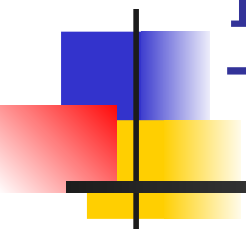
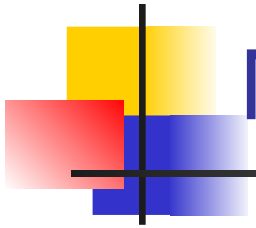


# IT4853
# Tìm kiếm và trình diễn thông tin

## Bài 19. Phân tích liên kết, PageRank

## IIR.C21. Link analysis

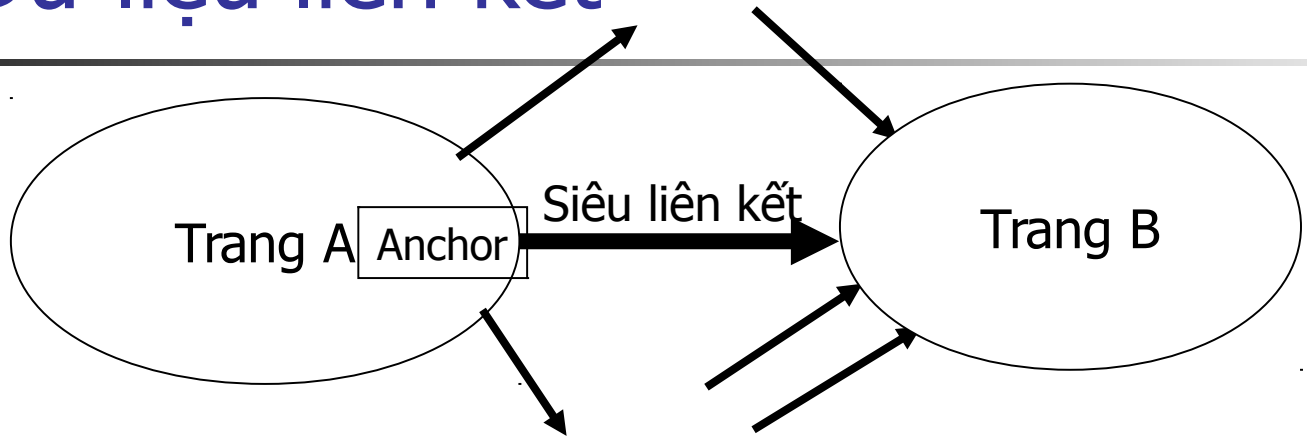
*Bộ môn Hệ thống thông tin*
*Viện CNTT & TT*



# Nội dung chính

- Dữ liệu liên kết
- Phân tích trích dẫn
- Giải thuật PageRank

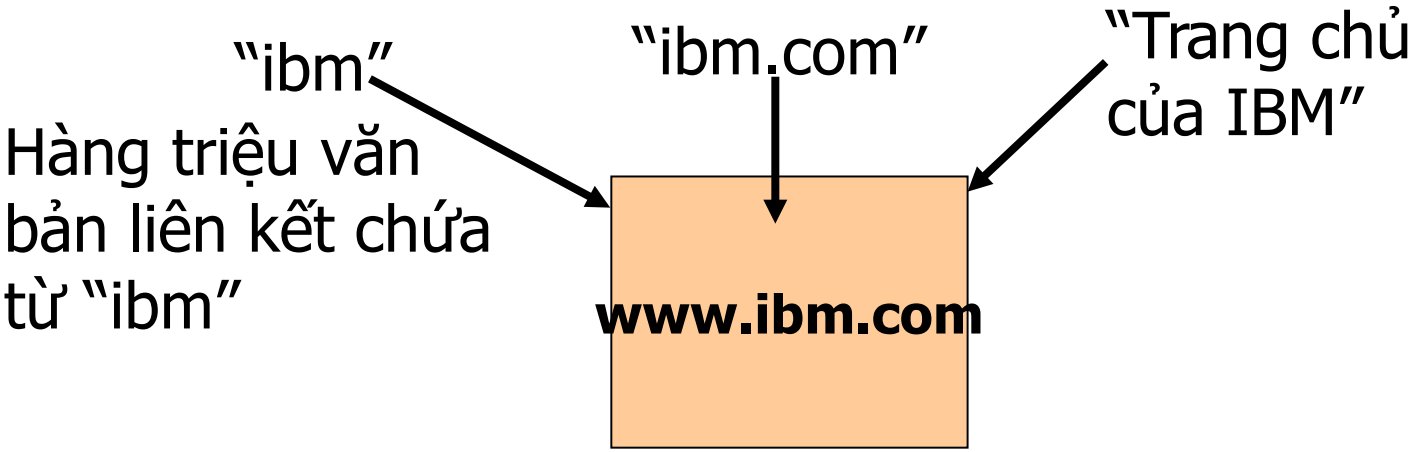
# Dữ liệu liên kết



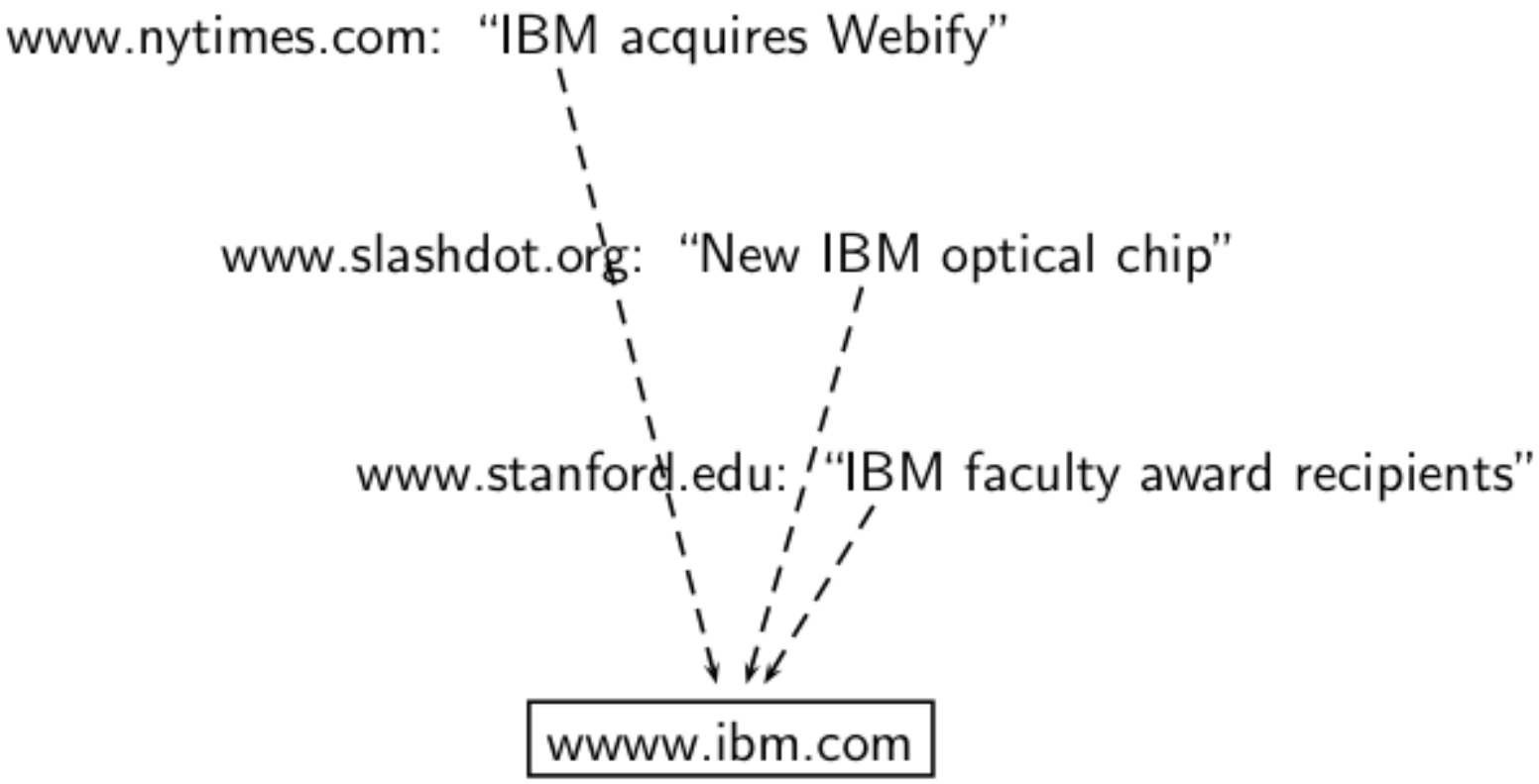

- **Giả thuyết 1:** Siêu liên kết là tín hiệu chất lượng
  - Siêu liên kết A → B là sự công nhận chất lượng trang B từ phía tác giả trang A.
- **Giả thuyết 2:** Văn bản liên kết mô tả trang B
  - Văn bản liên kết là văn bản xung quanh thẻ <a>
  - Ví dụ, Bạn có thể chọn xe máy <a href="...">ở đây</a>
  - Văn bản liên kết là "Bạn có thể chọn xe máy ở đây"

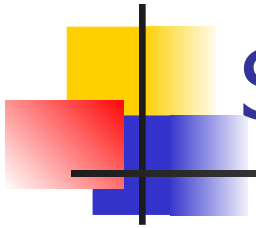
# Tìm kiếm bằng văn bản liên kết

- Ví dụ, trang www.ibm.com có nội dung đa phần là hình ảnh, rất ít từ ibm.
  - Tuy nhiên vẫn có thể tìm đến địa chỉ này bằng từ ibm.
- Tìm kiếm trên [nội dung] + [văn bản liên kết] sẽ hiệu quả hơn nếu chỉ tìm kiếm trên [nội dung]



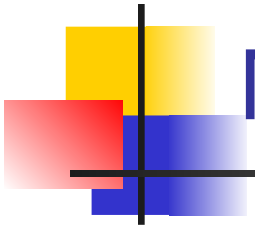
# Các văn bản liên kết của www.ibm.com chứa nhiều từ ibm





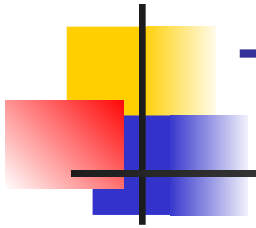
# Sử dụng văn bản liên kết

- Văn bản liên kết có thể mô tả trang web tốt hơn chính nội dung trang web đó.
- Có thể gán cho văn bản liên kết trọng số cao hơn chính nội dung trang web.



# Nội dung chính

- Dữ liệu liên kết
- Phân tích trích dẫn
- Giải thuật PageRank



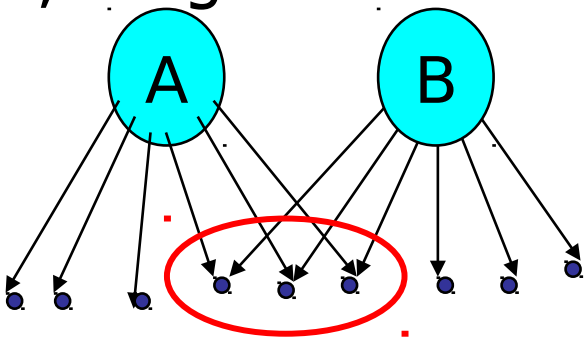
# Trích dẫn trong ấn phẩm in

- Đối với tài liệu là sách, báo, tạp trí v.v.
  - Một tài liệu có thể trích dẫn một tài liệu khác, ví dụ, tài liệu tham khảo.
- Ứng dụng:
  - Xác định độ tương đồng giữa các tài liệu
  - Đánh giá xếp hạng (impact factor) tạp trí
  - Xếp hạng tài liệu dựa trên phân tích dữ liệu liên kết
  - v.v.

Trích dẫn tài liệu có ý nghĩa tương tự siêu liên kết trong môi trường web

# Mức đồng tham khảo

- Mức đồng tham khảo của hai tài liệu *A* và *B* là số tài liệu tham khảo chung của *A* và *B*.
- Được sử dụng để đo độ tương đồng giữa các tài liệu, tác giả Kessler, công bố năm 1963.

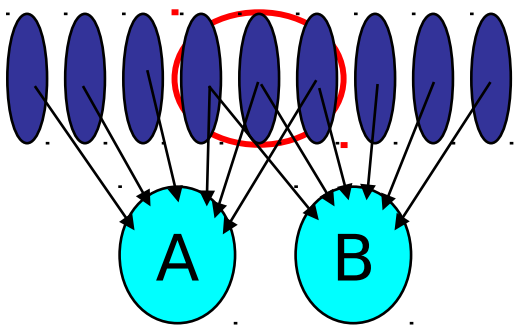



Mức đồng tham khảo: cocitation

Có nên chuẩn hóa theo số lượng trích dẫn?

# Mức đồng tham chiếu

- Mức đồng tham chiếu là số văn bản trích dẫn đồng thời cả *A* và *B*.
- Tương tự mức đồng tham khảo, được sử dụng để đánh giá độ tương đồng giữa hai tài liệu, tác giả Small, công bố năm 1973.




Có nên chuẩn hóa theo tổng số tài liệu trích dẫn A và số tài liệu trích dẫn B?

# Xếp hạng tạp trí theo impact factor

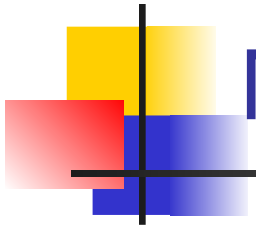
- Tác giả Garfield, công bố năm 1972
- Được tính và công bố thường niên bởi Institute for Scientific Information (ISI).
- Độ uy tín của một tạp trí *J* trong năm *Y* là số lượng trích dẫn trung bình từ các tài liệu được công bố trong năm Y tới tạp trí *J* trong năm *Y*–1 hoặc *Y*–2.
  - Không tính chất lượng của báo cáo chứa trích dẫn.

## Độ uy tín: impact factor

# Xếp hạng dựa trên phân tích trích dẫn

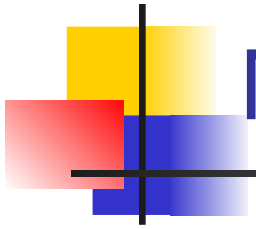
- Pinsker và Narin [1976], xếp hạng báo cáo khoa học dựa trên phân tích trích dẫn.

PageRank được phát triển theo phương pháp phân tích trích dẫn của Pinsker và Narin.



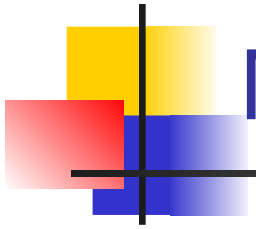
# Nội dung chính

- Dữ liệu liên kết
- Phân tích trích dẫn
- Giải thuật PageRank



# Mô hình duyệt Web ngẫu nhiên

- Quy tắc duyệt Web:
  - Bắt đầu với một trang Web bất kỳ
    - Lựa chọn ngẫu nhiên một địa chỉ để bắt đầu quá trình duyệt.
  - Lặp mở ngẫu nhiên một liên kết có trong trang hiện tại
    - Sau đó lại mở liên kết trong trang mới và cứ tiếp tục như vậy.
- Mục đích:
  - Quan sát tỉ lệ ghé thăm mỗi trang web sau một số bước đủ lớn.



# Mô hình duyệt Web ngẫu nhiên (2)

- Tỉ lệ ghé thăm mỗi trang web trong nhiều trường hợp sẽ là hằng số sau một số bước đủ lớn.
  - Không phụ thuộc vào việc lựa chọn trang bắt đầu;
  - Tỉ lệ này là PageRank của trang Web.

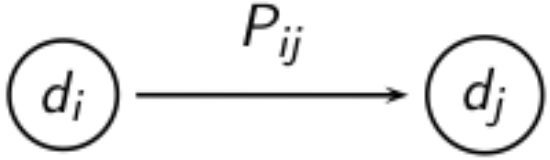
Điều kiện tồn tại tỉ lệ mở ổn định và không phụ thuộc vào trang bắt đầu là gì?

# Mô hình duyệt Web ngẫu nhiên với bước nhảy

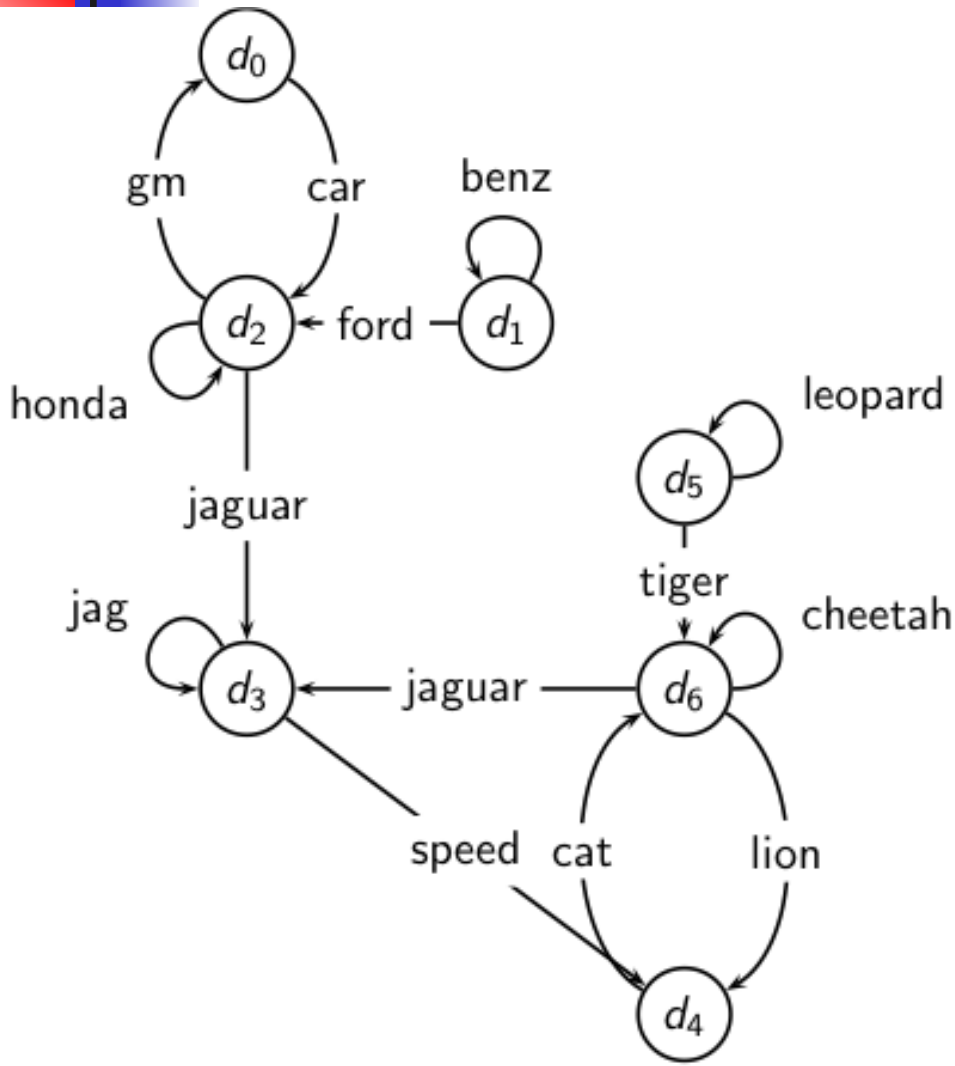
- Bước nhảy: Di chuyển đến một trang bất kỳ
  - Bổ xung thao tác này vào mô hình duyệt Web ngẫu nhiên.
- Mô hình duyệt Web ngẫu nhiên với bước nhảy:
  - Bắt đầu với một trang được lựa chọn ngẫu nhiên
  - Sau mỗi bước:
    - Nếu nút không có liên kết đi ra, thì thực hiện bước nhảy.
    - Nếu nút có liên kết đi ra, thì với xác xuất $\alpha$ thực hiện bước nhảy, với xác suất $1-\alpha$ di chuyển theo liên kết như bình thường.
- Đặt xác suất lựa chọn bước nhảy là $\alpha$, xác suất di chuyển theo liên kết là $1-\alpha$.

# Khái quát hóa quá trình duyệt Web bằng chuỗi Markov

- Chuỗi Markov gồm N trạng thái và ma trận xác suất chuyển trạng thái kích thước N x N:
  - Mỗi trạng thái tương ứng với một trang Web
  - $P_{ij}$ là xác suất chuyển từ trạng thái i sang trạng thái j, $1 \leq i, j \leq N$
    - $P_{ij}$ cũng chính là xác suất lựa chọn trang j khi đang ở trang i.
- Với i bất kỳ, $\sum P_{ij}=1$



# Ví dụ đồ thị Web



# 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 | 1 | 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 | 1 | 1 | 0 | 1 |

# Ma trận xác suất chuyển trạng thái

| | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|---|---|---|---|---|---|---|---|
| $d_0$ | 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 |

Mô hình duyệt web ngẫu nhiên

# Ma trận xác suất chuyển trạng thái (2)

| | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|---|---|---|---|---|---|---|---|
| $d_0$ | 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 |

Mô hình duyệt web ngẫu nhiên với bước nhảy với $\alpha = 0.1$

# Xác định ma trận xác suất chuyển trạng thái

- Đặt A là ma trận kề của đồ thị, dòng i cột j bằng 1 nếu có cạnh từ i tới j; Đặt $\alpha$ là xác suất nhảy ngẫu nhiên.
- Giải thuật xác định ma trận xác suất chuyển trạng thái trong trường hợp tổng quát gồm 4 bước sau:
  - 1) Nếu hàng i của A không chứa 1 thì thay thế 0 bằng 1/N, trong đó N là số lượng trang web.
  - 2) Đối với các hàng khác chia các giá trị 1 cho số lượng 1 trong hàng.
  - 3) Nhân ma trận kết quả sau khi thực hiện (1) và (2) với $1 - \alpha$
  - 4) Cộng mỗi phần tử của ma trận với $\alpha$ / N

Mô hình duyệt web ngẫu nhiên với bước nhảy

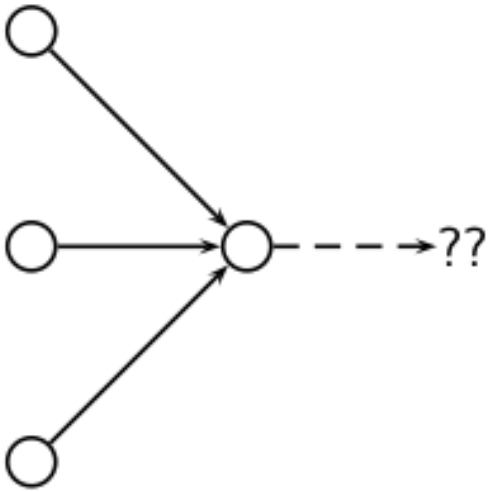
# Điều kiện dừng

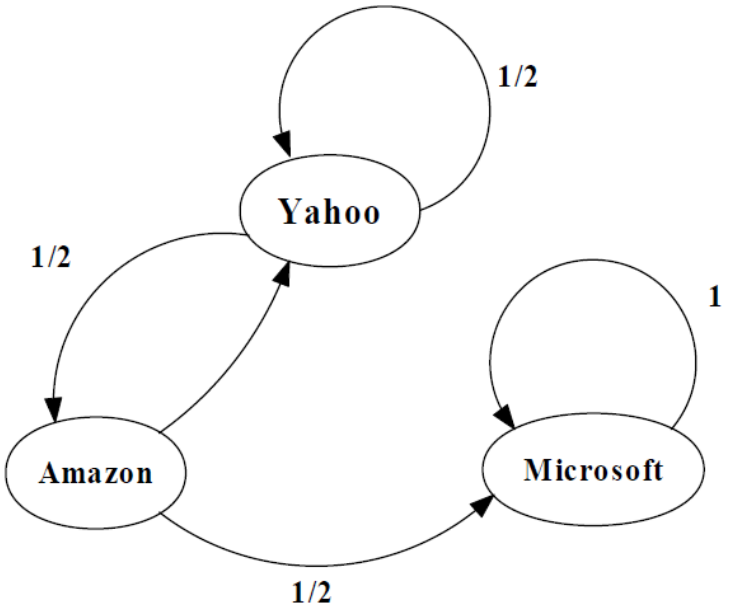
- Tồn tại phân bố xác suất ổn định nếu mô hình duyệt web thỏa mãn các điều kiện sau:
  - Luôn tồn tại đường đi giữa hai đỉnh bất kỳ: Có thể di chuyển từ một trang bất kỳ tới một trang bất kỳ;
  - Không có chu trình kín: Không thể chia các đỉnh của đồ thị thành nhiều nhóm sao cho quá trình duyệt Web trở thành tiến trình tuần tự và khép kín trong các nhóm này.

Nếu mô hình duyệt Web thỏa mãn các điều kiện nêu trên thì chuỗi Markov tương ứng cũng thỏa mãn điều kiện Ergodic

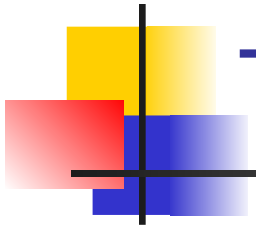
# Điều kiện dừng (2)

- Mô hình duyệt web ngẫu nhiên có thể không hội tụ
  - Ví dụ, khi tồn tại chu trình kín
  - Nút cụt (trang không có out-link)




Bổ xung bước nhảy giúp đảm bảo quá trình duyệt web ngẫu nhiên thoát khỏi nút cụt và chu trình kín.



# Tính PageRank

- Đặt $\vec{x} = (x_1, \ldots, x_N)$ là vec-tơ tỉ lệ mở liên kết, với $x_i$ là tỉ lệ mở trang thứ i.
  - $\Sigma x_i = 1$
- Biết ma trận xác suất chuyển trạng thái P, vec-tơ tỉ lệ mở liên kết ở bước tiếp theo là $\vec{x}P$.

# Tính PageRank (2)

- Đặt $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ là vec-tơ tỉ lệ mở liên kết ở trạng thái ổn định (đồng thời là vec-tơ PageRank)
  - $\vec{\pi} = \vec{\pi}P$
- Giải phương trình này cho chúng ta kết quả $\vec{\pi}$
- $\vec{\pi}$ đồng thời là vec-tơ riêng trái chính của P ...
  - ... là vec-tơ riêng có giá trị riêng lớn nhất
  - Ma trận xác suất chuyển trạng thái có giá trị riêng lớn nhất bằng 1.

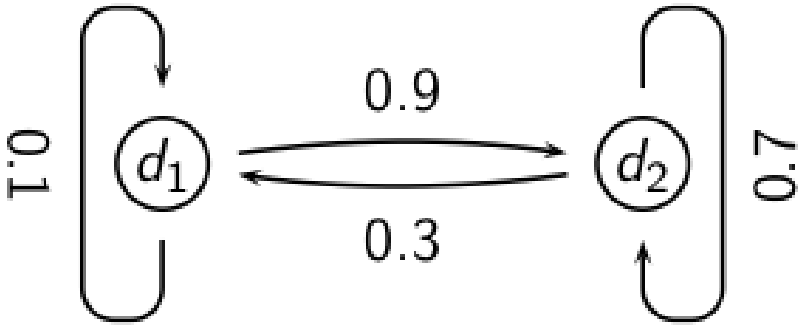
# Tính PageRank (3): Phương pháp lũy thừa

- Bắt đầu với vec-tơ x bất kỳ.
- Sau một bước chúng ta có xP.
- Sau hai bước chúng ta có $xP^2$.
- Sau k bước chúng ta có $xP^k$.
- Giải thuật: nhân x với lũy thừa tăng dần của P.
- Phương pháp này được gọi là phương pháp lũy thừa.

Không phụ thuộc vào giá trị vec-tơ x ban đầu, chúng ta luôn có giá trị không đổi ở trạng thái ổn định.

# Ví dụ phương pháp lũy thừa

- Tính PageRank cho đồ thị với các giá trị xác suất chuyển trạng thái như sau:



# Ví dụ phương pháp lũy thừa (2)

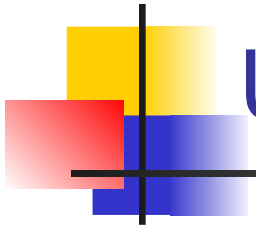
| | $x_1$ $P_t(d_1)$ | $x_2$ $P_t(d_2)$ | | | |
|---|---|---|---|---|---|
| | | | $P_{11} = 0.1$ $P_{21} = 0.3$ | $P_{12} = 0.9$ $P_{22} = 0.7$ | |
| $t_0$ | 0 | 1 | 0.3 | 0.7 | $= xP$ |
| $t_1$ | 0.3 | 0.7 | 0.24 | 0.76 | $= xP^2$ |
| $t_2$ | 0.24 | 0.76 | 0.252 | 0.748 | $= xP^3$ |
| $t_3$ | 0.252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

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

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

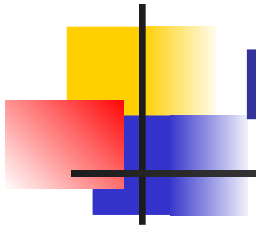
# Hạn chế của mô hình duyệt Web ngẫu nhiên

- Trong thực tế, người dùng không duyệt Web theo cách ngẫu nhiên:
  - Nút Back, danh sách trang ưa thích, tìm kiếm, v.v.
  - ➔ Mô hình chuỗi Markov không diễn tả hết được các tình huống thực tế
- Kết quả xếp hạng chỉ sử dụng PageRank có chất lượng không cao



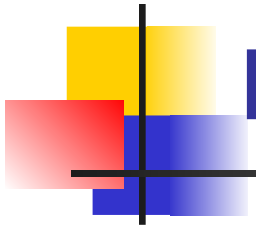
# Ứng dụng của giải thuật PageRank

- Điểm PageRank là một tín hiệu xếp hạng quan trọng đối với tìm kiếm trên Web.
- Trong thực tế điểm xếp hạng cuối cùng là tổng hợp của nhiều thành phần khác nhau, ngoài PageRank còn có văn bản liên kết, tần suất từ, khoảng cách v.v.



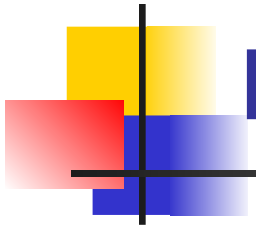
# Bài tập 24.1

- Xét một đồ thị Web đơn giản với ba đỉnh 1, 2, 3, với các liên kết như sau: 1 -> 2, 3 -> 2, 2 -> 1, 2 -> 3. Hãy viết ma trận xác suất chuyển trạng thái cho mô hình duyệt web ngẫu nhiên với bước nhảy trong ba trường hợp sau: $\alpha = 0$, $\alpha = 0.5$, $\alpha = 1$.
- Tính PageRank cho trường hợp $\alpha = 0.5$



# Bài tập 24.2

- Có thể đi từ một trang web đến một trang web bất kỳ khác hay không? Vì sao?



# Bài tập 24.3

- Hãy cung cấp một bộ văn bản liên kết cho một trang x và đề xuất một phương pháp gần đúng để lựa chọn một từ hoặc một câu mô tả tốt nhất trang x.
- Phương pháp đã đề xuất có tính đến những miền D lặp lại văn bản liên kết cho trang x từ nhiều trang của D hay không?

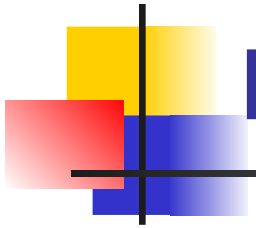
# Bài tập 24.4

- Xét một chuỗi Markov có ba trạng thái A, B và C. Xác suất chuyển trạng thái như sau:

$$p(A->B) = 1;\ p(B->A) = p_A;$$

$$p(B->C) = 1 - p_A;\ p(C->A) = 1.$$

Với những giá trị nào của $p_A$ trong khoảng [0, 1] thì chuỗi Markov đã cho thỏa mãn điều kiện ergodic?



# Bài tập 24.5

- Hãy chứng minh đối với mô hình duyệt web có bước nhảy, PageRank của trang bất kỳ không nhỏ hơn α/N với α là xác suất nhảy ngẫu nhiên. Giá trị PageRank của các trang khác biệt như thế nào khi α tiến đến 1?