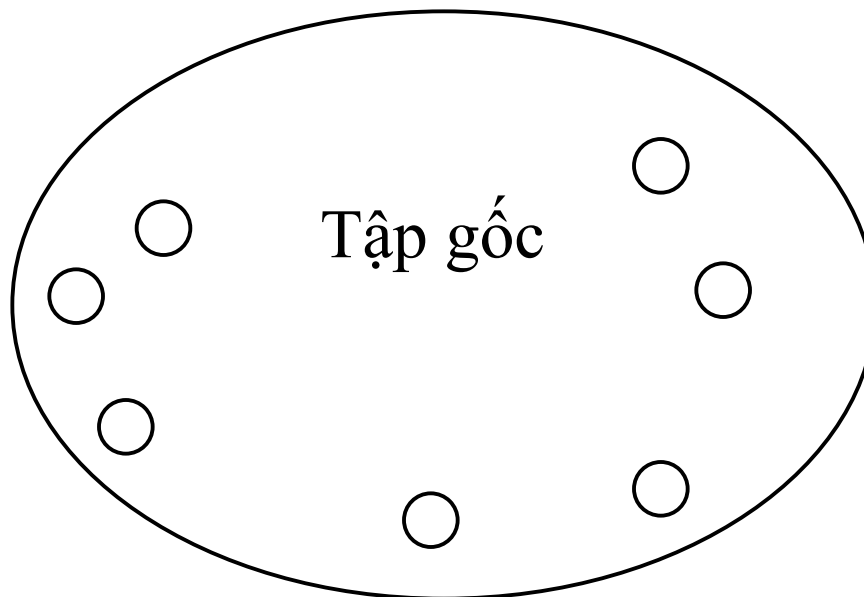
# Giải thuật HITS

- Giải thuật HITS chia tập trang Web (có thể là các kết quả tìm kiếm) thành 2 nhóm:
  - **Nhóm 1 - Hubs.** Các trang có điểm bao quát cao và có thể chứa nhiều liên kết có khả năng đáp ứng nhu cầu thông tin.
  - **Nhóm 2 - Authorities.** Các trang có điểm phổ biến cao, có thể chứa nhiều thông tin hữu ích đáp ứng nhu cầu thông tin.

*Hyperlink-Induced Topic Search (HITS), Klei98*

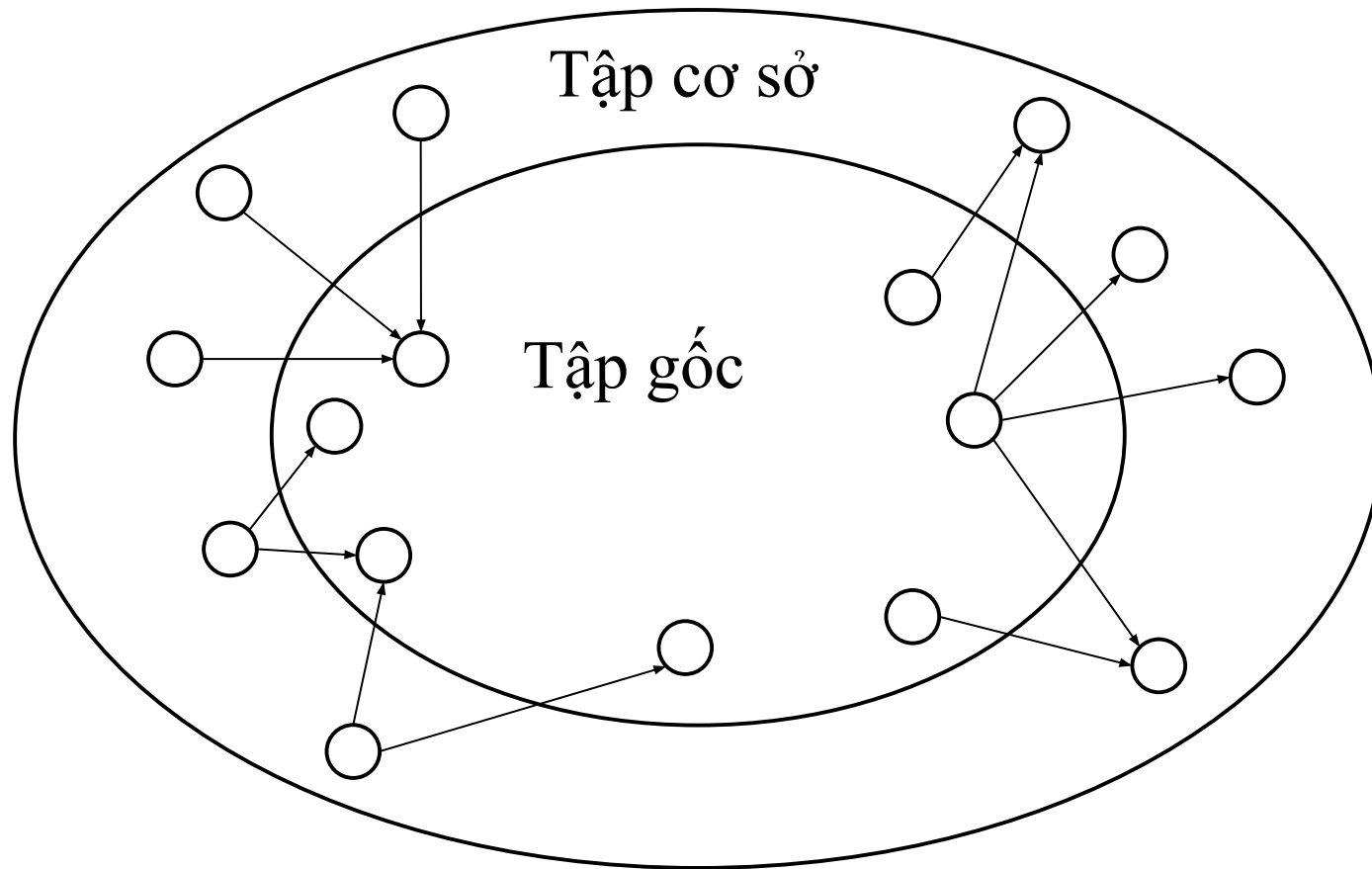
*(Ý tưởng hay tuy nhiên ít gặp trong các hệ thống tìm kiếm hoạt động thực tế).*

# Giải thuật HITS: Tập gốc



Là tập kết quả tìm kiếm thu được sau khi xử lý truy vấn.

# Giải thuật HITS: Tập cơ sở



Tính điểm bao quát và điểm phổ biến trên tập cơ sở (thu được sau khi mở rộng tập gốc)

# Kích thước tập gốc và tập cơ sở

- Tập gốc thường có 200-1000 trang
- Tập cơ sở có thể chứa tới 5000 trang
  - Kích thước không quá lớn, có thể xử lý trong thời gian xử lý truy vấn.

[Klei98]

# Ví dụ 11.6. Kết quả tìm kiếm với HITS

Truy vấn: Japan elementary schools

## Hubs

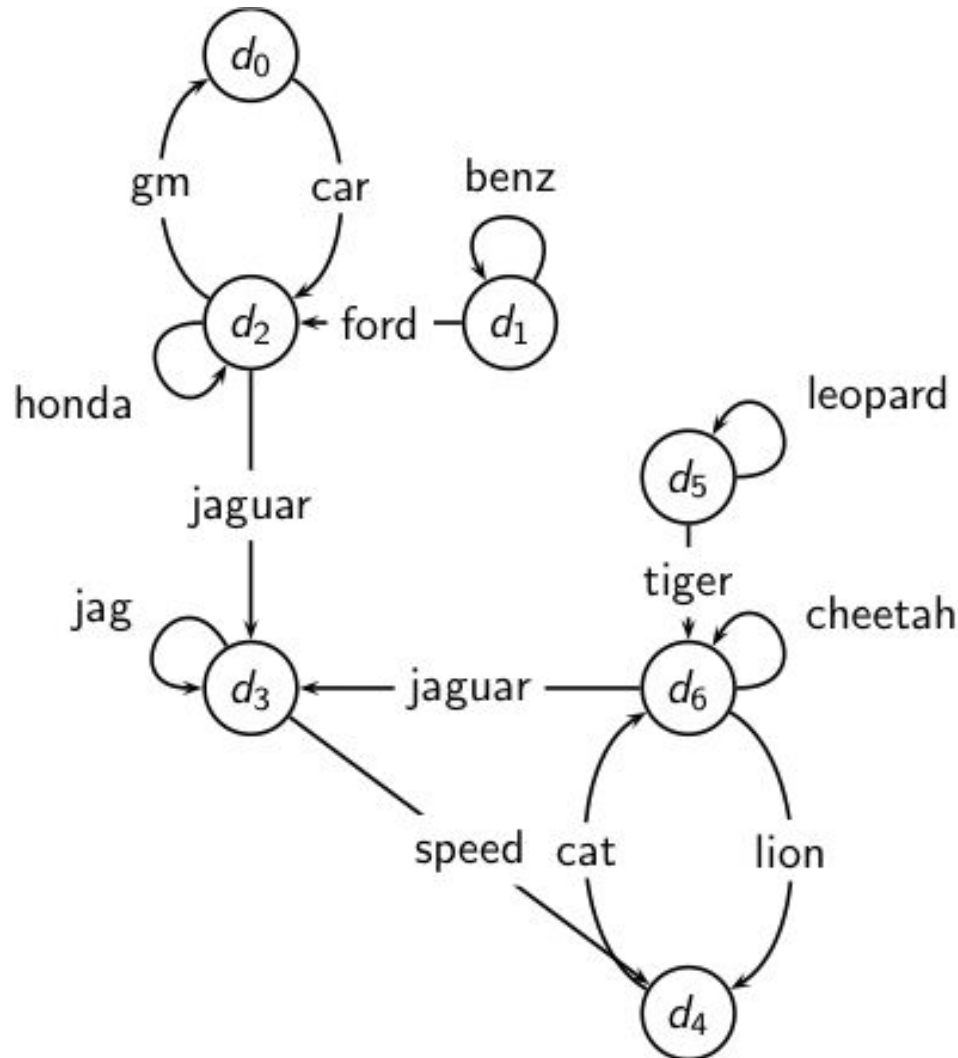
- schools
- LINK Page-13
- "ú-{,İŠw=Z
- =a%o,=ŕŠw=Zfz=[f=fy=[fW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...net and Education )
- <http://www...iglobe.ne.jp/~IKESAN>
- ,l,f,j=ŕŠw=Z,U"N,P'g•ŕŒê
- =ÒŠ—'ŕ—§=ÒŠ—ŕŒ=ŕŠw=Z
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- -y"i=ŕŠw=Z,İfz=[f=fy=[fW
- UNIVERSITY
- %oJ—ŕ=ŕŠw=Z DRAGON97-TOP
- =Â%o=ŕ=ŕŠw=Z,T"N,P'g fz=[f=fy=[fW
- ¶µ°é¼ÄÄ© ¥á¥Œ¥á¼ ¥á¥Œ¥á¼

## Authorities

- The American School in Japan
- The Link Page
- %o°=è=s—§"ä"=ŕŠw=Zfz=[f=fy=[fW
- Kids' Space
- "À=é=s—§"À=é=¼•"=ŕŠw=Z
- «{=é«ç'ăŠw•= '@=ŕŠw=Z
- KEIMEI GAKUEN Home Page ( Japanese )
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- =\_°P=iŒ§=E%oj•l=s—§'†=i=¼=ŕŠw=Z,İfy
- [http://www...p/~m\\_maru/index.html](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...

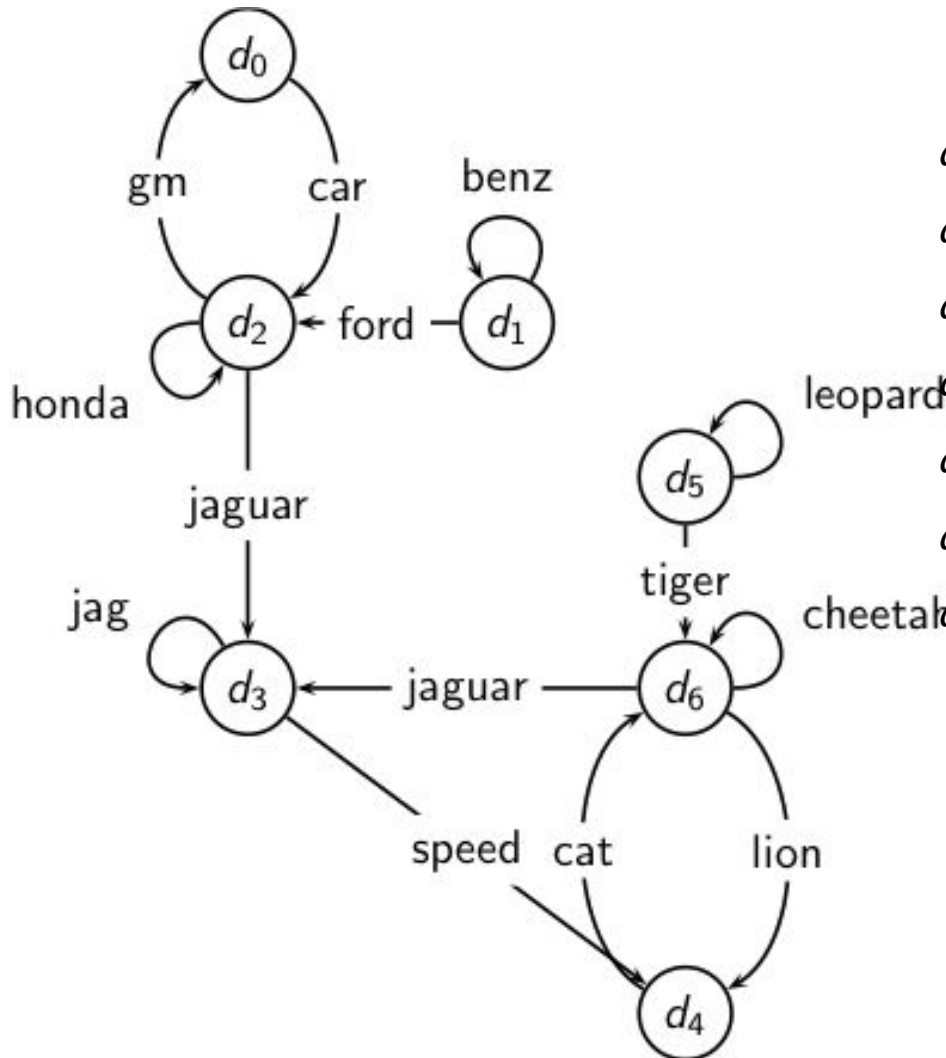
# Ví dụ 11.7. Áp dụng giải thuật HITS

*Xếp hạng các trang theo điểm bao quát và điểm phổ biến*



# Ví dụ 11.7. Áp dụng giải thuật HITS<sub>(2)</sub>

*Ma trận kề*



	$d_0$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
$d_0$	0	0	1	0	0	0	0
$d_1$	0	1	1	0	0	0	0
$d_2$	1	0	1	2	0	0	0
$d_3$	0	0	0	1	1	0	0
$d_4$	0	0	0	0	0	0	1
$d_5$	0	0	0	0	0	1	1
$d_6$	0	0	0	2	1	0	1

# Ví dụ 11.7. Áp dụng giải thuật HITS<sub>(3)</sub>

*Tính  $\vec{h}$*

	$\vec{h}_0$	$\vec{h}_1$	$\vec{h}_2$	$\vec{h}_3$	$\vec{h}_4$	$\vec{h}_5$
$d_0$	0.14	0.06	0.04	0.04	0.03	0.03
$d_1$	0.14	0.08	0.05	0.04	0.04	0.04
$d_2$	0.14	0.28	0.32	0.33	0.33	0.33
$d_3$	0.14	0.14	0.17	0.18	0.18	0.18
$d_4$	0.14	0.06	0.04	0.04	0.04	0.04
$d_5$	0.14	0.08	0.05	0.04	0.04	0.04
$d_6$	0.14	0.30	0.33	0.34	0.35	0.35



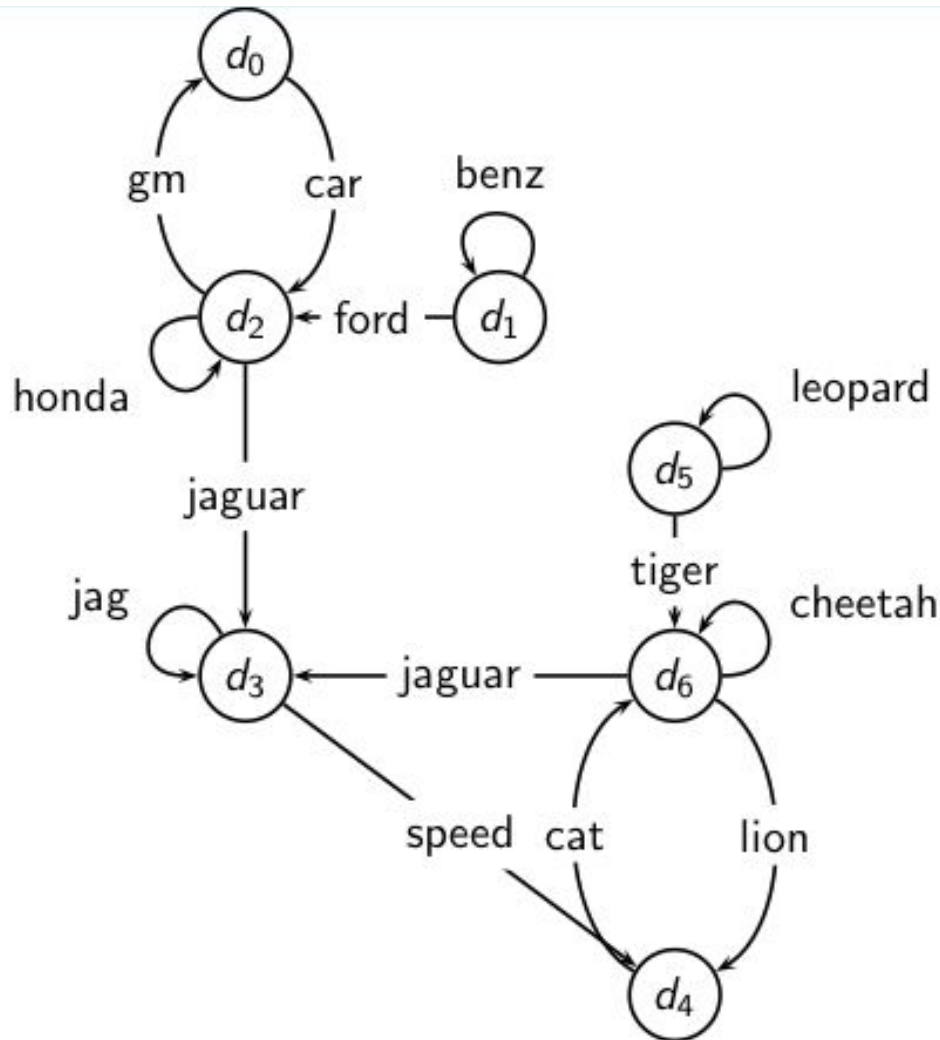
# Ví dụ 11.7. Áp dụng giải thuật HITS<sub>(4)</sub>

Tính  $\vec{a}$

	$\vec{a}_1$	$\vec{a}_2$	$\vec{a}_3$	$\vec{a}_4$	$\vec{a}_5$	$\vec{a}_6$	$\vec{a}_7$
$d_0$	0.06	0.09	0.10	0.10	0.10	0.10	0.10
$d_1$	0.06	0.03	0.01	0.01	0.01	0.01	0.01
$d_2$	0.19	0.14	0.13	0.12	0.12	0.12	0.12
$d_3$	0.31	0.43	0.46	0.46	0.46	0.47	0.47
$d_4$	0.13	0.14	0.16	0.16	0.16	0.16	0.16
$d_5$	0.06	0.03	0.02	0.01	0.01	0.01	0.01
$d_6$	0.19	0.14	0.13	0.13	0.13	0.13	0.13

# Ví dụ 11.7. Áp dụng giải thuật HITS<sub>(5)</sub>

Các giá trị  $\vec{a}$  và  $\vec{h}$



	$\vec{a}$	$\vec{h}$
$d_0$	0.10	0.03
$d_1$	0.01	0.04
$d_2$	0.12	0.33
$d_3$	0.47	0.18
$d_4$	0.16	0.04
$d_5$	0.01	0.04
$d_6$	0.13	0.35

# Ví dụ 11.7. Áp dụng giải thuật HITS<sub>(6)</sub>

*Xếp hạng các trang*

- Các trang có bậc-vào cao nhất:  $d_2$ ,  $d_3$ ,  $d_6$
- Các trang có bậc-ra cao nhất:  $d_2$ ,  $d_6$
- Các trang có PageRank cao nhất:  $d_6$
- Các trang có Hub cao nhất:  $d_6$  (tiếp sau là  $d_2$ )
- Các trang có Authority cao nhất:  $d_3$

# So sánh PageRank, Hub, và Authorities

- Cả PageRank, Hub, Authorities về bản chất toán học đều là các vec-tơ riêng của ma trận.
- PageRank được tính 1 lần trên tất cả dữ liệu trong thời gian tạo chỉ mục. Hubs và Authorities được tính cho từng truy vấn trong thời gian xử lý truy vấn.
  - Chi phí thực hiện HITS có thể quá cao
  - *(Có thể hoán đổi phạm vi áp dụng PageRank và HITS)*
- Trên Web các trang có điểm bao quát cao cũng thường có điểm phổ biến cao.
- Khác biệt giữa các kết quả xếp hạng theo PageRank và theo giải thuật HITS có thể không quá lớn.

