

Tài liệu giảng dạy môn Khai thác dữ liệu Web

TÌM KIẾM WEB

TS. Nguyễn Ngọc Thảo – ThS. Lê Ngọc Thành
Bộ môn Khoa học Máy tính, FIT HCMUS, VNUHCM

Thành phố Hồ Chí Minh, 02/2019

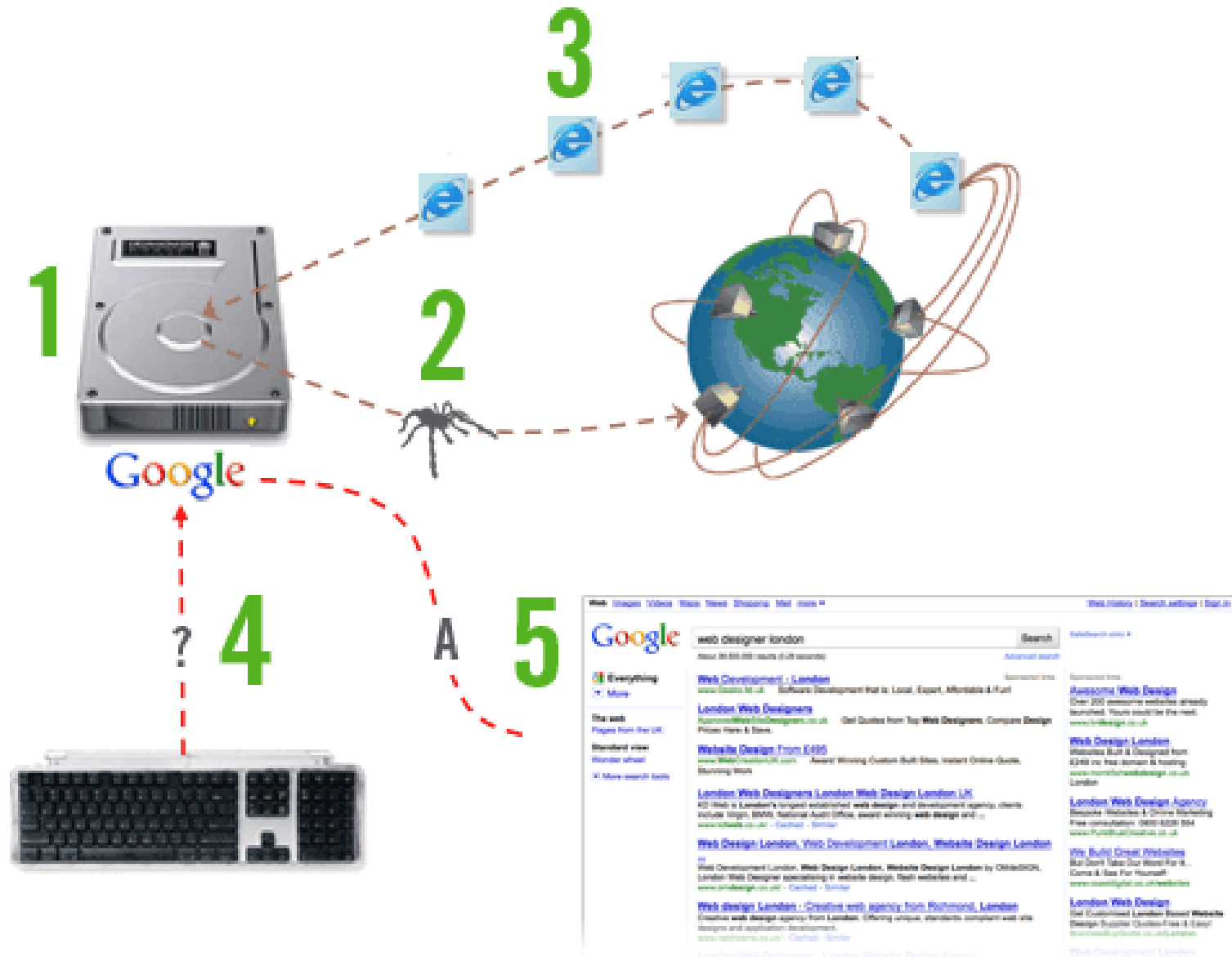
Nội dung bài giảng

- Tìm kiếm Web
- Siêu tìm kiếm và kết hợp đa xếp hạng
 - Kết hợp sử dụng độ đo độ tương tự
 - Kết hợp sử dụng vị trí xếp hạng
- Web Spamming



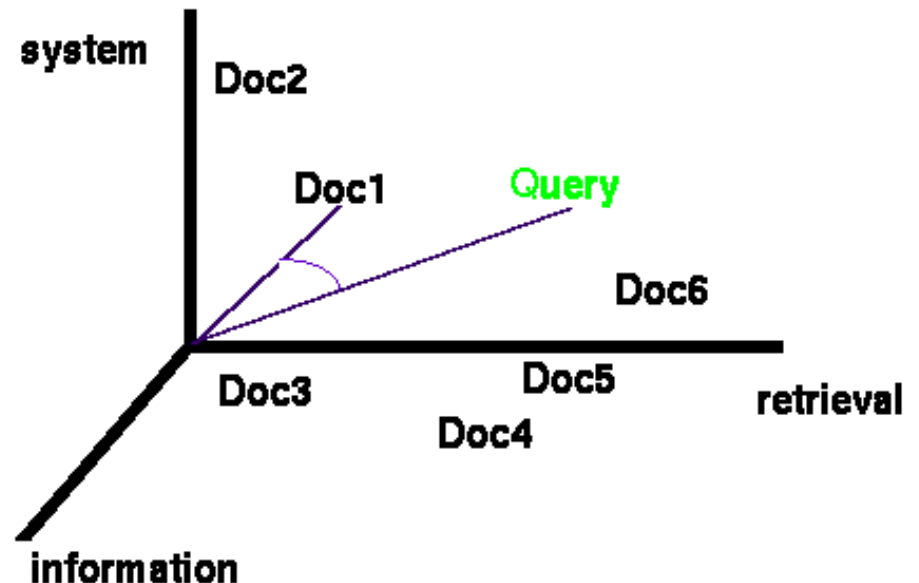
Tìm kiếm Web

Cơ chế của cỗ máy tìm kiếm



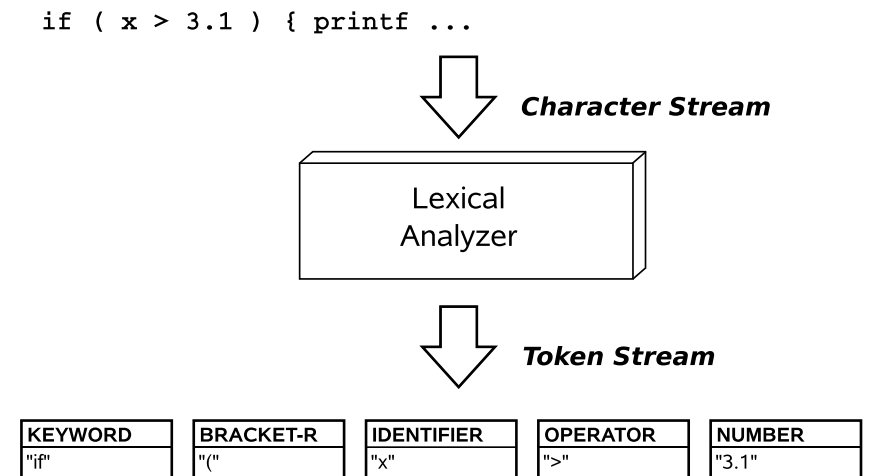
Cơ chế của cỗ máy tìm kiếm

- Chỉ mục ngữ nghĩa (latent semantic indexing) chưa được áp dụng trong Tìm kiếm Web do vấn đề hiệu suất.
- Các thuật toán tìm kiếm chủ yếu dựa trên mô hình không gian vector và so khớp từ.



Phân tích cú pháp (Parsing)

- Parser rút trích từ trang HTML đầu vào một chuỗi thuật ngữ để đánh chỉ mục.
 - Ví dụ: Lex & YACC, Flex – lexical analyzer generator,...



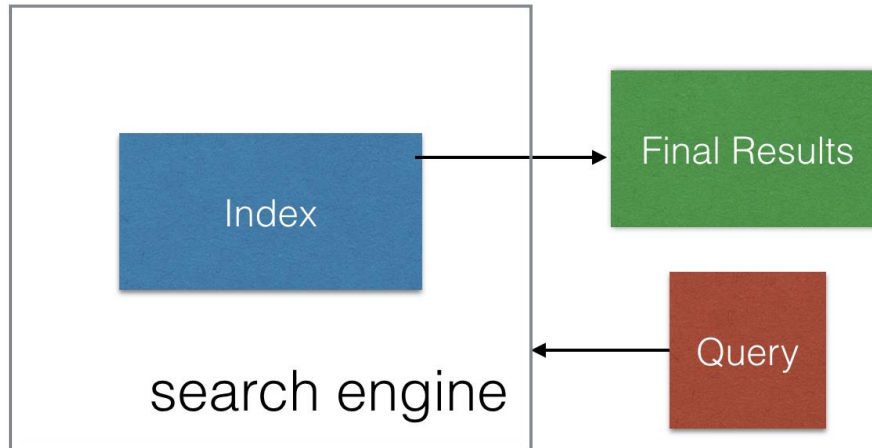
- Có thể áp dụng một vài kỹ thuật tiền xử lý vào trước hay sau khi phân tích cú pháp

Đánh chỉ mục (Indexing)

- Lập nhiều chỉ mục đảo để tăng hiệu suất truy vấn
 - Chỉ mục đảo nhỏ gồm các từ chỉ xuất hiện trong title và anchor text.
 - Chỉ mục đầy đủ chứa toàn bộ từ của trang, kể cả anchor text.
- Giải thuật tìm kiếm tiến hành trong chỉ mục nhỏ trước rồi đến chỉ mục đầy đủ.
 - Nếu tìm đủ số trang liên quan trong chỉ mục nhỏ thì hệ thống có thể không tìm tiếp trong chỉ mục đầy đủ.

Tìm kiếm và xếp hạng

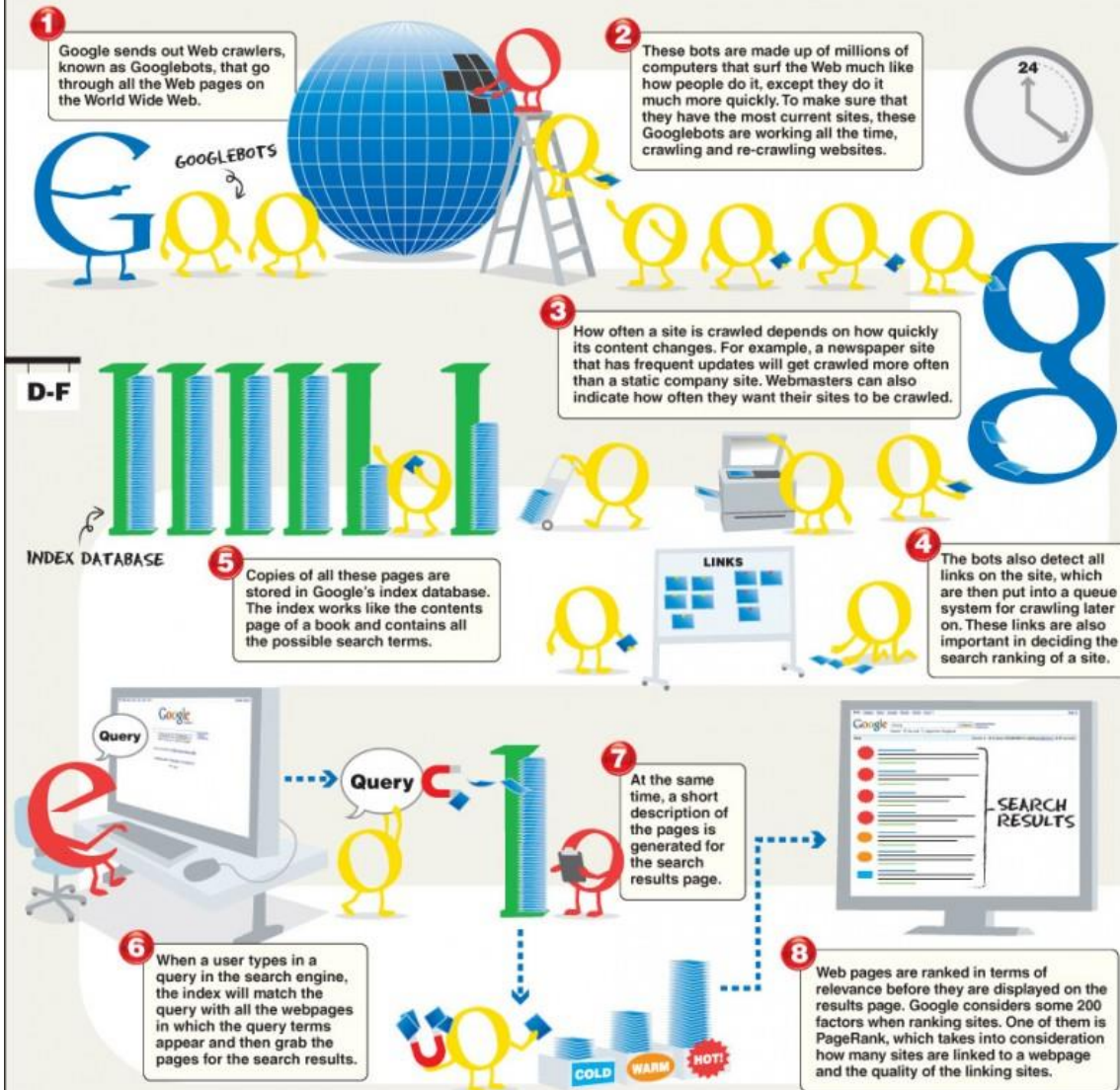
- Cho trước câu truy vấn, quá trình tìm kiếm bao gồm
 1. Tiền xử lý câu truy vấn, ví dụ loại stopwords, stemming, v.v
 2. Tìm trên chỉ mục đảo trang chứa tất cả (hay phần lớn) từ truy vấn
 3. Xếp hạng các trang tìm được và trả về cho người dùng.



- **Giải thuật xếp hạng** là cốt lõi của cỗ máy tìm kiếm

HOW GOOGLE SEARCH WORKS

Have you ever wondered what happens when you type in a query in Google's search field?
Tham Yuen-C and Quek Hong Shin go behind the scenes of the search engine



Google paper: Brin, S. and P. Lawrence. *The anatomy of a large-scale hypertextual web search engine*. Computer Networks, 1998, 30(1-7): p. 107-117

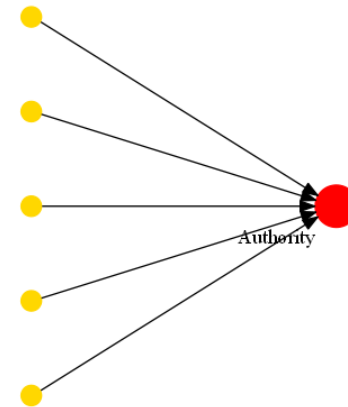
Xếp hạng văn bản và trang Web

- Hệ thống IR truyền thống xếp hạng tài liệu theo độ tương tự Cosine (hoặc độ đo khác tương đương).
- Những độ đo như thế vẫn chưa đủ với ngữ cảnh Web do chỉ xét nội dung văn bản.
 - Mỗi truy vấn trên Web đều có rất nhiều tài liệu liên quan. Ví dụ, Google trả về khoảng 46,500,000 kết quả cho truy vấn “web mining”.
- *Làm thế nào xếp hạng trang và cung cấp cho người dùng những trang tốt nhất ngay ở đầu danh sách?*

Chất lượng của trang Web

- Một trang được đánh giá là liên quan 100% vẫn có thể không phải là trang có chất lượng.
 - Ví dụ, tác giả không phải là chuyên gia trong lĩnh vực đó, thông tin được cung cấp bị sai lệch,....
- Một trang Web được đánh giá dựa trên **hệ số nội dung** (content factors) và **danh tiếng** (reputation).

```
<html>
  <head>
    <title>This is the Title of the Page</title>
  </head>
  <body>
    <h1>This is the Body of the Page</h1>
    <p>Anything within the body of a web page is
      displayed in the main browser window.</p>
  </body>
</html>
```



Đánh giá nội dung

- **Loại hình xuất hiện** (Occurrence Type): Từ truy vấn có thể xuất hiện trong trang theo nhiều hình thức khác nhau.

Title



URL

Body

font size bold/italic...

Đánh giá nội dung

- Loại hình xuất hiện (Occurrence Type)

The screenshot shows a web browser with multiple tabs. The active tab is 'Web mining - Wikipedia', displaying the article 'Web mining'. The URL bar shows 'https://en.wikipedia.org/wiki/Web_mining'. A red arrow points from the text 'web mining' in the article body to the 'Web mining' link in the browser's address bar, illustrating the concept of anchor text. The article text includes: 'Due to a lack of flexibilities in European copyright and d... works such as [web mining](#) without the permission of the database is pure data in Europe there... mining becomes subject to regula... greaves review this led to the UK government to... t mining as a limitation and exception. Only the se...'. The Wikipedia sidebar on the left includes the logo and navigation links: 'Main page', 'Contents', 'Featured content', 'Current events', and 'Random article'. A notice at the bottom of the article states: 'This article may **require** improve this article if you...'. The bottom of the article text reads: 'Web mining - is the application of data mining techniqu...'. A red arrow points from the text 'web mining' in the article body to the 'Web mining' link in the browser's address bar, illustrating the concept of anchor text.

Đánh giá nội dung

- **Đếm (Count):** Số lần xuất hiện của mỗi loại hình xuất hiện.
 - Ví dụ, từ truy vấn xuất hiện trong trường title 2 lần → title count = 2.
- **Vị trí (Position):** Vị trí của từ trong mỗi loại hình xuất hiện.
 - Dùng để đánh giá tính lân cận khi có nhiều từ truy vấn.
 - Các từ truy vấn nằm gần nhau hoặc xuất hiện thành chuỗi theo đúng thứ tự được xem là tốt hơn.

Truy vấn đơn từ

- Giả sử **reputation score** đã được tính cho mỗi trang.
- Mỗi loại hình xuất hiện được gán trọng số, **type weight**.
- Số lần xuất hiện được chuyển thành **count weight**.

	t_1	t_2	...	t_n
Type weight vector	tw_1	tw_2	...	tw_n
Count weight vector	cw_1	cw_2	...	cw_n

- $IR\ score = type\ weight\ vector \cdot count\ weight\ vector$
- Điểm cuối cùng để đánh giá trang là kết hợp của IR score và reputation score.

Truy vấn đa từ

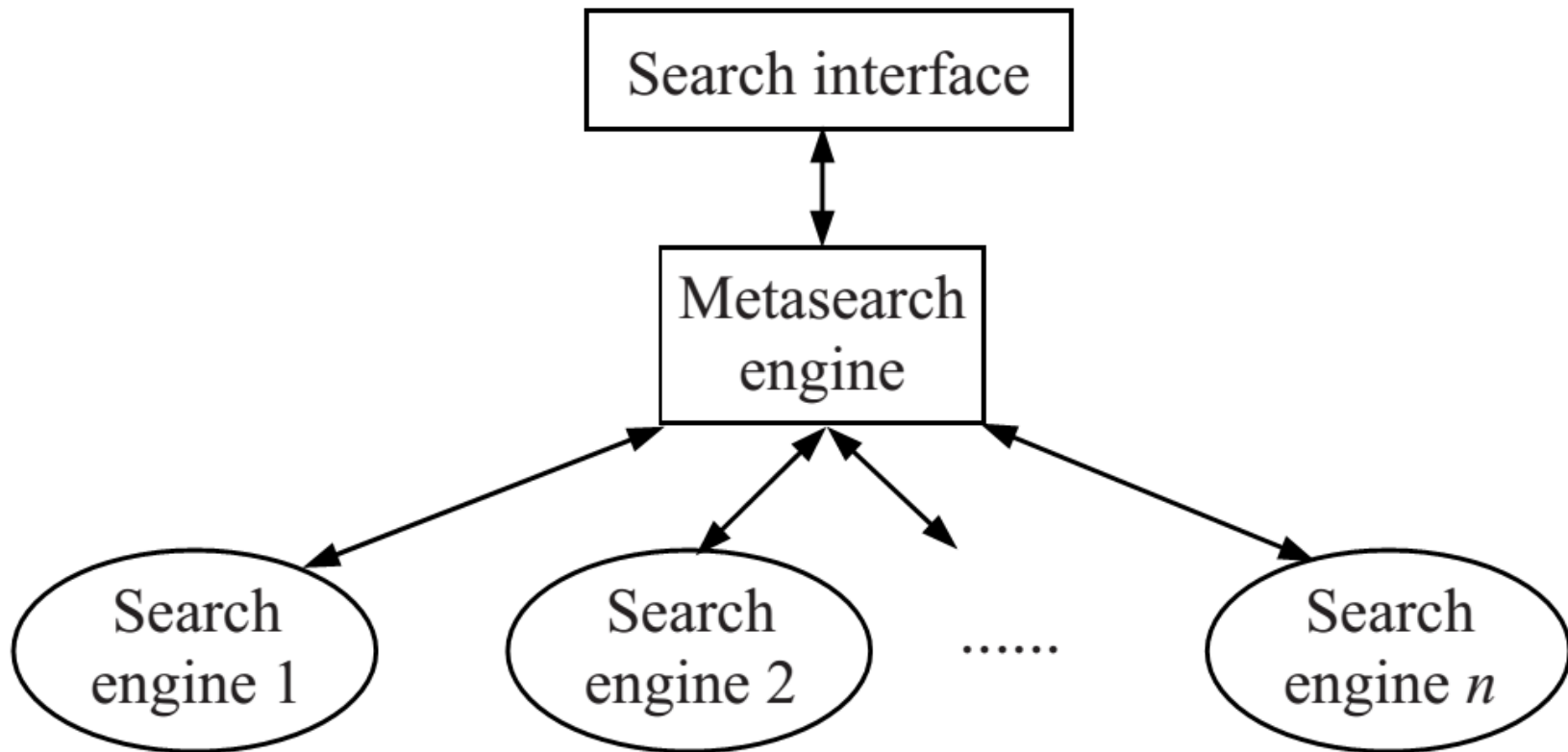
- Thực hiện tương tự như truy vấn đơn từ nhưng xét thêm độ lân cận và tính thứ tự
- Cụm từ xuất hiện gần nhau trong một trang cần được gán trọng số cao hơn những từ ở xa nhau.
 - Ước lượng độ lân cận cho mỗi cụm từ tìm được → mỗi cặp loại hình – độ lân cận có một trọng số **type-proximity weight**.
 - Tính giá trị count cho mỗi cặp loại hình – độ lân cận và chuyển thành count weight.



Siêu tìm kiếm và kết hợp đa xếp hạng

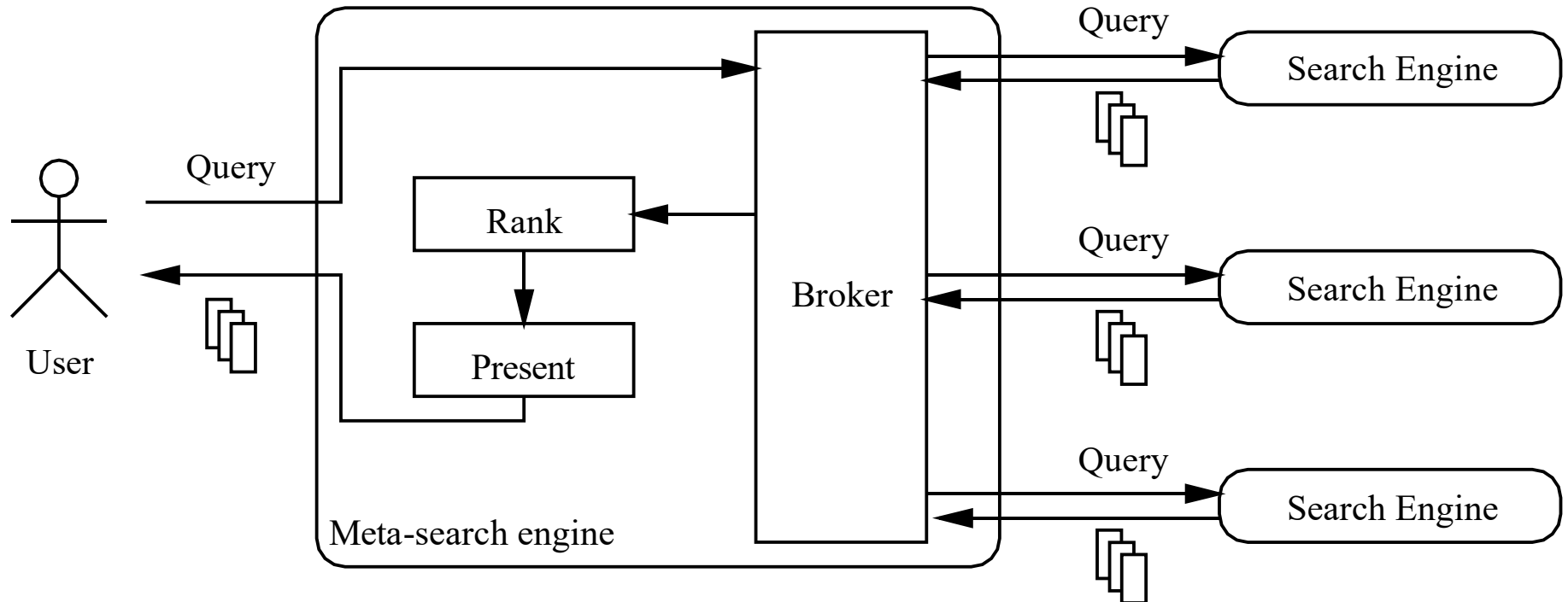
Cỗ máy siêu tìm kiếm

- Sử dụng nhiều cỗ máy tìm kiếm để tạo ra **cỗ máy siêu tìm kiếm** (meta-search engine)



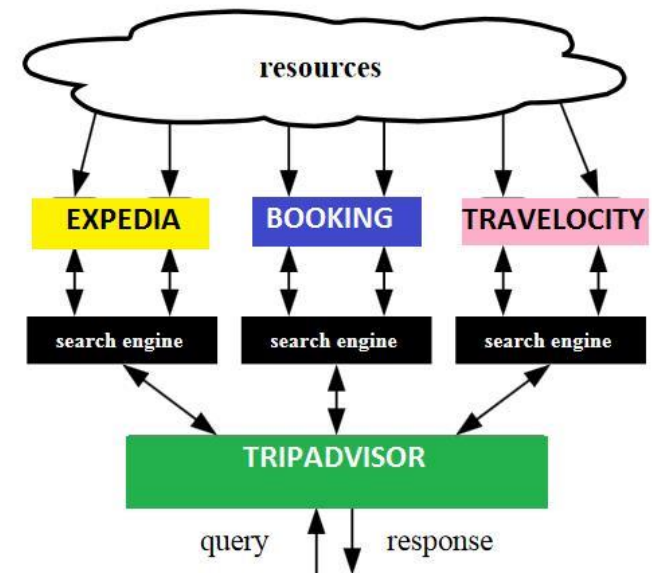
Cỗ máy siêu tìm kiếm

- Không tổ chức cơ sở dữ liệu riêng.
- Thay vào đó, tổng hợp kết quả tìm kiếm từ các cỗ máy tìm kiếm thông thường



Cỗ máy siêu tìm kiếm

- Siêu tìm kiếm giúp tăng độ bao phủ của tìm kiếm trên Web.
 - Mỗi cỗ máy tìm kiếm chỉ giải quyết một phần nhỏ trên kho thông tin Web khổng lồ.
- Siêu tìm kiếm tăng cường hiệu quả của phép tìm kiếm.
 - Mỗi cỗ máy tìm kiếm thường thiên vị cho một số loại trang/truy vấn.



Ví dụ về cỗ máy siêu tìm kiếm

- Cỗ máy siêu tìm kiếm trong dịch vụ du lịch

1. Airlines



2. Travel Agent



3. Travel Metasearch



Tác vụ chính trong siêu tìm kiếm

- Kết hợp kết quả xếp hạng từ các cỗ máy tìm kiếm thành phần để tạo ra một xếp hạng đơn.
1. Nhận diện trang giống nhau từ các cỗ máy tìm kiếm khác nhau để loại bỏ trùng.
 - Khó khăn: aliases, symbolic links, redirections,...→ Heuristics: so sánh tên domain của URL, tựa đề trang,...
 2. Tổng hợp kết quả xếp hạng từ các cỗ máy tìm kiếm đơn để tạo ra xếp hạng đơn.
 - Dựa vào điểm số tương tự hoặc vị trí xếp hạng được xác định bởi từng cỗ máy tìm kiếm.

Kết hợp sử dụng điểm tương tự

- Gọi tập tài liệu cần được xếp hạng là $D = \{d_1, d_2, \dots, d_N\}$.
- Có k hệ thống cơ sở (cỗ máy tìm kiếm thành phần hoặc kỹ thuật xếp hạng).
- Hệ thống i gán cho tài liệu d_j một điểm tương tự s_{ij} .
- **CombMIN:** Điểm tương tự kết hợp cho mỗi tài liệu d_j là giá trị nhỏ nhất trong mọi hệ thống cơ sở.

$$CombMIN(d_j) = \min(s_{1j}, s_{2j}, \dots, s_{kj})$$

- **CombMAX:**

$$CombMAX(d_j) = \max(s_{1j}, s_{2j}, \dots, s_{kj})$$

Kết hợp sử dụng điểm số tương tự

- **CombSUM:**

$$CombSUM(d_j) = \sum_{i=1}^k s_{kj}$$

- **CombANZ:**

$$CombANZ(d_j) = \frac{CombSUM(d_j)}{r_j}$$

- r_j là số lượng hệ thống đã truy vấn được d_j ($r_j \neq 0$)

- **CombMNZ:**

$$CombANZ(d_j) = CombSUM(d_j) \times r_j$$

Kết hợp sử dụng vị trí xếp hạng

- Dựa trên kỹ thuật **bỏ phiếu bầu cử** (voting in elections) của lĩnh vực lý thuyết chọn lựa xã hội (social choice theory).
- Xếp hạng Borda (Borda ranking)
- Xếp hạng Condorcet (Condorcet ranking)
- Xếp hạng thuận nghịch (Reciprocal ranking)



Xếp hạng trong siêu tìm kiếm: Ví dụ

- Xét cỗ máy siêu tìm kiếm có 5 hệ thống cơ sở.
- Các hệ thống xếp hạng 4 ứng cử viên a, b, c và d như sau
 - Hệ thống 1: a, b, c, d
 - Hệ thống 2: b, a, d, c
 - Hệ thống 3: c, b, a, d
 - Hệ thống 4: c, b, d
 - Hệ thống 5: c, b
- Gọi $Score(x)$ là điểm số cho mỗi ứng cử viên x .

Xếp hạng Borda

- 1770, Jean-Charles de Borda, “election by order of merit”.
- Giả sử có n ứng viên trong cuộc bầu cử.
- Với mỗi hệ thống, ứng viên xếp hạng đầu nhận n điểm, ứng viên xếp thứ hai nhận $n - 1$ điểm,...
- Nếu có ứng viên không được xếp hạng, số điểm còn lại chia đều cho các ứng viên chưa xếp hạng.
- Điểm số từ mọi hệ thống được cộng dồn thành điểm cuối cùng cho mỗi ứng viên.
- Ứng cử viên có điểm cao nhất sẽ thắng.

Xếp hạng Borda: Ví dụ

- Điểm số từ mỗi hệ thống cơ sở

	a	b	c	d
Hệ thống 1	4	3	2	1
Hệ thống 2	3	4	1	2
Hệ thống 3	2	3	4	1
Hệ thống 4	1	3	4	2
Hệ thống 5	1.5	3	4	1.5

- Như vậy,
 - $Score(a) = 4 + 3 + 2 + 1 + 1.5 = 11.5$
 - $Score(b) = 3 + 4 + 3 + 3 + 3 = 16$
 - $Score(c) = 2 + 1 + 4 + 4 + 4 = 15$
 - $Score(d) = 1 + 2 + 1 + 2 + 1.5 = 7.5$
- Xếp hạng cuối cùng là b, c, a, d .

Xếp hạng Condorcet

- 1785, Marquis de Condorcet.
- Ứng viên thắng mọi ứng viên khác trong cuộc so sánh tay đôi là người thắng cuộc.
- Nếu ứng viên không được bầu bởi một hệ thống nào đó, nó thua mọi ứng viên được bầu khác.
- Mọi ứng cử viên không được bầu đều hòa với nhau.

Xếp hạng Condorcet: Ví dụ

- Ma trận so sánh tay đôi (thắng:thua:hòa) giữa 4 ứng cử viên

	a	b	c	d
a	–	1:4:0	2:3:0	3:1:1
b	4:1:0	–	2:3:0	5:0:0
c	3:2:0	3:2:0	–	4:1:0
d	1:3:1	0:5:0	1:4:0	–

- Bảng thắng – thua – hòa:

	thắng	thua	hòa
a	1	2	0
b	2	1	0
c	3	0	0
d	0	3	0

- Xếp hạng cuối cùng là c, b, a, d .

Xếp hạng thuận nghịch

- Tương tự như xếp hạng Borda nhưng gán trọng số cao hơn cho các ứng cử viên nằm gần đầu danh sách.
- Với mỗi hệ thống, ứng viên xếp hạng đầu nhận 1 điểm, ứng cử viên xếp thứ hai nhận $1/2$ điểm, ứng cử viên xếp thứ ba nhận $1/3$ điểm...
- Nếu ứng viên không được xếp hạng bởi một hệ thống, bỏ qua ứng viên này trong lượt xét của hệ thống đó.

Xếp hạng thuận nghịch: Ví dụ

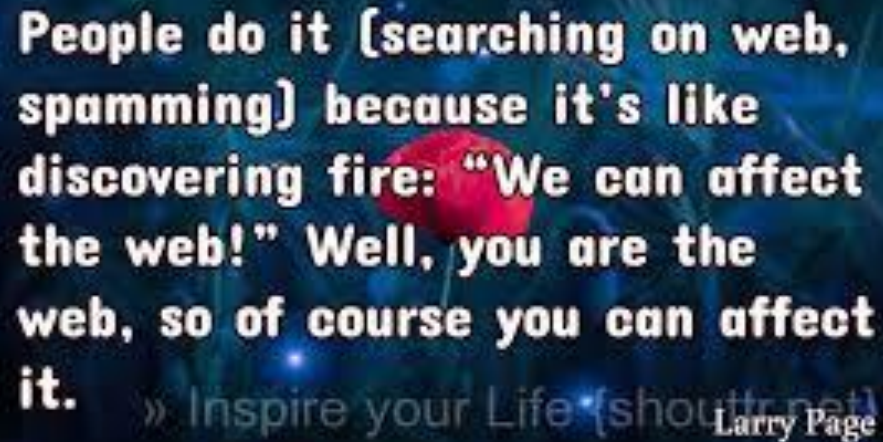
- Điểm số từ mỗi hệ thống cơ sở

	a	b	c	d
Hệ thống 1	1	1/2	1/3	1/4
Hệ thống 2	1/2	1	1/4	1/3
Hệ thống 3	1/3	1/2	1	1/4
Hệ thống 4	—	1/2	1	1/3
Hệ thống 5	—	1/2	1	—

- Như vậy,
 - $Score(a) = 1 + 1/2 + 1/3 = 1.83$
 - $Score(b) = 1/2 + 1 + 1/2 + 1/2 + 1/2 = 3$
 - $Score(c) = 1/3 + 1/4 + 1 + 1 + 1 = 3.55$
 - $Score(d) = 1/4 + 1/3 + 1/4 + 1/3 = 1.17$
- Xếp hạng cuối cùng là c, b, a, d .

Bài tập 1: Xếp hạng siêu tìm kiếm

- Xét cỗ máy siêu tìm kiếm có 5 hệ thống cơ sở, xếp hạng 4 tài liệu a , b , c và d như bên dưới.
 - Hệ thống 1: a, c, d, b
 - Hệ thống 2: c, a, b, d
 - Hệ thống 3: c, b, a, d
 - Hệ thống 4: a, c, b
 - Hệ thống 5: a, b
- Cho biết kết quả xếp hạng cuối cùng theo mỗi phương pháp xếp hạng đã học.



People do it (searching on web, spamming) because it's like discovering fire: "We can affect the web!" Well, you are the web, so of course you can affect it. » Inspire your Life (shouttr.net) Larry Page

Web Spamming

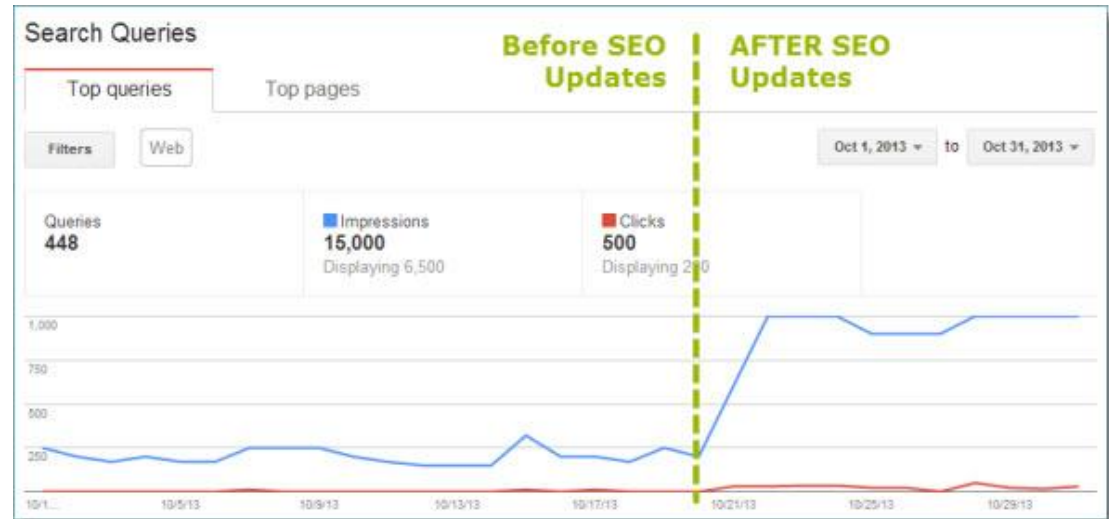
Spamming là gì?

- Trang xuất hiện nhiều trong những lượt tìm kiếm Web có thể đem lại sự tăng trưởng tài chính và danh tiếng đáng kể.
 - Thứ hạng của trang trong kết quả tìm kiếm là chỉ số quan trọng nhất.
- Cần hiểu rõ giải thuật xếp hạng của cỗ máy tìm kiếm để trình bày thông tin trên trang



Spamming là gì?

- **Spamming** là hoạt động chủ tâm dẫn dắt cỗ máy tìm kiếm xếp hạng một số trang cao hơn giá trị thực của chúng.
 - Khó phân định giữa spamming và tối ưu hóa trang.
 - Gây phiền toái cho người dùng, tổn hại đến môi trường Web và hoạt động của cỗ máy tìm kiếm.
- **SEO** (Search Engine Optimization): dịch vụ cải thiện thứ hạng tìm kiếm cho trang Web của khách hàng.



Content (Term) spamming

- Đa số cỗ máy tìm kiếm dùng các biến thể TF-IDF để đánh giá độ liên quan của một trang với câu truy vấn.
- **Content spamming** sắp đặt nội dung trong trường văn bản HTML sao cho trang spam trở nên liên quan hơn.
 - Từ spam có thể xuất hiện trong bất kỳ trường nào: title, meta tags (trọng số thấp hoặc bỏ qua), body, anchor text và URL.

Kỹ thuật content spamming

- Lặp lại một số thuật ngữ quan trọng

- Tăng điểm TF của từ spam trong tài liệu và nhờ đó tăng độ liên quan của tài liệu.
- Từ spam được đặt ngẫu nhiên trong các câu. Ví dụ, lặp từ “mining”
→ “the picture *mining* quality of this camera *mining* is amazing”

Online auto insurance quote

Michigan car insurance quote Nj car insurance quote Texas car insurance quote Health and life insurance quote online Fortis health ransamerica Banner insurance life quote Universal life insurance ar

- ♦ Online auto insurance quote
- ♦ Free car insurance quote
- ♦ Online auto insurance quote

Online auto insurance quote

Cheap car insurance quote On line car insurance quote Car insurance Uninsured car insurance quote Ohio car insurance quote Free instar insurance quote usaa Alberta car insurance quote Female car insurance quote Free online car insurance quote Car insurance mercury quote

Kỹ thuật content spamming

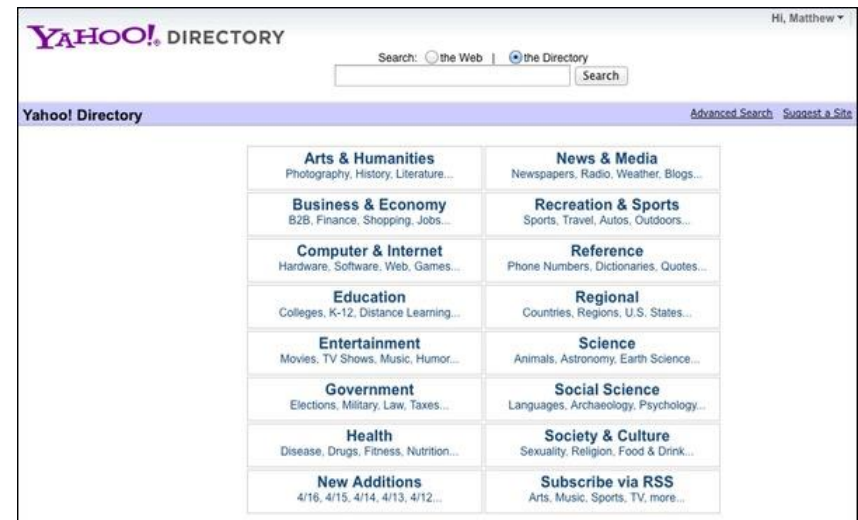
- Chèn vào nhiều thuật ngữ không liên quan
 - Sao chép câu từ những trang liên quan với trang đang xét.
 - Đặt vào trang những thuật ngữ thường được tìm kiếm trên Web.
- Ví dụ, để quảng cáo chương trình du lịch trên biển (cruise liners hay cruise holiday packages), đặt cụm từ “Tom Cruise” vào trang.



Link spamming

- Hyperlink đóng vai trò quan trọng trong việc xác định điểm reputation của một trang.
- **Out-Link spamming:** đặt out-link trỏ đến các trang authority để tăng điểm hub của trang đang xét.
 - **Directory cloning:** sao chép một lượng lớn từ thư mục chứa liên kết đến các trang Web (ví dụ, Yahoo! Directory).

Ref: <https://wordimpress.com/top-10-dmoz-competitors-open-project/>



Link spamming

- **In-Link spamming:** khó thực hiện hơn vì không thể tự ý đặt hyperlink lên trang của người khác.
 1. Tạo honey pot: xây dựng một số trang quan trọng chứa liên kết đến trang spam (ví dụ, trang glossary, FAQ, help,...).
 2. Thêm link vào các thư mục Web.
 3. Post link vào các nội dung do người dùng phát sinh (reviews, thảo luận trên forum, blogs,...).



The screenshot shows a Facebook post by "Mission Dolores Park" about a picnic. The post has 32,963 impressions and 0.04% feedback. Below the post, there are three comments. The first comment is by Saurabh Sharma, asking "What do we do with spam?". The second comment is by Xavier Llorà, suggesting to "Fry it and make a nice sandwich using a recently baked french baguette.". The third comment is by Saurabh Sharma, saying "That sounds magically delicious!!". The fourth comment is by Xavier Llorà, asking "Have you tried adding BACON to it.". A red arrow points to the text "Show comments removed as spam." in the comments section.

Mission Dolores Park – BYO–Picnic, for the afterparty !
A Park Clean-up and Picnic
Location: Dolores Park (behind the tennis courts)
Time: 3:00PM Monday, July 5th
32,963 Impressions · 0.04% Feedback
July 1, 2010 at 1:25pm · Like · Comment · Share
10 people like this.

Robert thanks laureen, i've got an ice chest.
July 1, 2010 at 1:29pm · Like

Trevor Bryant <http://www.porntubescan.com/>
50 seconds ago · Like

Saurabh Sharma · Yesterday 2:52 PM · Limited
What do we do with spam?
+1 · Comment · Hang out · Share
3 comments
[Show comments removed as spam.](#)

Xavier Llorà · Fry it and make a nice sandwich using a recently baked french baguette.
Yesterday 2:55 PM

Saurabh Sharma · That sounds magically delicious!!
Yesterday 3:09 PM · Edit

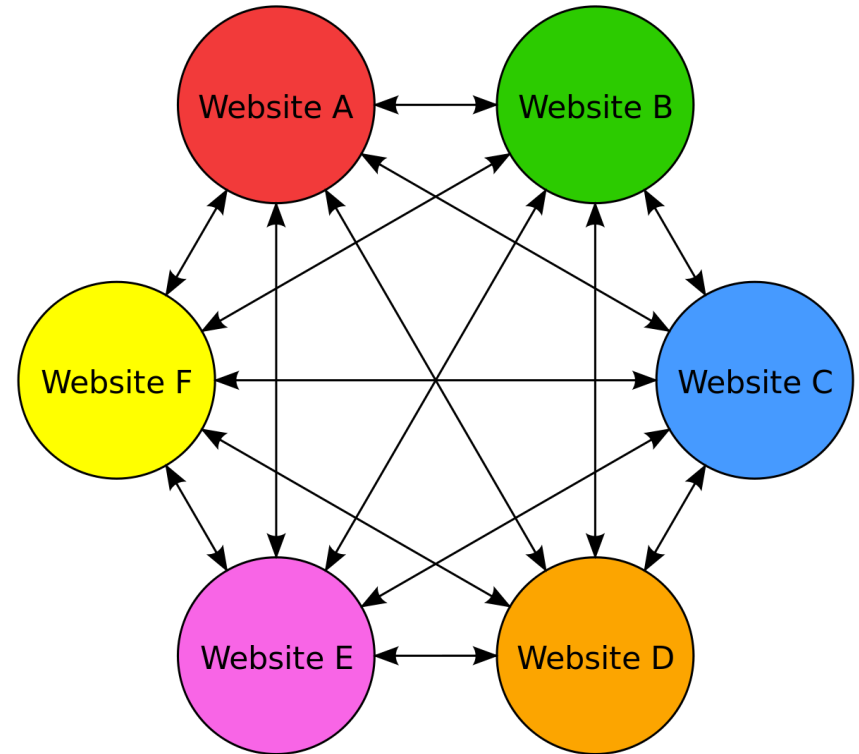
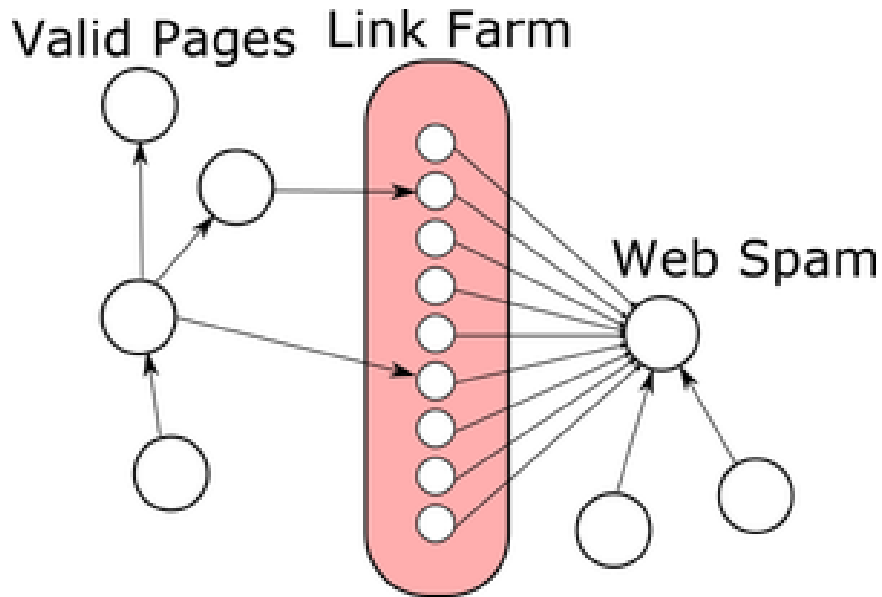
Xavier Llorà · Have you tried adding **BACON** to it.
Yesterday 3:29 PM

Add a comment...

Link spamming

- **In-Link spamming**

4. Tham gia vào việc trao đổi link.
5. Tự tạo spam farm.



Kỹ thuật che giấu

- Spammer luôn cần phải giấu hay ngụy trang các câu/từ spam và hyperlink để người dùng Web không thể nhận ra nội dung spam.
- **Giấu nội dung:** làm cho các mục spam trở nên vô hình.
 - Chọn màu chữ tiếp với màu nền

```
<body background = white>  
<font color = white> spam items</font>  
...  
</body>
```
 - Sử dụng ảnh trống và rất nhỏ để giấu hyperlink.

```
<a href = target.html"> </a>
```
 - Sử dụng script để dấu một vài thành phần thị giác trên trang.
 - Chỉnh sửa trên HTML style attribute

Kỹ thuật che giấu

- **Cloaking:** Spam Web server trả trang HTML có nội dung chính quy cho người dùng và trang spam cho Web crawler.
 - Duy trì danh sách địa chỉ IP của các cỗ máy tìm kiếm và nhận diện crawler bằng cách so khớp địa chỉ IP.
 - Nhận diện trình duyệt Web dựa trên trường user-agent trong thông điệp HTTP request.

GET /pub/WWW/TheProject.html HTTP/1.1

Host: www.w3.org

User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)

Kỹ thuật che giấu

- **Redirection:** Tự động hướng trình duyệt tới URL khác ngay khi trang vừa được tải.
 - Trang spam được đưa cho cỗ máy tìm kiếm để đánh chỉ mục (người dùng sẽ không bao giờ thấy nó) và trang mục tiêu được trình bày cho người dùng thông qua redirection.
 - “Refresh” meta-tag với thời gian refresh bằng 0, hoặc sử dụng script.



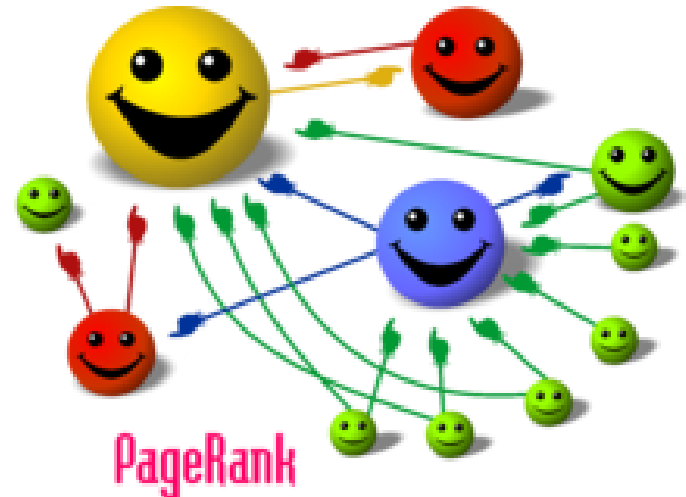
Chống redirection và cloaking spam

- **Redirection** bằng cách refresh meta-tag dễ bị phát hiện, trong khi redirection bằng script khó bị nhận diện hơn vì crawler không thực thi script.
- Crawler đôi lúc phải tự nhận mình là trình duyệt Web thông thường để chống **cloaking**.



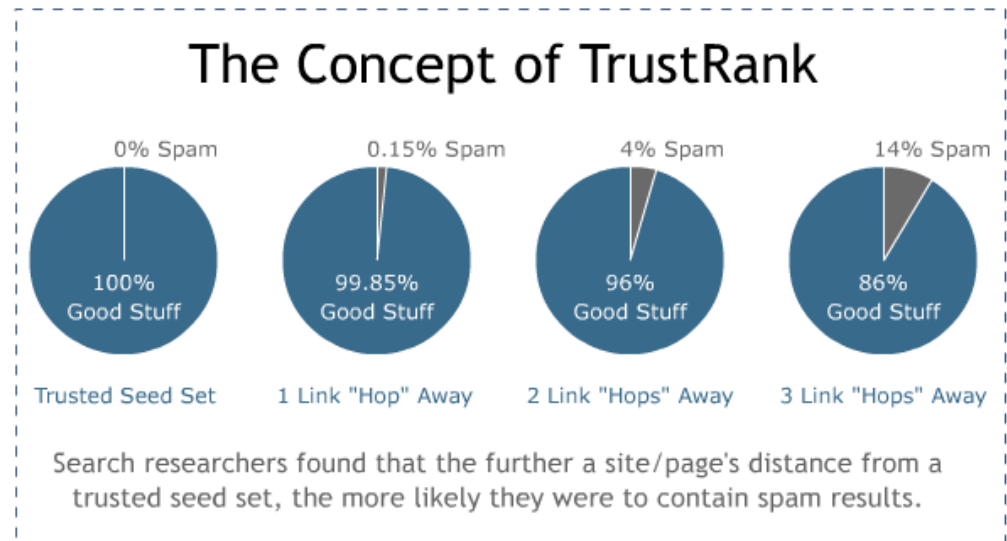
Chống spam bằng hyperlink

- Sử dụng anchor text của hyperlink trở đến trang cần xét.
 - Anchor text được gán trọng số cao hơn, đôi khi xét cả những từ lân cận anchor text.
- Giải thuật PageRank dựa trên chất lượng của liên kết trở đến trang mục tiêu.
 - Trang trở đến trang mục tiêu phải có danh tiếng hoặc có điểm PageRank cao.



Chống spam bằng phân tích liên kết

- Các phương pháp phân tích liên kết được sử dụng để tách biệt trang danh tiếng với các hình thức spam.
- Combating web spam with TrustRank (Gyöngyi, Zoltán, Hector Garcia-Molina, and Jan Pedersen, 2004)
 - Chuyên gia nhận diện một vài trang danh tiếng làm tập trang hạt giống. Crawlers mở rộng tìm kiếm từ tập hạt giống, độ tin cậy giảm dần theo khoảng cách giữa tài liệu và tập hạt giống.



Dự đoán trang spam/non-spam

- Chống spam là một hình thức của tác vụ phân lớp: dự đoán một trang có phải là spam hay không
- Các đặc trưng quan trọng để phát hiện nội dung spam
 1. Số từ trong trang
 2. Độ dài từ trung bình
 3. Số từ trong tiêu đề của trang
 4. Tỷ lệ nội dung hiển thị cho người dùng
- Đặc trưng khác
 - Số lượng anchor text, hệ số nén, tỷ lệ từ phổ biến toàn cục independent n-gram likelihoods, conditional n-gram likelihoods,...

Chống spam bằng phân khối trang

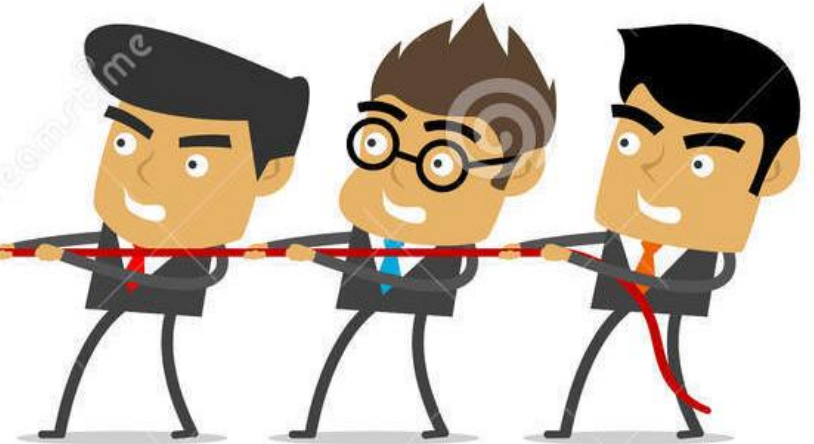
- Liên kết trong những khối nội dung ít quan trọng được gán xác suất chuyển PageRank thấp hơn.
 - Trang được trở bởi những link này có điểm PageRank thấp hơn.
- Chống spam dạng link exchange và honey pot
 - Các liên kết spam nằm trong khối ít quan trọng (ví dụ, cuối trang).
- Chống term spam
 - Từ trong khối ít quan trọng được gán trọng số rất thấp.

Chống spam

Spammers



Search engines



Chống spam là cuộc chiến không ngừng nghỉ.

Tài liệu tham khảo



- Bing Liu. 2007. *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*. Springer Series on Data-Centric Systems and Applications. **Chapter 6**.