

**TÀI LIỆU LÝ THUYẾT KHAI THÁC WEB**

**Chủ đề 3**

**KHAI THÁC NỘI DUNG WEB  
(PHẦN 2)**

Giảng viên: ThS. Lê Ngọc Thành  
Email: [lnthanh@fit.hcmus.edu.vn](mailto:lnthanh@fit.hcmus.edu.vn)

# NỘI DUNG

---

- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Truy vấn thông tin và tìm kiếm Web

---

- Tìm kiếm Web có nguồn gốc từ truy vấn thông tin (information retrieval – IR ), một lĩnh vực nghiên cứu giúp đỡ người sử dụng tìm kiếm thông tin cần thiết từ tập lớn các tài liệu văn bản.



<http://journals.sfu.ca/hypot/index.php/main/article/view/182>

# Truy vấn thông tin và tìm kiếm Web

- Truy vấn thông tin đơn giản nghĩa là tìm một tập các tài liệu mà liên quan đến câu truy vấn của người sử dụng.
- **Thứ hạng các tài liệu** được đánh giá theo độ liên quan của chúng tới câu truy vấn.
- Dạng câu truy vấn được sử dụng phổ biến nhất là một danh sách các **từ khóa**, mà cũng được gọi là thuật ngữ (term).



BUT DAD, THAT IS THE MOST SEARCHED  
KEYWORD ON SEARCH ENGINES...

<http://nilosarraf.com/2011/02/22/background-information-retrieval/>

# Truy vấn thông tin và tìm kiếm Web

- Truy vấn thông tin khác truy vấn dữ liệu (data retrieval)?

Truy vấn thông tin	Truy vấn dữ liệu
Câu truy vấn thường là ngôn ngữ tự nhiên	Câu truy vấn SQL
Thông tin không có cấu trúc hoặc bán cấu trúc	Thông tin được cấu trúc hóa cao
Trên Web thì có siêu liên kết nhưng rất tự do	Có các bảng quan hệ và các quan hệ được kiểm soát chặt chẽ

# Truy vấn thông tin và tìm kiếm Web

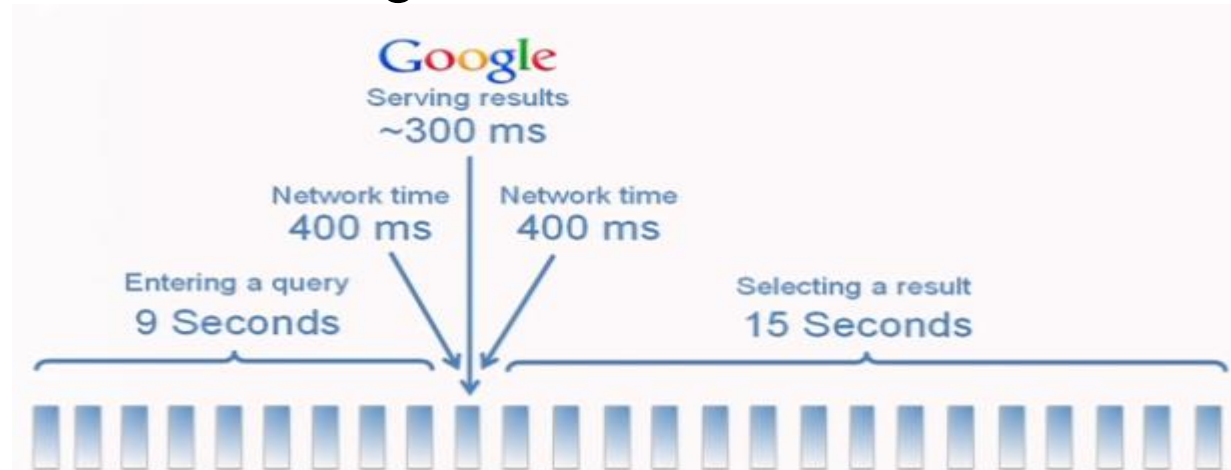
---

- Tìm kiếm Web là ứng dụng quan trọng nhất của IR.
- Tuy nhiên, tìm kiếm Web không đơn giản là một ứng dụng đơn thuần của mô hình IR truyền thống. Nó sử dụng một số kết quả của IR nhưng nó cũng có những kĩ thuật riêng của nó và đề xuất rất nhiều bài toán mới cho lĩnh vực nghiên cứu IR.

*Tại sao?*

# Truy vấn thông tin và tìm kiếm Web

- Bởi vì:
  - Hiệu suất là vấn đề tối cao của tìm kiếm Web, nhưng chỉ là thứ hai trong hệ thống IR vì tập hợp tài liệu trong hệ thống IR không quá lớn. Tuy nhiên, số lượng trang trên Web thì thật khổng lồ.
  - Ví dụ, Google đánh chỉ mục nhiều hơn 8 tỉ trang (2011). Người sử dụng Web cũng đòi hỏi sự hồi đáp cực nhanh. Nếu truy vấn không hiệu suất, ít người sẽ sử dụng nó.



<http://www.plumbersurplus.com/Blog/?tag=/search+engines>



# Truy vấn thông tin và tìm kiếm Web

---

- Bởi vì:
  - Trang web khác so với tài liệu văn bản.
    - Siêu liên kết (hyperlink) và chuỗi kí tự liên kết (anchor text): cực kì quan trọng trong thuật toán xếp hạng tìm kiếm.
    - Một trang Web có những phần và khối khác nhau như tựa đề, siêu dữ liệu, phần thân, .... Một số phần thì quan trọng và cái khác thì không (ví dụ như quảng cáo, điều khoản cá nhân, bản quyền,...). Phát hiện các khối nội dung chính của trang Web sẽ có ích cho tìm kiếm Web bởi vì các thuật ngữ xuất hiện trong những khối này có nhiều quan trọng hơn.



# Truy vấn thông tin và tìm kiếm Web

- Bởi vì:

- Dữ liệu rác (spamming): để cải thiện thứ hạng của một số trang, spamming thường được sử dụng để tăng thứ tự xếp hạng. Chính nó sẽ gây hại cho chất lượng của kết quả tìm kiếm và trải nghiệm tìm kiếm của người sử dụng.

- **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

<http://www.seobook.com/archives/01668.shtml>

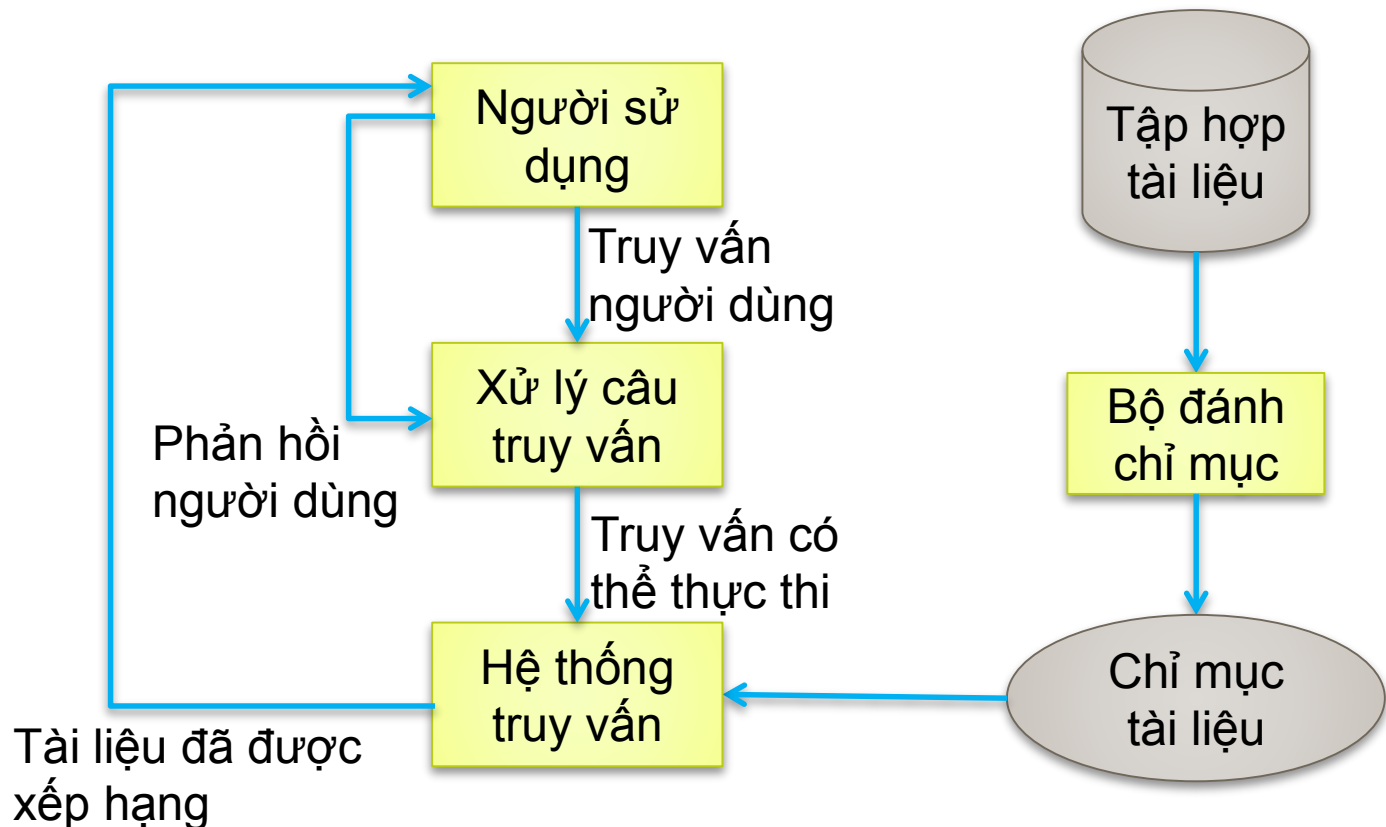
# NỘI DUNG

---

- Khái niệm khai thác nội dung Web
- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Hệ thống truy vấn thông tin

- Truy vấn thông tin (IR) là nghiên cứu để giúp đỡ người sử dụng để tìm thông tin trùng khớp với nhu cầu thông tin của họ.



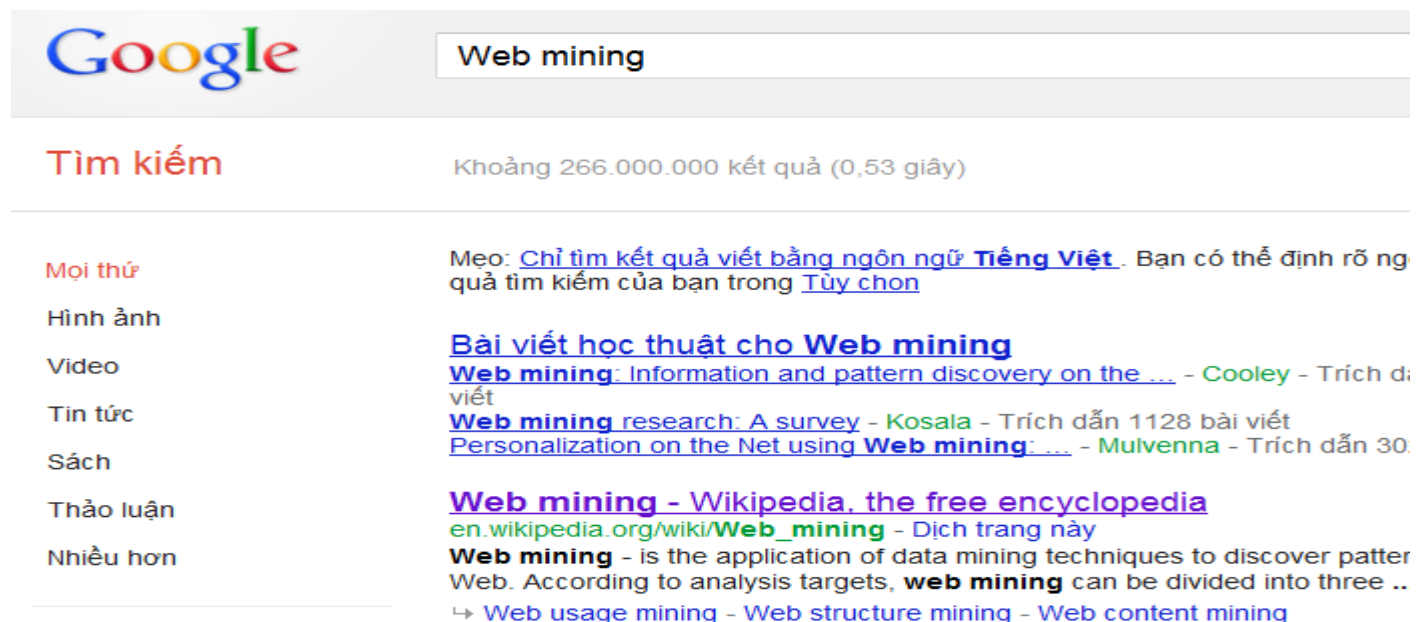
# Truy vấn người sử dụng

---

- Một truy vấn người sử dụng thể hiện nhu cầu thông tin của người sử dụng, nó có các dạng:
  - Truy vấn từ khóa
  - Truy vấn luận lý
  - Truy vấn cụm
  - Truy vấn lân cận
  - Truy vấn toàn tài liệu
  - Câu hỏi ngôn ngữ tự nhiên

# Truy vấn từ khóa

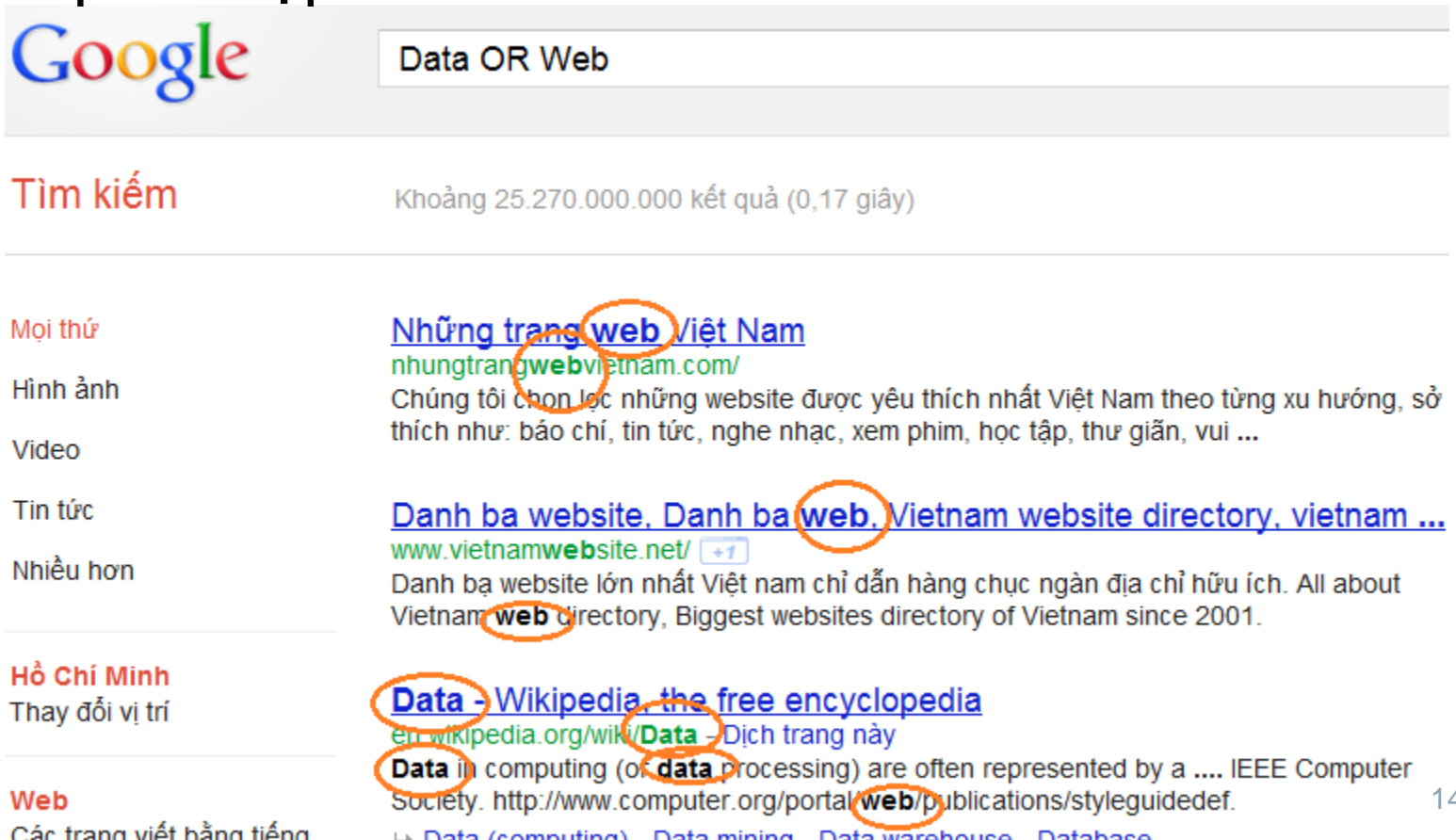
- Truy vấn từ khóa (keyword query): người sử dụng nhấn mạnh thông tin với một danh sách của các từ khóa hướng đến tìm các tài liệu mà chứa một số (ít nhất một) hay tất cả các thuật ngữ truy vấn.



The screenshot shows a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "Web mining". Below the search bar, the text "Tìm kiếm" (Search) is on the left, and "Khoảng 266.000.000 kết quả (0,53 giây)" (About 266,000,000 results in 0.53 seconds) is on the right. A vertical list of filters is on the left side: "Mọi thứ" (All), "Hình ảnh" (Images), "Video", "Tin tức" (News), "Sách" (Books), "Thảo luận" (Discussions), and "Nhiều hơn" (More). The main content area on the right displays search results. The first result is a snippet: "Mẹo: [Chỉ tìm kết quả viết bằng ngôn ngữ Tiếng Việt](#). Bạn có thể định rõ ng...  
quả tìm kiếm của bạn trong [Tùy chọn](#)". Below this is a link to "Bài viết học thuật cho Web mining" (Academic articles for Web mining), followed by a snippet: "[Web mining](#): Information and pattern discovery on the ... - Cooley - Trích di...  
viết [Web mining](#) research: A survey - Kosala - Trích dẫn 1128 bài viết  
[Personalization on the Net using Web mining](#): ... - Mulvenna - Trích dẫn 30...". The next result is "Web mining - Wikipedia, the free encyclopedia" with a snippet: "[en.wikipedia.org/wiki/Web\\_mining](#) - Dịch trang này  
**Web mining** - is the application of data mining techniques to discover patter...  
Web. According to analysis targets, **web mining** can be divided into three ..  
↳ [Web usage mining](#) - [Web structure mining](#) - [Web content mining](#)".

# Truy vấn luận lý

- Truy vấn luận lý (Boolean query): người sử dụng có thể sử dụng toán tử luận lý như AND, OR và NOT để hình thành truy vấn phức tạp.



The screenshot shows a Google search interface. The search bar contains the text "Data OR Web". Below the search bar, the text "Tìm kiếm" (Search) is on the left, and "Khoảng 25.270.000.000 kết quả (0,17 giây)" (About 25,270,000,000 results in 0.17 seconds) is on the right. The search results are displayed in a list format. The first result is titled "Những trang web Việt Nam" (Vietnam websites) with the URL "nhungtrangwebvietnam.com/". The second result is titled "Danh bạ website. Danh bạ web. Vietnam website directory. vietnam ..." with the URL "www.vietnamwebsite.net/". The third result is titled "Data - Wikipedia, the free encyclopedia" with the URL "en.wikipedia.org/wiki/Data". The fourth result is titled "Data in computing (or data processing) are often represented by a .... IEEE Computer Society." with the URL "http://www.computer.org/portals/web/publications/styleguidedef". The words "web", "Data", and "web" are circled in orange in the original image. On the left side of the search results, there are filters: "Mọi thứ" (Everything), "Hình ảnh" (Images), "Video", "Tin tức" (News), "Nhiều hơn" (More), "Hỗ Chí Minh" (Ho Chi Minh), "Thay đổi vị trí" (Change location), and "Web".

Google

Data OR Web

Tìm kiếm

Khoảng 25.270.000.000 kết quả (0,17 giây)

Mọi thứ

Hình ảnh

Video

Tin tức

Nhiều hơn

Hỗ Chí Minh

Thay đổi vị trí

Web

Những trang web Việt Nam  
nhungtrangwebvietnam.com/

Chúng tôi chọn lọc những website được yêu thích nhất Việt Nam theo từng xu hướng, sở thích như: báo chí, tin tức, nghe nhạc, xem phim, học tập, thư giãn, vui ...

Danh bạ website. Danh bạ web. Vietnam website directory. vietnam ...  
www.vietnamwebsite.net/ +1

Danh bạ website lớn nhất Việt nam chỉ dẫn hàng chục ngàn địa chỉ hữu ích. All about Vietnam web directory, Biggest websites directory of Vietnam since 2001.

Data - Wikipedia, the free encyclopedia  
en.wikipedia.org/wiki/Data Dịch trang này

Data in computing (or data processing) are often represented by a .... IEEE Computer Society. http://www.computer.org/portals/web/publications/styleguidedef.

# Truy vấn cụm

- Truy vấn cụm (phrase query): một truy vấn bao gồm tuần tự của các từ tạo thành một cụm. Mỗi tài liệu được trả về phải chứa ít nhất một thực thể của cụm.



"Web mining techniques and applications"

Tìm kiếm

Khoảng 53.500 kết quả (0,44 giây)

Mọi thứ

Hình ảnh

Video

Tin tức

Nhiều hơn

Hồ Chí Minh

Thay đổi vị trí

Web

Các trang viết bằng tiếng

[web-mining-and-social-networking-techniques-and-applications](#)

[www.scribd.com/.../web-mining-and-social-networkin...](#) - Dịch trang này

24 Mar 2011 - 3.0.2 The Evolution of Social Networks ..... 67.

Part I **Web Mining: Techniques and Applications** I Web Content Mining .

[Link Analysis in Web Mining: Techniques and Applications ...](#)

[onlinelibrary.wiley.com > ... > Book Home](#) - Dịch trang này

viết bởi P Desikan - Trích dẫn 1 bài viết - Bài viết có liên quan

20 Jul 2010 - How to Cite: Desikan, P., DeLong, G. and Srivastava, J. (2010) Link Analysis in **Web Mining: Techniques and Applications** in Semantic ...

[Advances in Web Mining and Web Usage Analysis](#)

[www.springer.com > ... > Artificial Intelligence](#) - Dịch trang này

The enhanced papers show that **Web mining techniques and applications** have to more effectively integrate a variety of types of data across multiple channels ...



# Truy vấn lân cận

---

- Truy vấn lân cận (proximity query): là một phiên bản nới lỏng của truy vấn cụm và có thể xem là một sự liên kết giữa thuật ngữ và cụm. Truy vấn lân cận tìm kiếm các thuật ngữ truy vấn bên trong lân cận gần với cái khác.
- Tính gần được sử dụng như một nhân tố trong xếp hạng tài liệu trả về.
- Ví dụ, một tài liệu chứa tất cả các thuật ngữ truy vấn gần với nhau được xem như liên quan hơn một tài liệu mà các thuật ngữ truy vấn nằm xa nhau.

# Truy vấn toàn tài liệu

- Truy vấn toàn tài liệu (full document query): câu truy vấn là một tài liệu đầy đủ, người sử dụng muốn tìm các tài liệu khác giống với tài liệu được truy vấn.

Google Web Images Groups News Froogle<sup>News</sup> more »  
related:www.consumerreports.org/ Search Advanced Search Preferences

**Web** Results 1 - 30 of about 31 similar to www.consumerreports.org/. (0)

[ConsumerReports.org: Unbiased product Ratings from the experts at ...](#)  
Home Customer service My account, SUBSCRIBE, LOGIN. Autos, Appliances, Electronics & computers, Home & garden, Health & fitness, Personal ...  
www.consumerreports.org/ - 60k - [Cached](#) [Similar pages](#)

[Consumer World: Everything Consumer](#)  
Consumer World is a public service, non-commercial\* guide with over 2000 of the most useful consumer resources. FREE WEEKLY CONSUMER WORLD® EMAIL NEWSLETTER. ...  
www.consumerworld.org/ - 42k - [Cached](#) [Similar pages](#)

[BBB](#)  
CHECK IT OUT: BBB reports provide information on over two million organizations. It is a good idea to check before you invest or give. Business. Charity. ...  
www.bbb.org/ - 13k - [Cached](#) [Similar pages](#)

# Câu hỏi ngôn ngữ tự nhiên



## START's reply

==> what is ipod?

iPod

The iPod is a combination portable digital media player and hard drive from Apple Com. 60 GB of hard drive capacity and color screens capable of playing television shows, vid

An iPod can be connected to a computer with either a [FireWire](#) or [USB](#) port, with

- Câu hỏi ngôn ngữ tự nhiên (natural language question): đây là trường hợp phức tạp nhất, và cũng là trường hợp lý tưởng. Người sử dụng nhấn mạnh thông tin cần như một câu hỏi ngôn ngữ tự nhiên. Hệ thống sau đó sẽ tìm câu trả lời. Tuy nhiên, những truy vấn này rất khó để giải quyết bởi vì không dễ để hiểu ngôn ngữ tự nhiên.

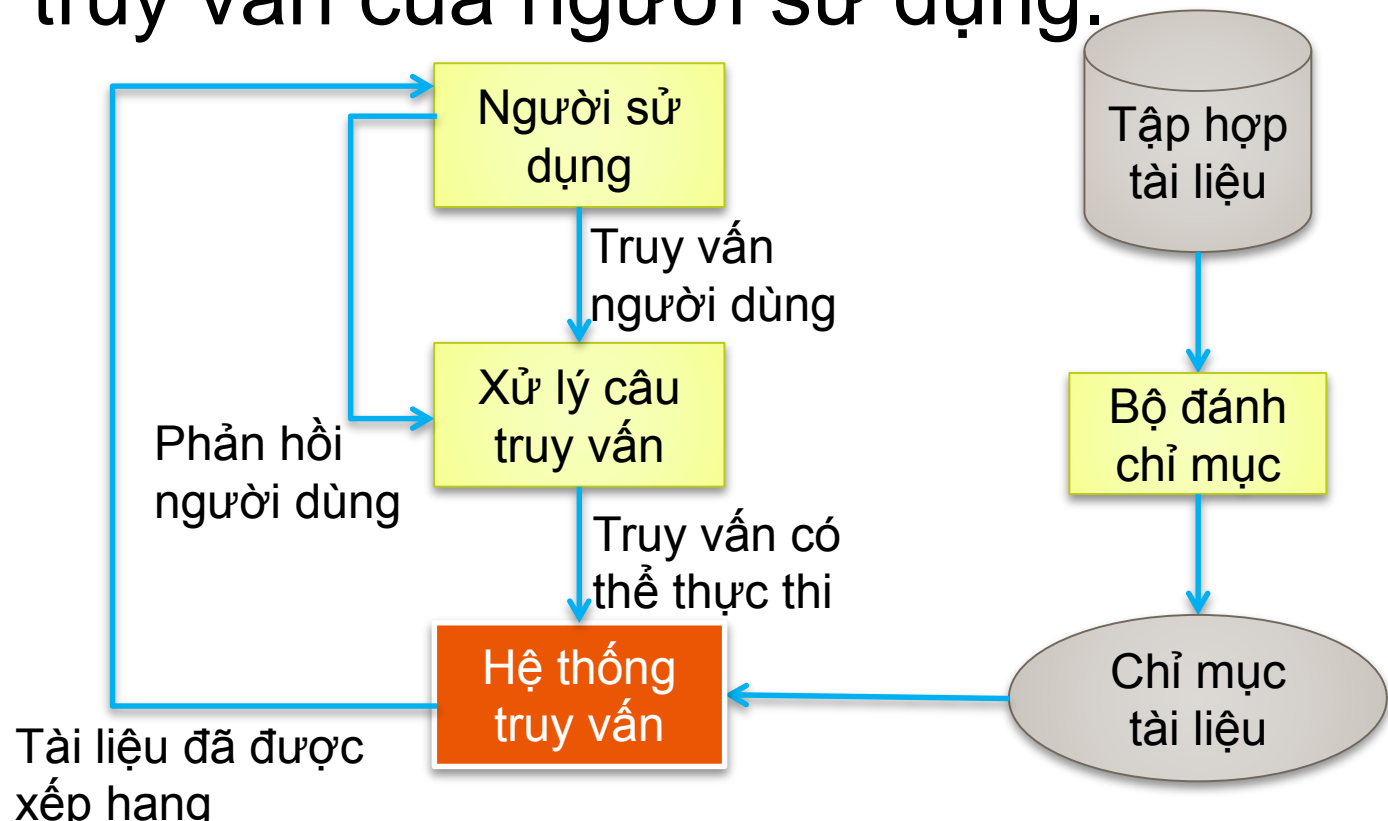
# NỘI DUNG

---

- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Giới thiệu mô hình truy vấn

- Mô hình IR chi phối như thế nào một tài liệu và một truy vấn được thể hiện và sự liên quan của một văn bản với truy vấn của người sử dụng.



# Các mô hình truy vấn

---

- Có 4 mô hình IR:
  - Mô hình luận lý (Boolean model)
  - Mô hình không gian vector (vector space model)
  - Mô hình ngôn ngữ (language model)
  - Mô hình xác suất (probabilistic model).
- 3 mô hình đầu là ba mô hình được sử dụng phổ biến nhất trong hệ thống IR và trên Web.

# Nền tảng của các mô hình

---

- Mặc dù các mô hình này khác nhau, nhưng chúng sử dụng cùng một nền tảng:
  - Mỗi tài liệu hay truy vấn như một túi của từ hay thuật ngữ (bag of words).
  - Thứ tự và vị trí của từ được bỏ qua.
  - Một tài liệu được mô tả bởi một tập các thuật ngữ phân biệt.
  - Nghĩa của các thuật ngữ giúp gợi lại chủ đề chính của tài liệu.
  - Mỗi thuật ngữ được gán một trọng số.



# Thẻ hiện tài liệu và văn bản

- Gọi  $D$  là một tập tài liệu  $\{d_1, d_2, \dots, d_N\}$
- $V = \{t_1, t_2, \dots, t_{|V|}\}$  là tập các thuật ngữ phân biệt trong tập tài liệu, trong đó  $t_i$  là một thuật ngữ.  $|V|$  là kích thước.
- Một trọng số  $w_{ij} > 0$  được gán cho mỗi thuật ngữ  $t_i$  của tài liệu  $d_j$  thuộc  $D$  xác định mức quan trọng của  $t_i$  trong tài liệu  $d_j$ . Thuật ngữ không xuất hiện trong  $d_j$ ,  $w_{ij} = 0$ .
- Vì vậy, mỗi tài liệu  $\mathbf{d}_j$  được thể hiện dưới dạng vector:

$$\mathbf{d}_j = (w_{1j}, w_{2j}, \dots, w_{|V|j})$$

# Ví dụ thể hiện TL và VB

Terms ↓	d1 ↓	d2 ↓	d2 ↓	q ↓
a	1	1	1	0
arrived	0	1	1	0
damaged	1	0	0	0
delivery	0	1	0	0
fire	1	0	0	0
gold	1	0	1	1
in	1	1	1	0
of	1	1	1	0
shipment	1	0	1	0
silver	0	2	0	1
truck	0	1	1	1

**D =**

**q =**

**Tập tài liệu**      **Câu truy vấn**

Một tập hợp của tài liệu được thể hiện như một ma trận. Mỗi thuật ngữ là một thuộc tính và mỗi trọng số là một giá trị thuộc tính. Trong các mô hình truy vấn khác nhau,  $w_{ij}$  được tính toán khác nhau.

# NỘI DUNG

---

- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
    - Mô hình luận lý
    - Mô hình không gian vector
    - Mô hình ngôn ngữ
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Mô hình luận lý

---

- Mô hình luận lý (Boolean model) là một trong những mô hình truy vấn thông tin đơn giản và sớm nhất.
- Mô hình sử dụng khái niệm của so khớp chính xác để so khớp tài liệu với câu truy vấn người sử dụng.
- Cả hai, câu truy vấn và việc truy vấn, đều dựa trên đại số luận lý.

# Mô hình luận lý (tt)

---

- Thể hiện tài liệu:
  - Tài liệu và câu truy vấn được xem như một tập các thuật ngữ.
  - Mỗi thuật ngữ chỉ được xem xét xuất hiện hay không xuất hiện trong một tài liệu.
  - Trọng số của thuật ngữ  $t_i$  trong tài liệu  $d_j$ :

$$w_{ij} = \begin{cases} 1 & \text{nếu } t_i \text{ xuất hiện trong } d_j \\ 0 & \text{ngược lại} \end{cases}$$

# Mô hình luận lý (tt)

---

- Câu truy vấn luận lý:
  - Các thuật ngữ trong câu truy vấn được liên kết một cách logic sử dụng toán tử luận lý AND, OR và NOT.
  - Ví dụ:
    - Truy vấn  $((x \text{ AND } y) \text{ AND } (\text{NOT } z))$  nói rằng một tài liệu tìm được phải chứa cả hai thuật ngữ  $x$  và  $y$  nhưng không chứa  $z$ .
    - Truy vấn  $(x \text{ OR } y)$  chỉ ra rằng ít nhất một trong các thuật ngữ này phải ở trong mỗi tài liệu truy vấn được.

# Mô hình luận lý (tt)

---

- Truy vấn tài liệu:
  - Với một câu truy vấn luận lý, hệ thống sẽ truy tìm các tài liệu làm cho câu truy vấn đúng về mặt logic.
  - Vì vậy, việc truy vấn dựa trên tiêu chí quyết định nhị phân, tài liệu hoặc liên quan hoặc không liên quan. Đây được gọi là ***so khớp chính xác*** (exact matching)



Bất lợi?



# Ví dụ mô hình luận lý

---

- Cho tập  $D = \{d_1, d_2, d_3\}$ 
  - $d_1 = \{\text{Bayes' Principle, probability}\}$
  - $d_2 = \{\text{probability, decision-making}\}$
  - $d_3 = \{\text{probability, Bayesian Epistemology}\}$
- Câu truy vấn:
  - $q = \text{probability AND decision-making}$
- Tập chứa thuật ngữ đầu:  $S_1 = \{d_1, d_2, d_3\}$
- Tập chứa thuật ngữ thứ hai:  $S_2 = \{d_2\}$
- Nên tài liệu truy vấn được sẽ là  $\{d_1, d_2, d_3\} \cap \{d_2\} = \{d_2\}$

# NỘI DUNG

---

- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
    - Mô hình luận lý
    - **Mô hình không gian vector**
    - Mô hình ngôn ngữ
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Mô hình không gian vector

---

- Mô hình này là mô hình nổi tiếng và được sử dụng rộng rãi nhất trong IR.
- Thể hiện tài liệu:
  - Thể hiện như một vector trọng số.
  - Mỗi trọng số được tính dựa trên lược đồ TF (term frequency) hoặc lược đồ TFXIDF (term frequency x inverse document frequency).

# Mô hình không gian vector (tt)

---

- Lược đồ thuật ngữ phổ biến (TF): trọng số của một thuật ngữ  $t_i$  trong tài liệu  $d_j$  là số lần mà  $t_i$  xuất hiện trong tài liệu  $d_j$ , kí hiệu  $f_{ij}$ .
  - Lược đồ TF không xem xét trường hợp một thuật ngữ xuất hiện trong nhiều tài liệu của tập hợp. Những thuật ngữ như vậy có thể không rõ ràng.
- Lược đồ TF x IDF: TF vẫn là độ phổ biến thuật ngữ và IDF là độ phổ biến của tài liệu đảo (inverse document frequency).

# Mô hình không gian vector (tt)

- Độ TF của thuật ngữ  $t_i$  trong tài liệu  $d_j$ :

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \dots, f_{|V|j}\}}$$

$f_{ij}$  là đếm độ phổ biến của thuật ngữ  $t_i$  trong tài liệu  $d_j$ .

- Độ phổ biến tài liệu đảo (kí hiệu là  $idf_i$ ) của thuật ngữ  $t_i$  được cho bởi:

$$idf_i = \log\left(\frac{N}{df_i}\right)$$

- Nếu một thuật ngữ xuất hiện trong nhiều tài liệu, nó có thể không quan trọng hay không rõ ràng.
- Trọng số thuật ngữ TF x IDF cuối cùng được cho bởi công thức:

$$wi_j = tf_{ij} \times idf_i$$

# Ví dụ TFxIDF

---

- Giả sử có  $N = 10$  triệu tài liệu.
- Tài liệu  $d_1$  có từ “ipad” xuất hiện 3 lần, từ có số lần xuất hiện nhiều nhất là 100.

$$tf_{\text{ipad}} = 3/100 = 0.03$$

- Trong toàn bộ  $N$  tài liệu có 1000 tài liệu chứa từ “ipad”. Nên:

$$idf_{\text{ipad}} = \log(10\,000\,000/1\,000) = 4$$

- Vậy trọng số của từ “ipad” trong tài liệu  $d_1$  là:  
 $0.03 \times 4 = 0.12$

# Mô hình không gian vector (tt)

- Câu truy vấn:
  - Giống với thể hiện tài liệu.
  - Trọng số cũng được tính tương tự. Trong một số công trình có đề xuất một số cách tính khác. Ví dụ (Salton và Buckley):

$$w_{iq} = \left( 0.5 + \frac{0.5 * f_{iq}}{\max\{f_{1q}, f_{2q}, \dots, f_{|v|q}\}} \right) \times \log\left(\frac{N}{df_i}\right)$$

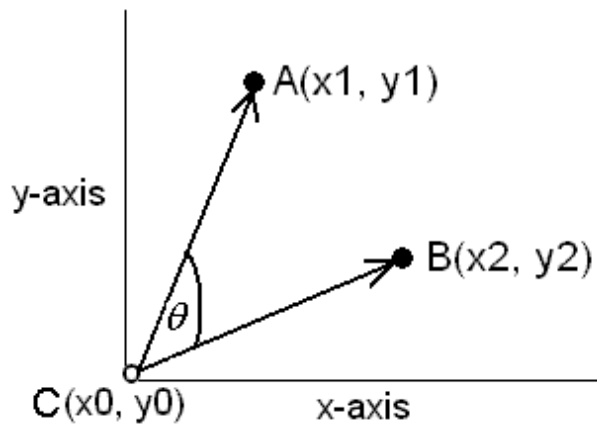


# Mô hình không gian vector (tt)

- Truy vấn tài liệu và xếp hạng:
  - Tính toán độ tương tự của truy vấn  $q$  đến mỗi tài liệu  $d_j$  trong tập hợp tài liệu  $D$ . Có rất nhiều cách tính.
  - Một trong số đó là độ tương tự cosin. Độ đo này là cosin của góc hợp giữa vector truy vấn  $q$  và vector tài liệu  $d_j$ :

$$\cosin(d_j, q) = \frac{\langle d_j \cdot q \rangle}{\|d_j\| \times \|q\|}$$

$$= \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$



# Ví dụ độ đo cosin

- So sánh độ tương quan giữa tài liệu d1 và câu truy vấn sau:

d1: *Julie loves me more than Linda loves me*

Truy vấn: *Jane likes me more than Julie loves me*

- Ta đếm số từ xuất hiện trong d1 và câu truy vấn: me 2 2, Julie 1 1, likes 0 1, loves 2 1, Jane 0 1, Linda 1 0, than 1 1, more 1 1.
- Giả sử vector trọng số chính là đếm xuất hiện của các từ, ta có:  
d1: [2, 1, 0, 2, 0, 1, 1, 1]  
Truy vấn: [2, 1, 1, 1, 1, 0, 1, 1]
- Độ đo cosin hợp bởi hai vector này là 0.822.

# Mô hình không gian vector (tt)

---

- Truy vấn tài liệu và xếp hạng:
  - Một độ đo khác có nhiều hiệu quả hơn độ đo cosin đối với câu truy vấn ngắn là Okapi

# NỘI DUNG

---

- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
    - Mô hình luận lý
    - Mô hình không gian vector
    - **Mô hình ngôn ngữ**
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Mô hình ngôn ngữ

---

- Mô hình ngôn ngữ thống kê (gọi tắt là mô hình ngôn ngữ) dựa trên xác suất và lý thuyết thống kê.
- Đầu tiên ước lượng một mô hình ngôn ngữ cho mỗi tài liệu và sau đó xếp hạng tài liệu bởi khả năng câu truy vấn được cho bởi mô hình ngôn ngữ.
- Truy vấn thông tin sử dụng mô hình ngôn ngữ đầu tiên được đưa ra bởi Ponte và Croft.

# Mô hình ngôn ngữ (tt)

---

- Xét câu truy vấn  $q$  là tuần tự của các thuật ngữ,  $q = q_1 q_2 \dots q_m$  và tập hợp  $D$  là tập các tài liệu,  $D = \{d_1, d_2, \dots, d_N\}$ .
- Để xếp hạng tài liệu, chúng ta ước lượng xác suất hậu nghiệm  $\Pr(d_j|q)$ . Sử dụng luật Bayes:

$$\Pr(d_j|q) = \frac{\Pr(q|d_j) \Pr(d_j)}{\Pr(q)}$$

- $\Pr(q)$  không cần xét vì nó giống nhau ở mỗi tài liệu.  $\Pr(d_j)$  thường được xem như đồng nhất và vì vậy không ảnh hưởng đến xếp hạng. Chúng ta chỉ cần tính toán  $\Pr(q|d_j)$ .

# Mô hình ngôn ngữ (tt)

- Mô hình ngôn ngữ được sử dụng ở đây dựa trên đơn từ (unigram). Mô hình xem mỗi thuật ngữ (từ) được phát sinh một cách độc lập, về bản chất là một phân phối đa thức trên từ:

$$\Pr(q = q_1 q_2 \dots q_m | d_j) = \prod_{i=1}^m \Pr(q_i | d_j) = \prod_{i=1}^{|V|} \Pr(t_i | d_j)^{f_{iq}}$$

Trong đó  $f_{iq}$  là số lần mà thuật ngữ  $t_i$  xảy ra trong  $q$ , và  $\sum_{i=1}^{|V|} \Pr(t_i | d_j) = 1$ .

- Bài toán truy vấn được chuyển thành ước lượng  $\Pr(t_i | d_j)$ :

$$\Pr(t_i | d_j) = \frac{f_{ij}}{|d_j|}$$

$f_{ij}$  là số lượng lần thuật ngữ  $t_i$  xuất hiện trong tài liệu  $d_j$ ,  $|d_j|$  là tổng số từ trong  $d_j$ .

# Ví dụ phân phối đa thức trên unigram

Bảng ước lượng xác suất  $\Pr(t_i|d_1)$  và  $\Pr(t_i|d_2)$

Thuật ngữ	Xác suất trong $d_1$	Xác suất trong $d_2$
a	0.1	0.3
world	0.2	0.1
likes	0.05	0.03
we	0.05	0.02
share	0.3	0.2
...	...	...

Câu truy vấn  $q$ : *world share*

Vậy,  $\Pr(q|d_1) = \Pr(\text{world}|d_1) \times \Pr(\text{share}|d_1) = 0.2 \times 0.3 = 0.06$

$\Pr(q|d_2) = \Pr(\text{world}|d_2) \times \Pr(\text{share}|d_2) = 0.1 \times 0.2 = 0.02$



# Mô hình ngôn ngữ (tt)

---

- Một vấn đề với ước lượng này là một thuật ngữ không xuất hiện trong  $d_j$  có xác suất là 0.
- Kỹ thuật làm mịn (smoothing) được dùng điều chỉnh xác suất 0. Độ smoothing phụ thêm là:

$$Pr_{add}(tj|dj) = \frac{\lambda + f_{ij}}{\lambda|V| + |dj|}$$

- Khi  $\lambda = 1$ , đó là bộ làm mịn Laplace và khi  $0 < \lambda < 1$  đó là làm mịn Lidstone. Ngoài ra còn rất nhiều phương pháp làm mịn khác.

# NỘI DUNG

---

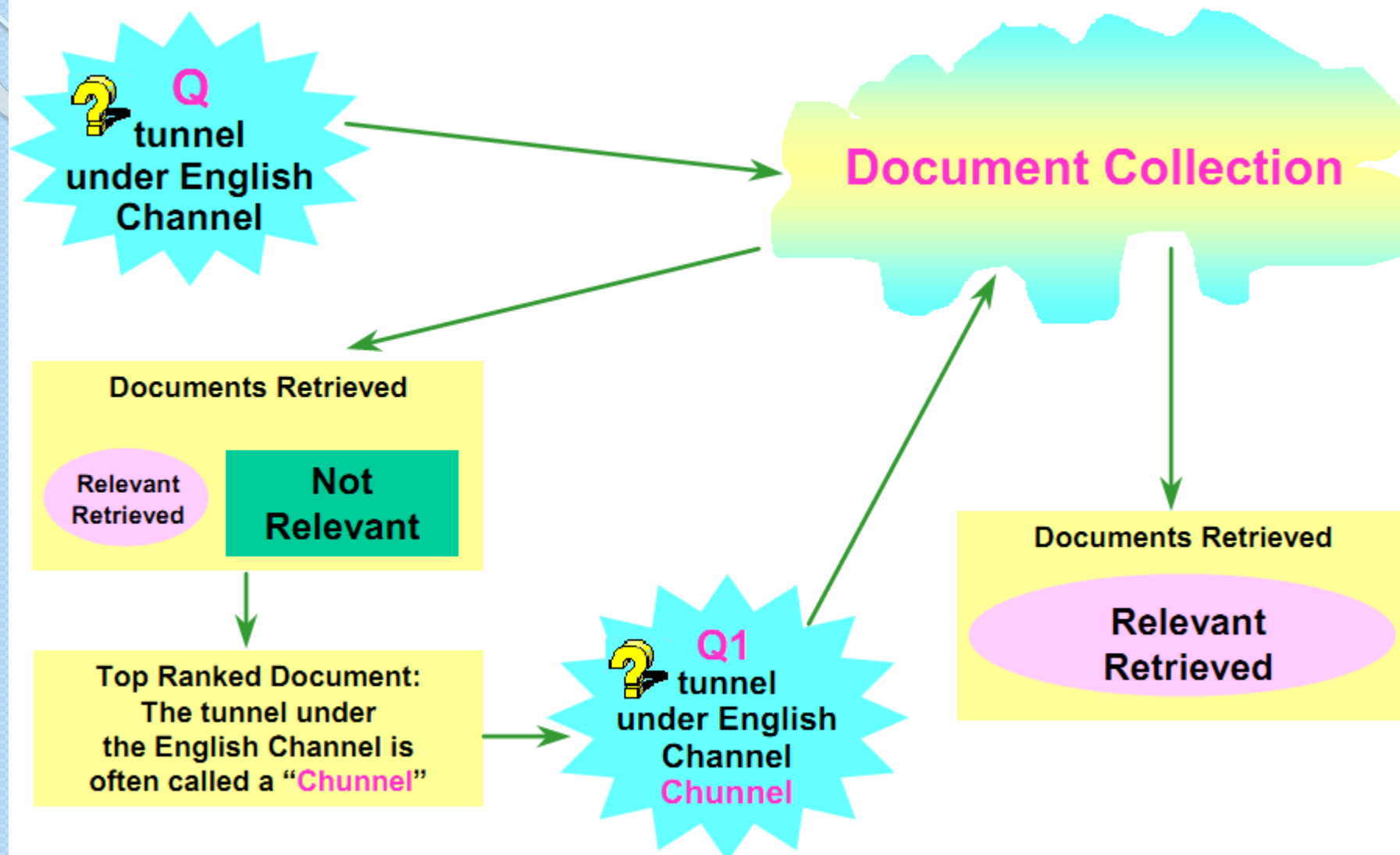
- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Phản hồi liên quan

---

- **Phản hồi liên quan** (relevant feedback - RF) là một tiến trình:
  - *Người sử dụng xác định* các tài liệu liên quan và không liên quan.
  - Sau đó, hệ thống *tạo một truy vấn mở rộng* (bằng cách trích ra một số thuật ngữ từ tài liệu liên quan và không liên quan) và đưa vào vòng truy vấn sau.
  - Hệ thống có thể cũng đưa ra *mô hình phân lớp* để phân lớp các tài liệu.
  - Tiến trình lặp lại cho đến khi *người sử dụng thỏa mãn* với kết quả nhận được.

# Ví dụ RF



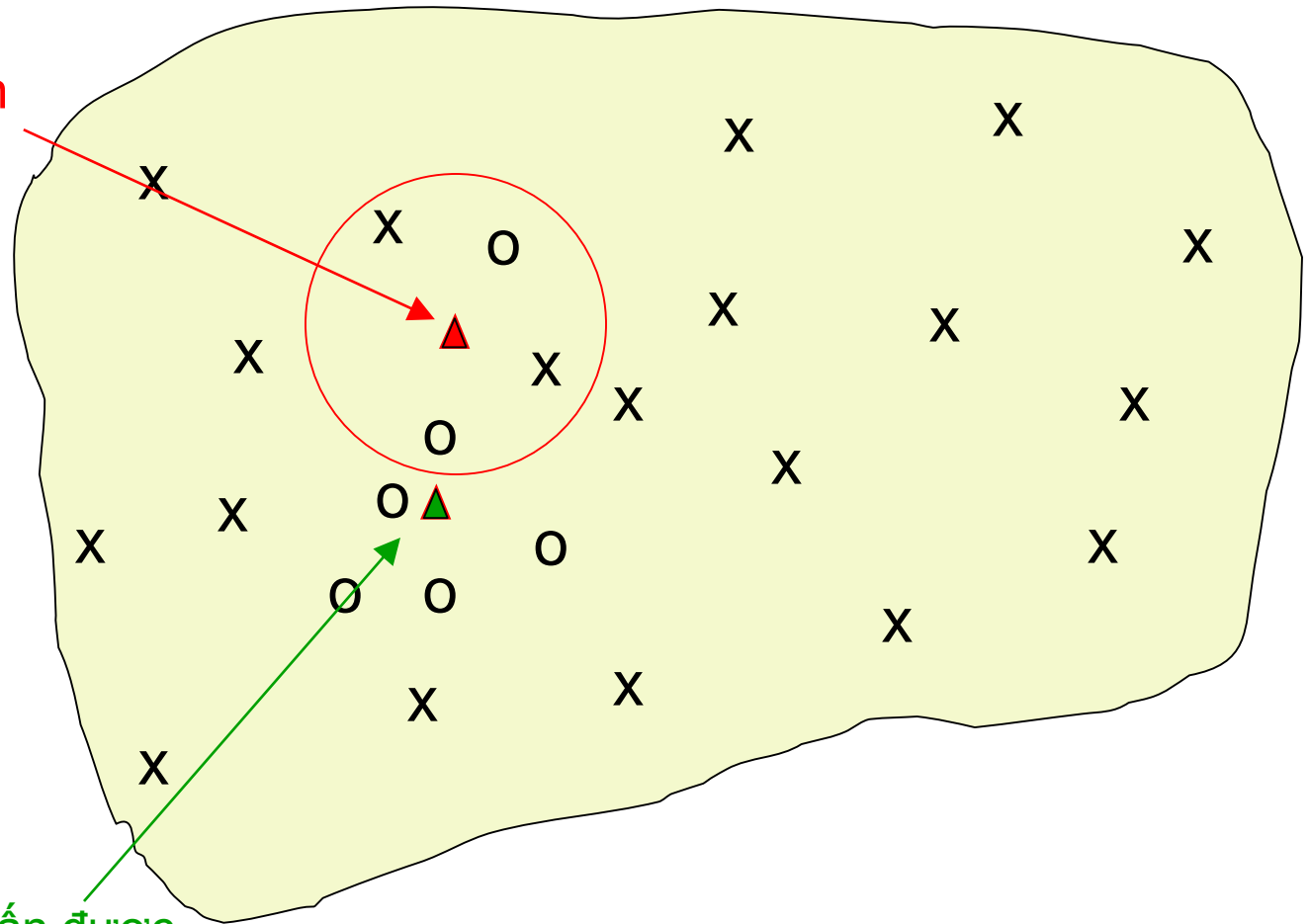
# Các thuật toán RF

---

- **Phương pháp Rocchio**
  - Sử dụng các tài liệu liên quan và không liên quan do người dùng đánh giá để mở rộng câu truy vấn gốc.
  - Câu truy vấn mới sau đó được sử dụng để thực hiện việc truy vấn lại.

# Các thuật toán RF - Rocchio

Truy vấn ban đầu



Truy vấn được mở rộng

x tài liệu không liên quan  
o tài liệu liên quan

# Các thuật toán RF– Rocchio(tt)

---

- Vector truy vấn gốc là  $q$
- Tập các tài liệu liên quan  $D_r$
- Tập các tài liệu không liên quan là  $D_{ir}$
- Câu truy vấn mở rộng  $q_e$  được tính:

$$q_e = \alpha q + \frac{\beta}{|D_r|} \sum_{d_r \in D_r} d_r - \frac{\gamma}{|D_{ir}|} \sum_{d_{ir} \in D_{ir}} d_{ir}$$

trong đó  $\alpha$ ,  $\beta$  và  $\gamma$  là các tham số được rút ra từ thực nghiệm.

# Các thuật toán RF– Rocchio(tt)

---

- Nhận xét:
  - *Câu truy vấn gốc  $q$  vẫn sử dụng lại* vì nó phản ánh trực tiếp nhu cầu thông tin của người dùng.
  - *Phép cộng làm tăng vector truy vấn gốc  $q$*  với những thuật ngữ thêm từ các tài liệu liên quan.
  - *Phép trừ làm giảm ảnh hưởng của những thuật ngữ không rõ ràng* (xuất hiện hiện ở cả hai tài liệu liên quan và không liên quan) và các thuật ngữ chỉ xuất hiện trong tài liệu không liên quan.



# Các thuật toán RF– Học Máy

---

- **Phương pháp học máy:**
  - Do có một tập các tài liệu liên quan và không liên quan, chúng ta có thể hình thành nên mô hình phân lớp từ chúng.
    - Bài toán phản hồi liên quan trở thành bài toán học máy.
  - Bất kì phương pháp học giám sát nào đều có thể sử dụng được, ví dụ phân lớp naïve Bayesian hay SVM.
  - Không cần so sánh độ tương tự với câu truy vấn gốc.

# Các thuật toán RF– Học Máy (tt)

- Một dạng khác của phương pháp Rocchio, gọi là *phương pháp phân lớp Rocchio*, có thể cũng được sử dụng.
- Bộ phân lớp Rocchio hình thành vector mẫu  $c_i$  cho mỗi lớp  $i$  (lớp liên quan hoặc không liên quan) với công thức:

$$c_i = \frac{\alpha}{|D_i|} \sum_{d \in D_i} \frac{d}{||d||} - \frac{\beta}{|D - D_i|} \sum_{d \in D - D_i} \frac{d}{||d||}$$

trong đó  $D_i$  là tập các tài liệu của lớp  $i$  và  $\alpha$ ,  $\beta$  là các tham số. Sử dụng lược đồ đánh trọng TFxIDF,  $\alpha = 16$  và  $\beta = 4$  thường có kết quả tốt.

# Các thuật toán RF– Học Máy (tt)

---

- Mỗi tài liệu kiểm thử  $d_t$  sẽ được so sánh với vector mẫu  $c_i$  sử dụng *độ tương tự cosin*. Và  $d_t$  được gán đến phân lớp với giá trị tương tự cao.

## Algorithm

```
1  for each class  $i$  do
2      construct its prototype vector  $c_i$  using Equation
3  endfor
4  for each test document  $d_t$  do
5      the class of  $d_t$  is  $\arg \max_i \text{cosine}(d_t, c_i)$ 
6  endfor
```

# Các thuật toán RF – Học Máy (tt)

---

- Các phương pháp học máy khác:
  - Học từ các mẫu đánh nhãn và chưa đánh nhãn (LU Learning) .
  - Học từ mẫu dương và chưa đánh nhãn (PU learning).
  - Sử dụng SVM xếp hạng (ranking SVM)
  - Mô hình ngôn ngữ (language model)

# Phản hồi liên quan ....giả

---

- Phản hồi liên quan giả (pseudo-relevance feedback): *trích ra một số thuật ngữ từ các tài liệu xếp hạng cao* và thêm chúng vào truy vấn gốc để hình thành câu truy vấn mới cho vòng truy vấn lần sau.
- Tiến trình có thể lặp lại cho đến khi người sử dụng thỏa mãn.

# NỘI DUNG

---

- Truy vấn thông tin và tìm kiếm Web
  - Khái niệm
  - Các mô hình truy vấn thông tin
  - Phản hồi liên quan
  - Đánh giá các độ đo

# Đánh giá các độ đo

---

- $D$ : tập hợp các tài liệu (kích thước  $N$ )
- $q$ : câu truy vấn người dùng
- $R_q = \langle d_1, d_2, \dots, d_N \rangle$ : là thứ tự xếp hạng các tài liệu do hệ thống truy vấn trả về
- $D_q (\subseteq D)$ : tập các tài liệu thực sự liên quan đến truy vấn  $q$

# Đánh giá các độ đo (tt)

---

- Độ phủ (recall) ở hạng  $i$  hay tài liệu  $d_i$ :

$$r(i) = \frac{s_i}{|Dq|}$$

- Độ chính xác (precision) ở hạng  $i$  hay tài liệu  $d_i$ :

$$p(i) = \frac{s_i}{i}$$

với  $s_i$  là số lượng tài liệu thực sự liên quan từ  $d_1$  đến  $d_i$  trong  $R_q$



# Ví dụ đánh giá các độ đo

- Tập D có 20 tài liệu.
- Với  $q$  cho trước, giả sử có 8 tài liệu thực sự liên quan đến  $q$  (kí hiệu dấu '+')

Rank $i$	+/-	$p(i)$	$r(i)$
1	+	$1/1 = 100\%$	$1/8 = 13\%$
2	+	$2/2 = 100\%$	$2/8 = 25\%$
3	+	$3/3 = 100\%$	$3/8 = 38\%$
4	-	$3/4 = 75\%$	$3/8 = 38\%$
5	+	$4/5 = 80\%$	$4/8 = 50\%$
6	-	$4/6 = 67\%$	$4/8 = 50\%$
7	+	$5/7 = 71\%$	$5/8 = 63\%$
8	-	$5/8 = 63\%$	$5/8 = 63\%$
9	+	$6/9 = 67\%$	$6/8 = 75\%$
10	+	$7/10 = 70\%$	$7/8 = 88\%$
11	-	$7/11 = 63\%$	$7/8 = 88\%$
12	-	$7/12 = 58\%$	$7/8 = 88\%$
13	+	$8/13 = 62\%$	$8/8 = 100\%$
14	-	$8/14 = 57\%$	$8/8 = 100\%$
15	-	$8/15 = 53\%$	$8/8 = 100\%$
16	-	$8/16 = 50\%$	$8/8 = 100\%$
17	-	$8/17 = 53\%$	$8/8 = 100\%$
18	-	$8/18 = 44\%$	$8/8 = 100\%$
19	-	$8/19 = 42\%$	$8/8 = 100\%$
20	-	$8/20 = 40\%$	$8/8 = 100\%$

# Đánh giá các độ đo (tt)

---

- Độ chính xác trung bình: để so sánh các thuật toán truy vấn khác nhau trên một query  $q$ , chúng ta cần phải có một *độ chính xác đơn*.

$$p_{avg} = \frac{\sum_{d_i \in D_q} p(i)}{|D_q|}$$

# Ví dụ đánh giá các độ đo (tt)

$$p_{avg} =$$

$$= \frac{100\% + 100\% + 100\% + 80\% + 71\% + 67\% + 70\% + 62\%}{8}$$

$$= 81\%$$

Rank $i$	+/-	$p(i)$	$r(i)$
1	+	1/1 = 100%	1/8 = 13%
2	+	2/2 = 100%	2/8 = 25%
3	+	3/3 = 100%	3/8 = 38%
4	-	3/4 = 75%	3/8 = 38%
5	+	4/5 = 80%	4/8 = 50%
6	-	4/6 = 67%	4/8 = 50%
7	+	5/7 = 71%	5/8 = 63%
8	-	5/8 = 63%	5/8 = 63%
9	+	6/9 = 67%	6/8 = 75%
10	+	7/10 = 70%	7/8 = 88%
11	-	7/11 = 63%	7/8 = 88%
12	-	7/12 = 58%	7/8 = 88%
13	+	8/13 = 62%	8/8 = 100%
14	-	8/14 = 57%	8/8 = 100%
15	-	8/15 = 53%	8/8 = 100%
16	-	8/16 = 50%	8/8 = 100%
17	-	8/17 = 53%	8/8 = 100%
18	-	8/18 = 44%	8/8 = 100%
19	-	8/19 = 42%	8/8 = 100%
20	-	8/20 = 40%	8/8 = 100%

# Đường cong độ chính xác – độ phủ

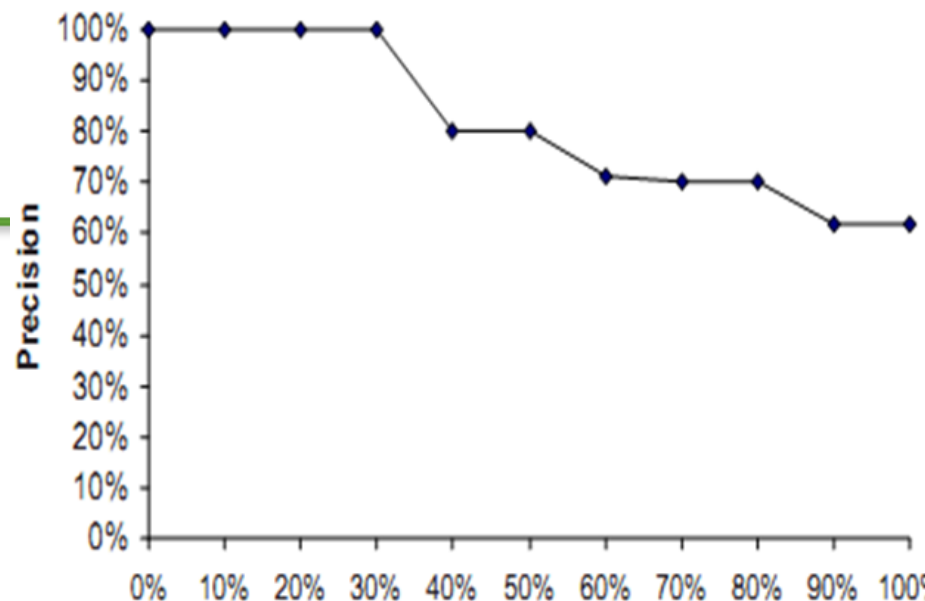
---

- Đường cong độ chính xác – độ phủ: thể hiện sự tương quan giữa độ chính xác và độ phủ.
  - Trong đó, lấy trục biểu diễn độ phủ làm chuẩn với 10 mức  $r_i \in \{0\%, 10\%, 20\%, \dots, 100\%\}$ .
- Vì không có độ chính xác ở từng mức bao phủ cụ thể, chúng ta phải thực hiện phép nội suy độ chính xác:

$$p(r_i) = \max_{r_i \leq r \leq r_{10}} p(r)$$

# Ví dụ

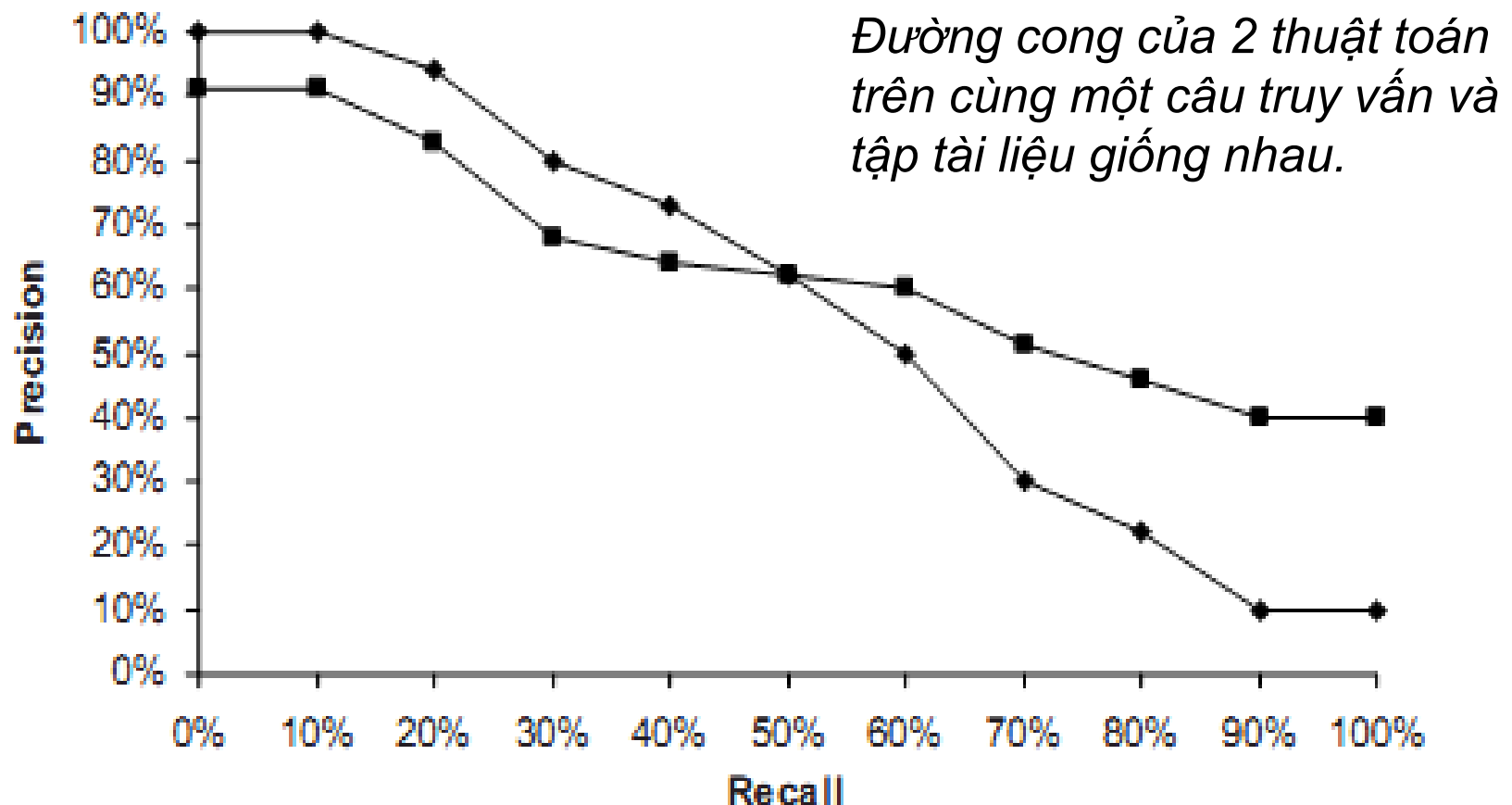
Rank $i$	+/-	$p(i)$	$r(i)$
1	+	1/1 = 100%	1/8 = 13%
2	+	2/2 = 100%	2/8 = 25%
3	+	3/3 = 100%	3/8 = 38%
4	-	3/4 = 75%	3/8 = 38%
5	+	4/5 = 80%	4/8 = 50%
6	-	4/6 = 67%	4/8 = 50%
7	+	5/7 = 71%	5/8 = 63%
8	-	5/8 = 63%	5/8 = 63%
9	+	6/9 = 67%	6/8 = 75%
10	+	7/10 = 70%	7/8 = 88%
11	-	7/11 = 63%	7/8 = 88%
12	-	7/12 = 58%	7/8 = 88%
13	+	8/13 = 62%	8/8 = 100%
14	-	8/14 = 57%	8/8 = 100%
15	-	8/15 = 53%	8/8 = 100%
16	-	8/16 = 50%	8/8 = 100%
17	-	8/17 = 53%	8/8 = 100%
18	-	8/18 = 44%	8/8 = 100%
19	-	8/19 = 42%	8/8 = 100%
20	-	8/20 = 40%	8/8 = 100%



$i$	$p(r_i)$	$r_i$
0	100%	0%
1	100%	10%
2	100%	20%
3	100%	30%
4	80%	40%
5	80%	50%
6	71%	60%
7	70%	70%
8	70%	80%
9	62%	90%
10	62%	100%

Recall

# So sánh các thuật toán



- Nhận xét: độ chính xác của thuật toán này tốt hơn cái kia ở mức phủ thấp nhưng lại kém hơn trong mức phủ cao.

# Đánh giá trên nhiều truy vấn

---

- Độ chính xác toàn cục ở mỗi mức phủ  $r_i$ :

$$\bar{p}(r_i) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} p_j(r_i)$$

với  $Q$  là tập các câu truy vấn

# Nhận xét về độ phủ, độ chính xác

---

- Độ chính xác và độ phủ có sự đánh đổi (trade-off).
- Trên Web,  $D_q$  hầu như không thể xác định bởi vì có quá nhiều trang. Không có  $D_q$ , độ phủ không thể tính toán.
- Trong thực tế, độ phủ không gây ảnh hưởng nhiều cho tìm kiếm Web bởi vì hiếm khi người sử dụng nhìn vào các trang xếp hạng dưới 30.
- Tuy nhiên, độ chính xác thì quan trọng.



# Độ đo khác

---

- Độ đo F (F-score): là trung bình điều hòa của độ chính xác và độ phủ:

$$F(i) = \frac{2}{\frac{1}{r(i)} + \frac{1}{p(i)}} = \frac{2p(i)r(i)}{p(i) + r(i)}$$

# TÀI LIỆU THAM KHẢO

---

- Tài liệu bài giảng môn học
- **Chapter 3.** B. Liu, *Web Data Mining- Exploring Hyperlinks, Contents, and Usage Data*, Springer Series on Data-Centric Systems and Applications, 2007.

# KẾT THÚC PHẦN 2

