

Tài liệu giảng dạy môn Khai thác dữ liệu Web

TRUY VẤN THÔNG TIN

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

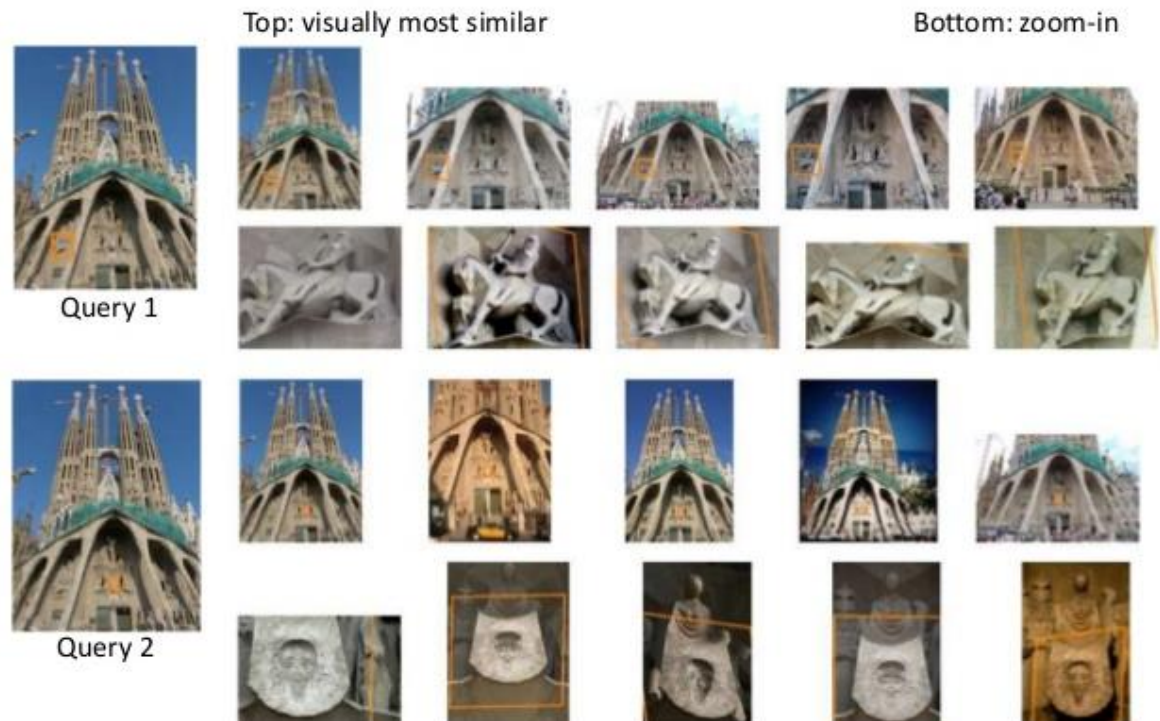
- Truy vấn thông tin và Tìm kiếm Web
- Mô hình truy vấn thông tin
 - Mô hình luận lý Boolean
 - Mô hình không gian vector
 - Mô hình ngôn ngữ
- Phản hồi liên quan
- Độ đo đánh giá



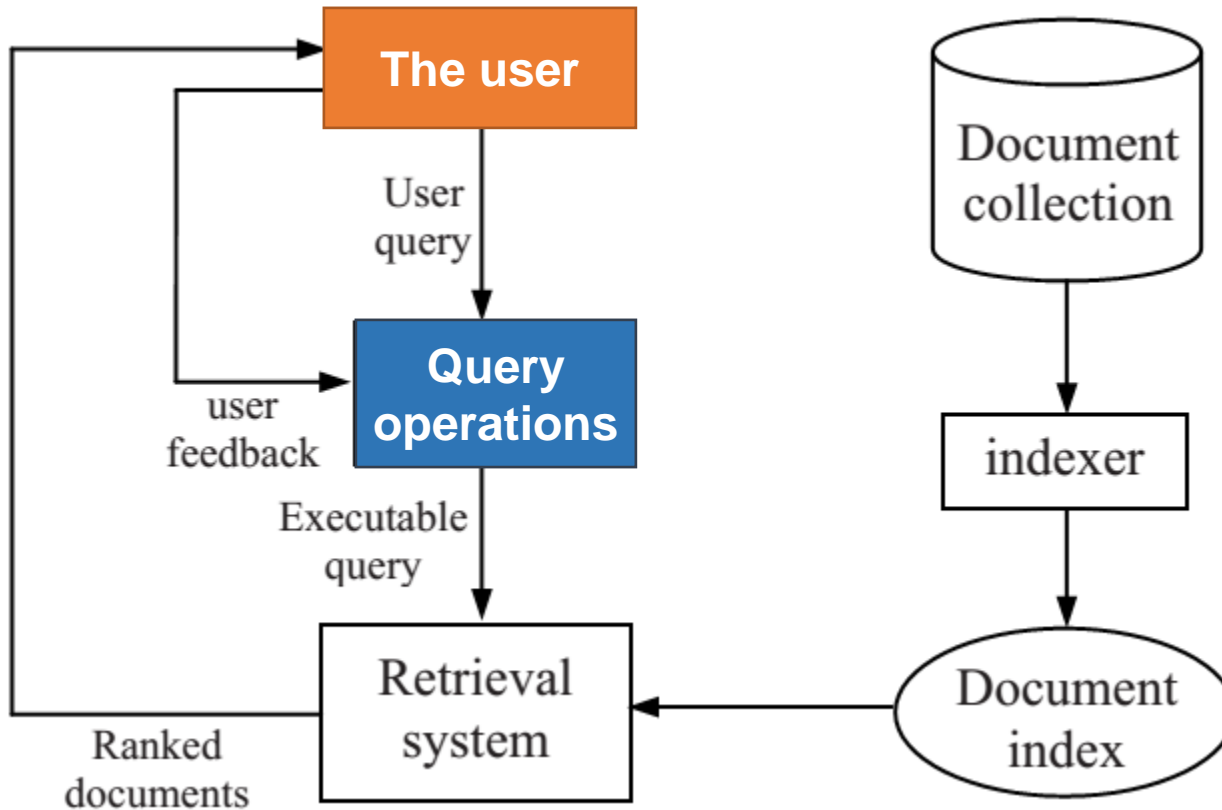
Truy vấn thông tin

Truy vấn thông tin (IR)

- **Truy vấn thông tin** là nghiên cứu nhằm hỗ trợ người dùng tìm kiếm thông tin trùng khớp với nhu cầu thông tin của họ.
 - Thu thập, tổ chức, lưu trữ, truy vấn và phân bố thông tin.

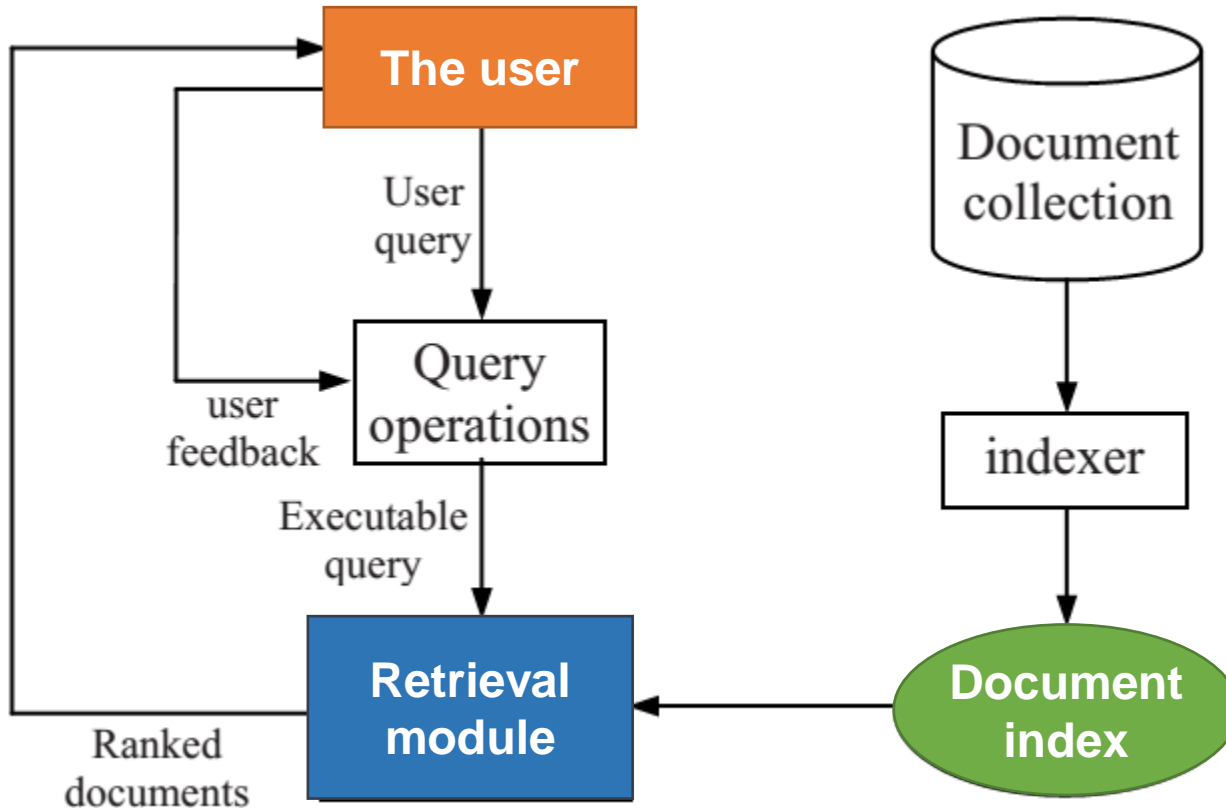


Kiến trúc hệ thống IR tổng quát



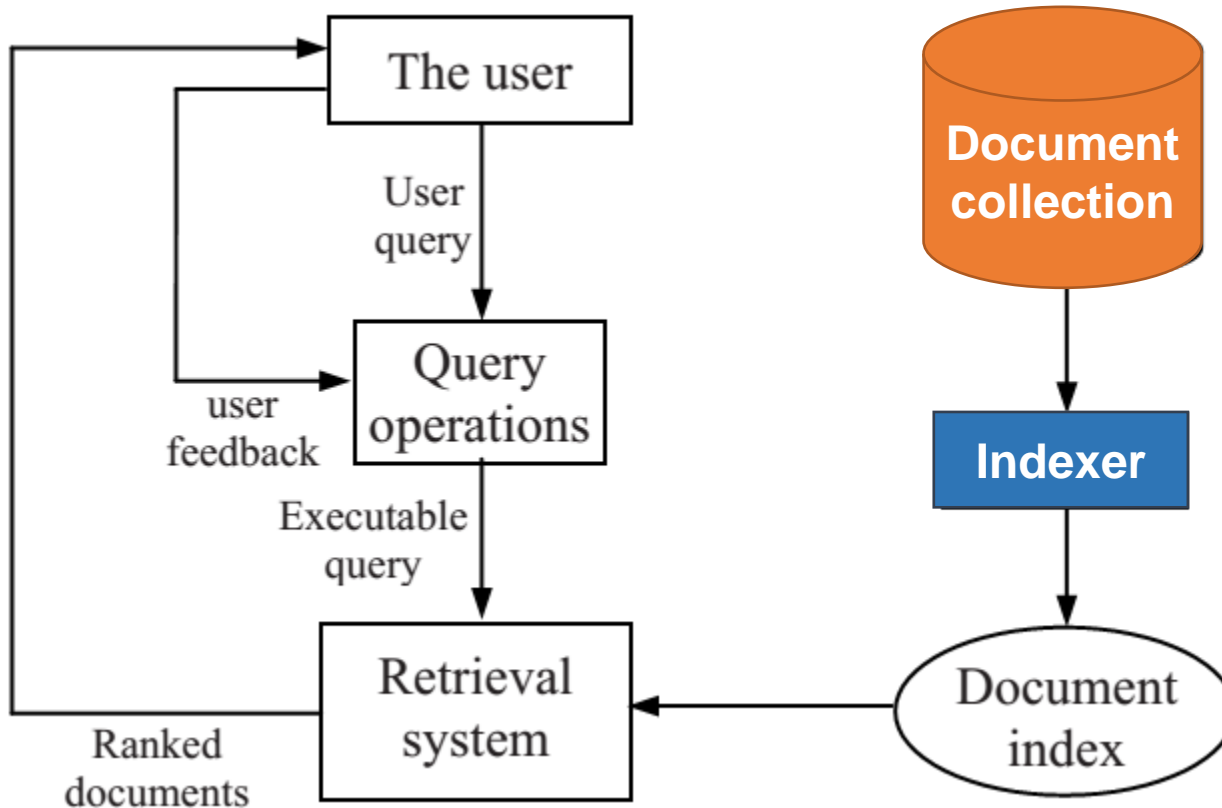
- Người dùng gửi truy vấn (**user query**) đến hệ thống truy vấn (**retrieval system**) thông qua module thực thi truy vấn (**query operations**).

Kiến trúc hệ thống IR tổng quát



- Module truy vấn sử dụng bảng chỉ mục tài liệu (**document index**) để lấy các tài liệu có chứa từ khóa rồi tính điểm số liên quan cho mỗi tài liệu.
- Tài liệu được xếp hạng theo điểm số liên quan và trả về cho người dùng.

Kiến trúc hệ thống IR tổng quát



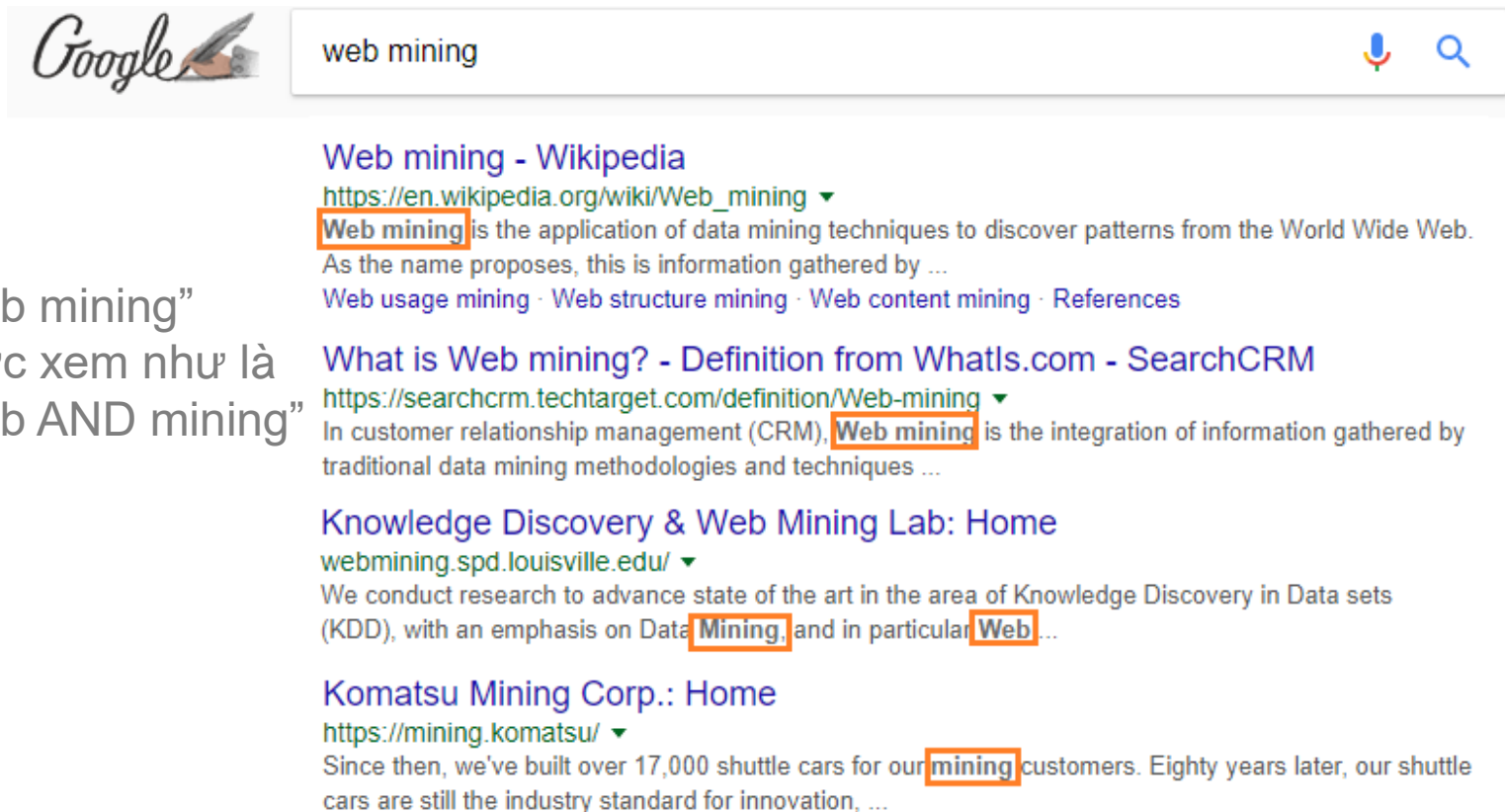
- Tập hợp tài liệu (**document collection**) còn được gọi là cơ sở dữ liệu văn bản (**text database**), được đánh chỉ mục bằng bộ **indexer** để truy vấn hiệu quả.

Information retrieval và Data retrieval

| | Information retrieval | Data retrieval |
|--------------------------|--|---|
| Dữ liệu | Văn bản tự do, phi cấu trúc | Bảng cơ sở dữ liệu, có cấu trúc |
| Truy vấn | Từ khóa, ngôn ngữ tự nhiên | SQL, đại số quan hệ |
| Results | So khớp tương đối, sắp xếp theo độ liên quan | So khớp tuyệt đối, không có thứ tự |
| Khả năng tiếp cận | Người dùng không phải chuyên gia | Người dùng có kiến thức, tiến trình tự động |

User query: Truy vấn từ khóa

- Người dùng cung cấp danh sách từ khóa để tìm tài liệu chứa một vài hoặc mọi từ khóa
- Thứ tự của các từ có thể ảnh hưởng đến kết quả truy vấn.



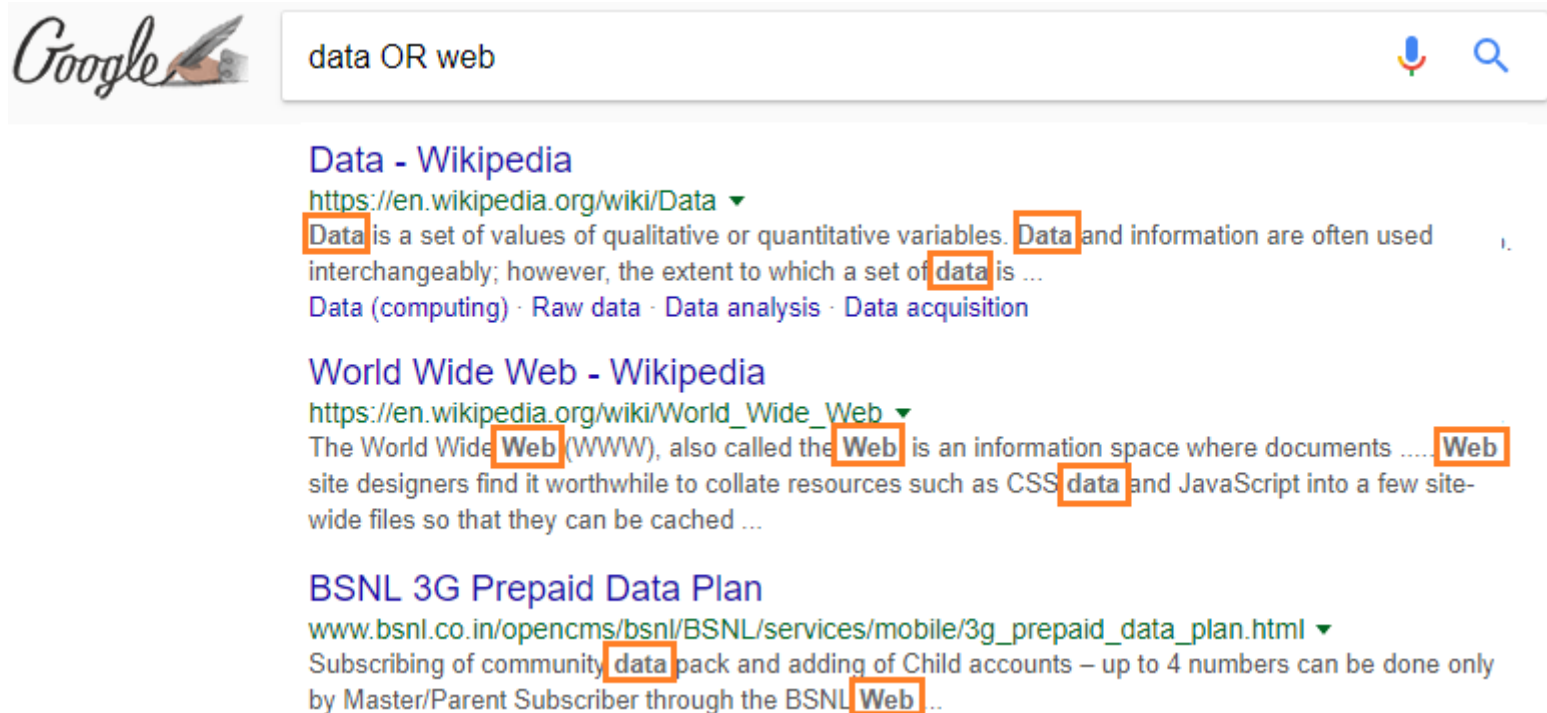
The screenshot shows a Google search interface with the query "web mining". The search results are as follows:

- Web mining - Wikipedia**
https://en.wikipedia.org/wiki/Web_mining ▼
Web mining is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by ...
Web usage mining · Web structure mining · Web content mining · References
- What is Web mining? - Definition from WhatIs.com - SearchCRM**
<https://searchcrm.techtarget.com/definition/Web-mining> ▼
In customer relationship management (CRM), Web mining is the integration of information gathered by traditional data mining methodologies and techniques ...
- Knowledge Discovery & Web Mining Lab: Home**
webmining.spd.louisville.edu/ ▼
We conduct research to advance state of the art in the area of Knowledge Discovery in Data sets (KDD), with an emphasis on Data Mining, and in particular Web ...
- Komatsu Mining Corp.: Home**
<https://mining.komatsu/> ▼
Since then, we've built over 17,000 shuttle cars for our mining customers. Eighty years later, our shuttle cars are still the industry standard for innovation, ...

“Web mining”
được xem như là
“Web AND mining”

User query: Truy vấn luận lý

- Toán tử luận lý (AND, OR, và NOT, v.v.) được sử dụng để xây dựng các truy vấn phức tạp.
- **So khớp tuyệt đối:** Trả về trang có biểu thức luận lý đúng



The screenshot shows a Google search interface with the query "data OR web" entered in the search bar. The search results are displayed below the bar, featuring three entries:

- Data - Wikipedia**
<https://en.wikipedia.org/wiki/Data> ▼
Data is a set of values of qualitative or quantitative variables. Data and information are often used interchangeably; however, the extent to which a set of data is ...
Data (computing) · Raw data · Data analysis · Data acquisition
- World Wide Web - Wikipedia**
https://en.wikipedia.org/wiki/World_Wide_Web ▼
The World Wide Web (WWW), also called the Web is an information space where documents ... Web
site designers find it worthwhile to collate resources such as CSS data and JavaScript into a few site-wide files so that they can be cached ...
- BSNL 3G Prepaid Data Plan**
www.bsnl.co.in/opencms/bsnl/BSNL/services/mobile/3g_prepaid_data_plan.html ▼
Subscribing of community data pack and adding of Child accounts – up to 4 numbers can be done only by Master/Parent Subscriber through the BSNL Web ..

User query: Truy vấn cụm

- Câu truy vấn là chuỗi gồm nhiều từ.
- Tài liệu trả về phải chứa ít nhất một thực thể của cụm.



"web mining techniques and applications"



[Web mining techniques and applications: Literature review and a ...](https://ieeexplore.ieee.org/document/8354043/)

<https://ieeexplore.ieee.org/document/8354043/>

by K Sellamy - 2018 - [Related articles](#)

Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco.

[PDF] [Semantic Web Requirements through Web Mining Techniques - arXiv](https://arxiv.org/pdf/1208.0690)

<https://arxiv.org/pdf/1208.0690> ▼

by H Hassanzadeh - 2012 - [Cited by 9](#) - [Related articles](#)

Jun 2, 2012 - **Web mining techniques and applications.** There are three factors affecting the way a user perceives and values a site: content, Web page ...

(PDF) [Review on Web Content Mining Techniques - ResearchGate](https://www.researchgate.net/.../276928728)

<https://www.researchgate.net/.../276928728> [Review on Web Content Mining Tech...](#) ▼

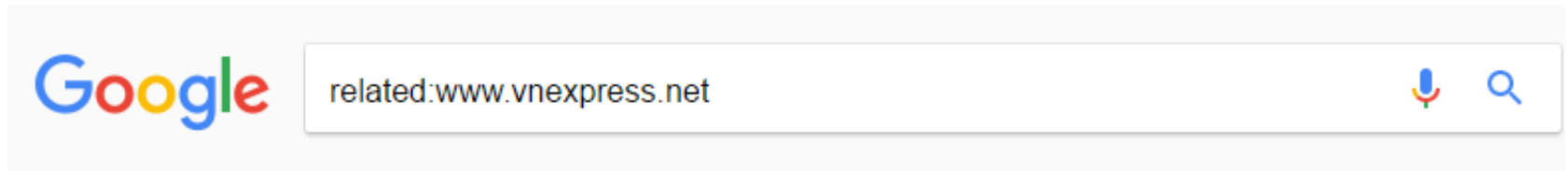
May 21, 2015 - **Web mining techniques and applications:** Literature review and a proposal approach to improve perform... April 2018. View the large amount of ...

User query: Truy vấn lân cận

- Phiên bản yếu (relaxed) của truy vấn cụm
- Tổ hợp của từ và cụm từ
- **Tính gần** (closeness) giữa các từ truy vấn được xem là hệ số để sắp xếp tài liệu trả về.
 - Ví dụ, tài liệu chứa tất cả các từ khóa nằm gần với nhau được xem là liên quan hơn tài liệu mà trong đó các từ khóa nằm xa nhau
- Hầu hết cỗ máy tìm kiếm xét cả *tính gần của từ* và *thứ tự của từ* khi truy vấn.

User query: Truy vấn toàn tài liệu

- Truy vấn là một tài liệu đầy đủ
- Tìm các tài liệu tương tự với tài liệu mẫu được đưa ra



VietNamNet: Tin tức, Đọc báo Online, Tin tức trong ngày 24h

vietnamnet.vn/ ▾ [Translate this page](#)

Tin tức trong ngày 24h, VietNamNet cập nhật nhanh nhất & mới nhất. Đọc báo tin tức online về thời sự, pháp luật, giải trí, ... hay nhất 24h qua.

Cached

Similar

Báo Dân trí | Tin tức Việt Nam và quốc tế nóng, nhanh, cập nhật 24h

dantri.com.vn/ ▾ [Translate this page](#)

Tin tức, sự kiện Việt Nam và quốc tế nhanh, chính xác và cập nhật mới online 24h /7: Xã hội, Pháp luật, Kinh doanh, Giải trí, Thể thao, Đời sống, Sức khỏe, Giáo dục ...

Tin tức, tin nóng, đọc báo điện tử - Tuổi Trẻ Online

tuoitre.vn/ ▾ [Translate this page](#)

Tin tức nhanh - mới - nóng nhất đang diễn ra về: kinh tế, chính trị, xã hội, thể giới, giáo dục, thể thao, văn hóa, giải trí, công nghệ.

User query: Câu hỏi ngôn ngữ tự nhiên

- Tình huống phức tạp nhất, mang nhiều tính lý tưởng
- Người dùng mô tả nhu cầu thông tin → hệ thống tìm đáp án
 - Lĩnh vực nghiên cứu hỏi đáp (question answering).
 - Khó giải quyết vì ngôn ngữ tự nhiên không dễ hiểu cho máy.

STAP
Natura
what is web mini

Google Assistant

amazon alexa

ion Answering System

States is Barack Hussein Obama.

ing and Web

patterns from data captures the

what is w

Web mining

Web mining - is Web mining can structure mining

Web Usage Mining

Web data in order to identify or origin of new users along with their browsing behavior in a web site.

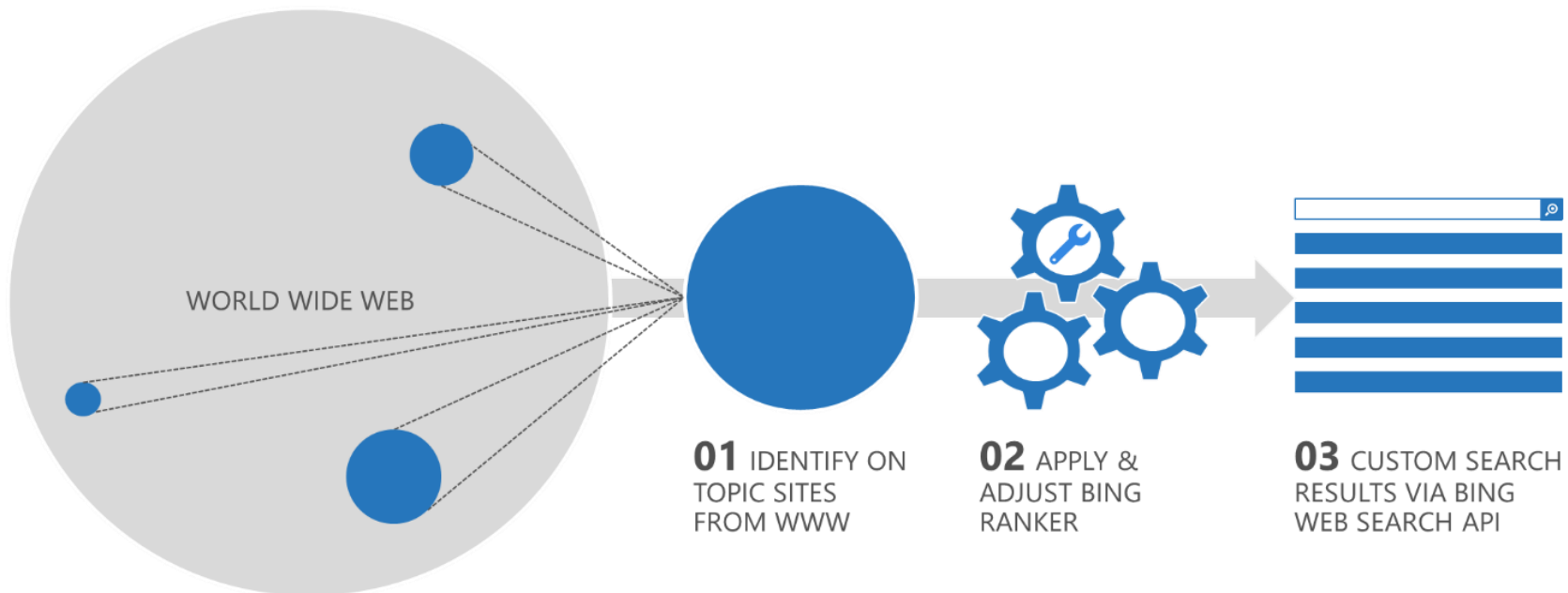
Web usage mining itself can be classified further depending on the kind of usage data considered:

Hi, I'm Cortana.

Hey Siri

Tìm kiếm Web

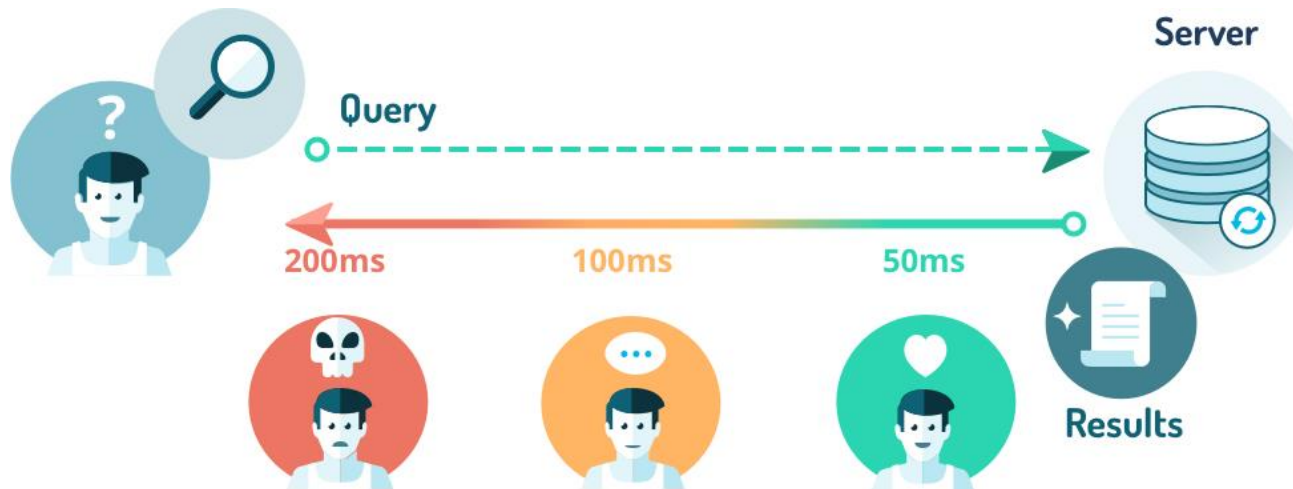
- **Tìm kiếm Web** (Web search) là một trong những ứng dụng IR quan trọng nhất, trong đó tài liệu là những trang Web.
- Sử dụng một số thành tựu IR, đồng thời có kỹ thuật riêng
- Đề ra nhiều bài toán mới thú vị cho nghiên cứu IR



Đặc điểm của Tìm kiếm Web

1

- Hiệu suất là vấn đề tối cao của tìm kiếm Web.
 - Số lượng trang Web khổng lồ, ví dụ Google vào năm 2017 đã đánh chỉ mục hơn 130×10^{12} trang.
 - Người dùng Web đòi hỏi hồi đáp cực nhanh.



- Tuy nhiên, vấn đề này chỉ là thứ yếu trong hệ thống truy vấn thông tin vì tập hợp tài liệu không quá lớn.

Đặc điểm của Tìm kiếm Web

2

- Trang Web hoàn toàn khác với tài liệu văn bản truyền thống trong những hệ thống IR cơ bản.
 - **Hyperlinks** là nhân tố cốt yếu cho các giải thuật xếp hạng tìm kiếm.
 - **Anchor text** kết hợp với một hyperlink thường là mô tả chính xác về trang mà hyperlink trở đến.
- Trang Web có tính bán cấu trúc.
 - Nhiều trường khác nhau đa dạng (ví dụ, title, metadata, body, v.v.)
 - Thông tin chứa trong một số trường (ví dụ title) quan trọng hơn trong những trường khác (ví dụ., quảng cáo, privacy policy, copyright notices, v.v.).

3

- Spamming là vấn đề nghiêm trọng trên Web, trong khi đó IR truyền thống ít quan tâm đến vấn đề này.
 - Trang có chất lượng thấp (kể cả không liên quan) lọt vào những hạng đầu → ảnh hưởng kết quả tìm kiếm và trải nghiệm người dùng.

Index of /movies/avi4/08

| Name |
|---|
| Parent Directory |
| Remembrance of Love.avi |
| How the Grinch Stole Christmas.avi |
| Breathless.avi |
| ALAIN DELON - GRAND JOULET EN LA COTE D'AZUR (DVD) REGION 2 PAL Region 2 Import.avi |
| Love Song Remains.avi |
| Legend of the Blue Wolves.avi |
| Sgt. Pepper's Lonely Hearts Club Band.avi |
| Reis Segunda.avi |
| 17th Delta Force La Sierrita.avi |
| Yachting: Legends.avi |
| Star 1 Movie.avi |
| Tragedy of The Third Eye.avi |
| Escoria Mexicana.avi |
| Circle Remains Who Can Recall His Past Lives.avi |
| Royal Wedding.avi |
| RAH!.avi |
| Blue Murder At St. Trinian's.avi |
| The Cuban Connection.avi |
| Sailed of a Soldier (The Criterion Collection).avi |
| Swing Vote.avi |
| Happy Ghost, The (DVD).avi |
| How to Succeed in Business Without Really Trying (DVD).avi |
| Yachting - Fastlane Against Fastlane 1 & 2 Policies Apache 1 & 2 REGION 2 PAL Region 2 Import |
| Main Menu (2) DVD Region 2 PAL |
| Euphon Complex.avi |
| Mississippi Marauder.avi |
| 2 St. Blues.avi |
| American Eagle.avi |
| Rabbit Hole.avi |
| Supergirl.avi |
| The Creation Adventure Team: A Jurassic Ark Mystery.avi |
| TEAM - GAT by David Lowery (DVD) REGION 2 PAL Region 2 Import |
| Wild Willie Weekend.avi |
| Victorian Romance.avi |
| The Texas Chainsaw Massacre.avi |
| How the Grinch Stole Christmas (DVD) Region 2 PAL |
| Football.avi |
| Shadow Box on Sacred Ground.avi |
| Travelling the Undiscoverable.avi |
| How the Grinch Stole Christmas (DVD).avi |
| Heart of Dragon (DVD) First Season (Region 2 / DVD) (DVD).avi |
| A Night to Remember.avi |
| The Offenders (Dante Desiderio).avi |
| Red A Rock.avi |
| Love & Marriage: Nothing to Declare (Region 2).avi |
| The Baby Maker.avi |
| Fun Grassy Fun.avi |

These pages are examples of 'pure spam.' They appear to use aggressive spam techniques such as automatically generated gibberish, cloaking and scraping content from other websites.

Screenshot of the removed page

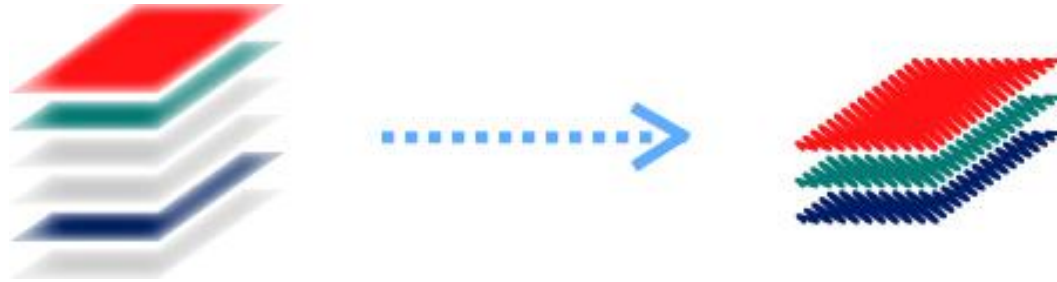
Snippet of the removed page

Remove from search results an hour ago

Index of /movies/avi4/08

http://topicalarticles.info/movies/avi4/08/

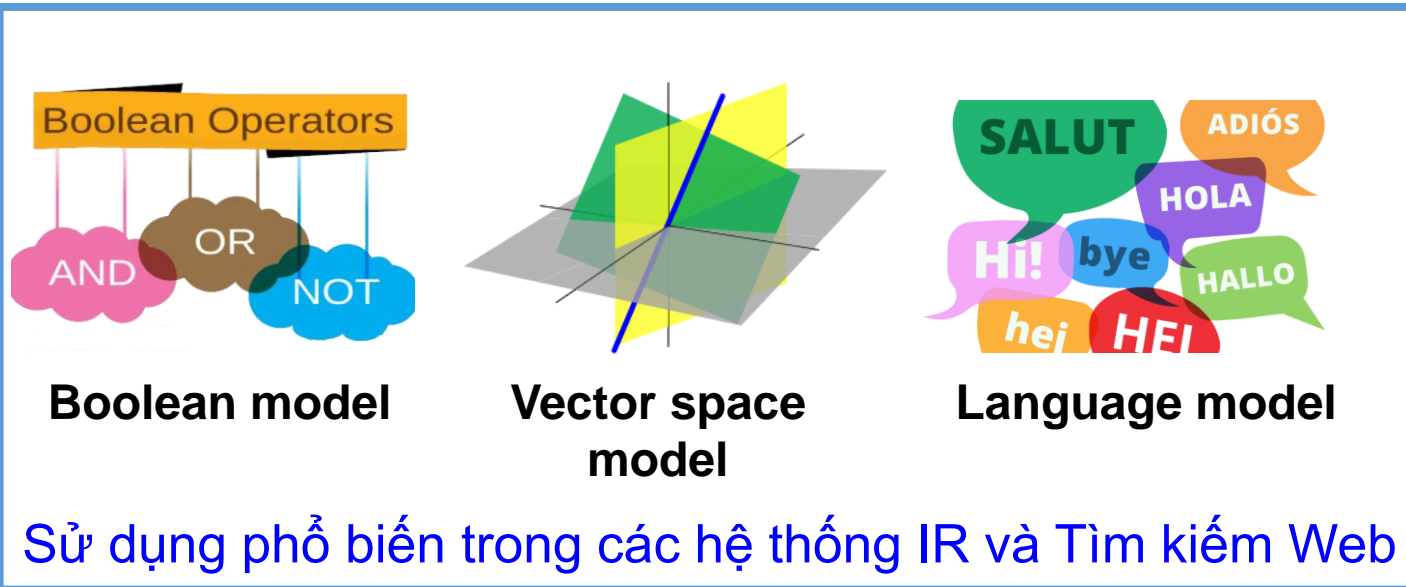
Parent Directory Remembrance of Love.avi 15-May-2011 2760M How the Grinch Stole Christmas.avi 29-Nov-2011 1626M Breathless.avi 02-Mar-2012 3470M ...



Mô hình truy vấn thông tin

Mô hình truy vấn thông tin

- Chi phối cách thức thể hiện tài liệu và câu truy vấn, định nghĩa sự liên quan của một tài liệu với câu truy vấn.
- Các mô hình IR chính



Mô hình truy vấn thông tin

- Các mô hình IR khác nhau biểu diễn tài liệu và câu truy vấn theo cách thức khác nhau nhưng chia sẻ cùng framework.
- Mỗi tài liệu (hay truy vấn) được biểu diễn bằng một tập hợp thuật ngữ, gọi là “**bag of words**”.
 - Thứ tự và vị trí của các thuật ngữ trong tài liệu bị bỏ qua.
- Thuật ngữ hay từ (**term**) được sử dụng phải có ngữ nghĩa gợi nhớ chủ đề chính của tài liệu.
 - Ví dụ, sport (match, goal, coach, v.v.), education (class, lecture, v.v.)

Biểu diễn tài liệu trong IR

- Cho tập hợp tài liệu D .
- Gọi $V = \{t_1, t_2, \dots, t_{|V|}\}$ là tập hợp gồm các từ t_i phân biệt có trong D .
 - V được gọi là **tập ngữ vựng của D** , có kích thước $|V|$.
- Độ quan trọng của $t_i \in V$ trong tài liệu $d_j \in D$ được lượng hóa bằng trọng số $w_{ij} > 0$.
 - $w_{ij} = 0$ khi từ t_i không xuất hiện trong tài liệu d_j .
 - w_{ij} được tính toán theo nhiều cách khác nhau, tùy mô hình IR.
- Mỗi tài liệu d_j được biểu diễn bằng vector từ

$$d_j = (w_{1j}, w_{2j}, \dots, w_{|V|j})$$

Biểu diễn tài liệu trong IR

- Tập hợp tài liệu được thể hiện như một ma trận.
 - Mỗi từ là một thuộc tính, và mỗi trọng số là một giá trị thuộc tính.
- Ví dụ

| | intelligent | applications | creates | business | processes | bots | are | i | do | intelligence |
|-------|-------------|--------------|---------|----------|-----------|------|-----|---|----|--------------|
| Doc 1 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Doc 2 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| Doc 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |

Doc 1 = Intelligent applications creates intelligent business processes

Doc 2 = Bots are intelligent applications

Doc 3 = I do business intelligence

Mô hình luận lý Boolean

- Một trong những mô hình IR đơn giản nhất và sớm nhất
- **Biểu diễn tài liệu:** Mỗi từ chỉ được xét là xuất hiện/không xuất hiện trong một tài liệu.

$$w_{ij} = \begin{cases} 1 & \text{nếu } t_i \text{ có trong } \mathbf{d}_j \\ 0 & \text{ngược lại} \end{cases}$$

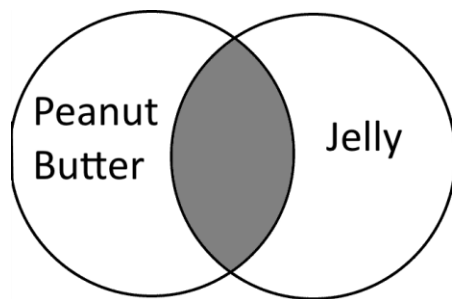
the dog is on the table

| | | | | | | | |
|-----|-----|-----|----|-----|----|-------|-----|
| 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| are | cat | dog | is | now | on | table | the |

- **Truy vấn tài liệu:** Hệ thống lấy mọi tài liệu làm cho câu truy vấn Boolean đúng về mặt luận lý.
 - So khớp chính xác, liên quan/không liên quan
 - Không so khớp một phần hay xếp hạng tài liệu tìm thấy

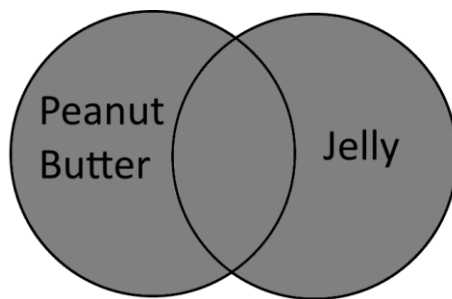
Mô hình luận lý Boolean

- **Câu truy vấn Boolean:** Các từ khóa được kết hợp luận lý bằng toán tử Boolean (**AND**, **OR** và **NOT**).
 - Ví dụ, câu truy vấn $((x \text{ AND } y) \text{ AND } (\text{NOT } z))$ chỉ định tìm tài liệu chứa cả hai từ x và y nhưng không chứa z .



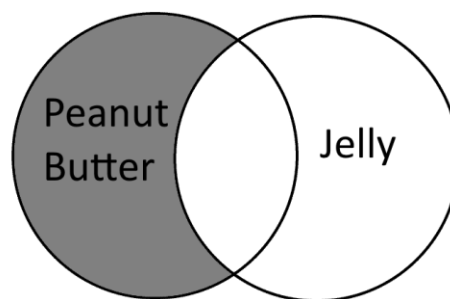
AND

Using AND, this search would only retrieve results with Peanut Butter and Jelly.



OR

Using OR, this search would retrieve results with peanut butter, with jelly, and with both.



NOT

Using NOT, this search would retrieve results with peanut butter, and exclude those with jelly or PB with jelly.

Mô hình luận lý Boolean: Ví dụ

- Cho tập tài liệu $D = \{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$, trong đó nội dung của mỗi tài liệu là
 - $\mathbf{d}_1 = \{\text{Bayes' Principle, probability}\}$
 - $\mathbf{d}_2 = \{\text{probability, decision-making}\}$
 - $\mathbf{d}_3 = \{\text{probability, Bayesian Epistemology}\}$
- Câu truy vấn $\mathbf{q} = \text{probability AND decision-making}$

| | Bayes' Principle | probability | decision- making | Bayesian Epistemology |
|----------------|------------------|-------------|------------------|-----------------------|
| \mathbf{d}_1 | 1 | 1 | 0 | 0 |
| \mathbf{d}_2 | 0 | 1 | 1 | 0 |
| \mathbf{d}_3 | 1 | 1 | 0 | 1 |
| \mathbf{q} | 0 | 1 | 1 | 0 |

- Tài liệu chứa từ *probability*: $S_1 = \{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$
- Tài liệu chứa từ *decision-making*: $S_2 = \{\mathbf{d}_2\}$
- Các tài liệu được trả về: $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\} \cap \{\mathbf{d}_2\} = \{\mathbf{d}_2\}$.

Mô hình không gian vector

- Mô hình IR nổi tiếng nhất và được sử dụng rộng rãi nhất
- **Biểu diễn tài liệu:** Tài liệu được biểu thành vector trọng số, mỗi thành phần được tính toán theo một biến thể nào đó của **lược đồ TF** hoặc **TF-IDF**.

$$VSM = \begin{bmatrix} w(1,1) & \cdots & w(1,d_j) & \cdots & w(1,N) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w(q_i, 1) & \cdots & w(q_i, d_j) & \cdots & w(q_i, N) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w(M, 1) & \cdots & w(M, d_j) & \cdots & w(M, N) \end{bmatrix}$$

Lược đồ TF

- Term Frequency (TF) scheme
- Trọng số của từ t_i trong tài liệu d_j là số lần t_i xuất hiện trong tài liệu d_j , được kí hiệu là f_{ij} .
- TF có thể được chuẩn hóa bằng công thức sau

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

- Không xét trường hợp từ xuất hiện ở nhiều tài liệu

Lược đồ TF – IDF

- Term Frequency – Inverse Document Frequency (TF-IDF)
- Lược đồ trọng số phổ biến nhất, có nhiều biến thể
- **TF** được định nghĩa trong slide trước
- Gọi N là số tài liệu trong cơ sở dữ liệu và df_i là số tài liệu mà trong đó từ t_i xuất hiện ít nhất một lần.
- IDF của từ t_i là

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

- Như vậy, TF – IDF là của từ t_i đối với tài liệu d_j là

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

Lược đồ TF – IDF: Ví dụ

- Cho tập hợp gồm $N = 10,000,000$ tài liệu.
- Giả sử rằng từ “iPad” xuất hiện 3 lần trong tài liệu document \mathbf{d}_1 , và số lần xuất hiện tối đa của một từ nào đó trong \mathbf{d}_1 là 100. Như vậy,

$$tf_{iPad} = \frac{3}{100} = 0.03$$

- Từ “iPad” xuất hiện trong $1,000 / N$ tài liệu. Do đó,

$$idf_{iPad} = \log\left(10,000,000 / 1,000\right) = 4$$

- Cuối cùng, trọng số của từ “iPad” trong tài liệu \mathbf{d}_1 là

$$w_{iPad1} = 0.03 \times 4 = 0.12$$

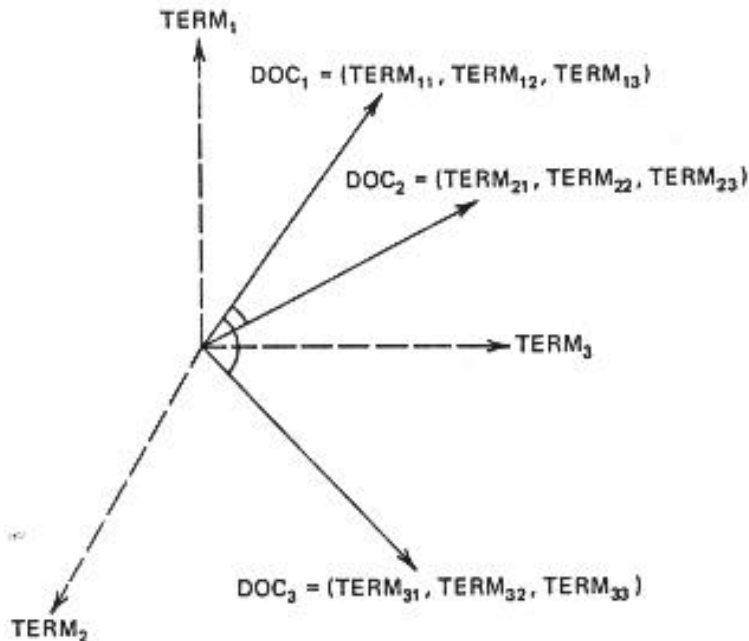
Mô hình không gian vector

- **Truy vấn:** q được biểu diễn theo đúng cách mà các tài liệu được biểu diễn.
- Trọng số w_{iq} của mỗi từ t_i trong q cũng được tính tương tự như trong tài liệu, có thể hơi khác tùy biến thể.
 - Ví dụ, công thức Salton and 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

- **Truy vấn tài liệu và Sắp xếp theo theo độ liên quan:** Các tài liệu tìm thấy được xếp hạng theo mức độ liên quan của chúng với truy vấn.
 - Tính toán độ tương tự (ví dụ, Cosine similarity) của truy vấn q với mỗi tài liệu d_j trong tập hợp tài liệu D .



$$\begin{aligned} \text{cosine}(d_j, q) &= \frac{\langle d_j \cdot q \rangle}{\|d_j\| \times \|q\|} \\ &= \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}} \end{aligned}$$

Cosine similarity: Ví dụ

- Cho tài liệu d_1 và truy vấn q

d_1 : *Julie loves me more than Linda loves me*

q : *Jane likes me more than Julie loves me*

- Đếm số lần xuất hiện của mỗi từ trong d_1 và q

| | me | Julie | likes | loves | Jane | Linda | than | more |
|-------|----|-------|-------|-------|------|-------|------|------|
| d_1 | 2 | 1 | 0 | 2 | 0 | 1 | 1 | 1 |
| q | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

- Giả sử rằng mỗi thành phần trong vector trọng số là số lần xuất hiện của từ tương ứng. Như vậy,

$$d_1 = [2, 1, 0, 2, 0, 1, 1, 1]$$

$$q = [2, 1, 1, 1, 1, 0, 1, 1]$$

- Cuối cùng, $\text{cosine}(d_j, q) = 0.822$.

Phương pháp Okapi

- Phương pháp Okapi và các biến thể thường hiệu quả hơn Cosine similarity đối với truy vấn ngắn.

$$okapi(d_j, q) = \sum_{t_i \in q, d_j} \ln \frac{N - df_i + 0.5}{df_i + 0.5} \times \frac{(k_1 + 1)f_{ij}}{k_1 \left(1 - b + b \frac{dl_j}{avdl} \right) + f_{ij}} \times \frac{(k_2 + 1)f_{iq}}{k_2 + f_{iq}}$$

- Các tham số: $k_1 = 1.0 - 2.0$, b thường là 0.75, và $k_2 = 1 - 1000$.
- dl_j là độ dài tài liệu (theo byte) của d_j
- $avdl$ là độ dài tài liệu trung bình của toàn tập dữ liệu

Pivoted normalization weighting

- Phương pháp pivoted normalization weighting được tính là

$$pnw(d_j, q) = \sum_{t_i \in q, d_j} \frac{1 + \ln(1 + \ln(f_{ij}))}{(1 - s) + s \frac{dl_j}{avdl}} f_{iq} \times \ln \frac{N + 1}{df_i}$$

- Trong đó s là tham số thường được thiết lập bằng 0.2.

Bài tập 1: Mô hình không gian vector

- Cho tập hợp các tài liệu \mathbf{d}_i và truy vấn \mathbf{q} như sau
 - \mathbf{d}_1 : new york times
 - \mathbf{d}_2 : new york post
 - \mathbf{d}_3 : los angeles times
 - \mathbf{q} : new new times
- Xếp hạng các tài liệu theo giá trị **Cosine similarity** của mỗi tài liệu đối với \mathbf{q} trong **mô hình không gian vector**.

Mô hình ngôn ngữ thống kê

- Dựa trên xác suất của lý thuyết thống kê.
- **Ý tưởng cơ bản:** Ước lượng một mô hình ngôn ngữ cho mỗi tài liệu rồi xếp hạng tài liệu theo khả năng (likelihood) truy vấn được cho bởi mô hình đó.
 - Những ý tưởng tương tự đã được sử dụng trước đó trong xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói.
- Truy vấn thông tin sử dụng mô hình ngôn ngữ được đề xuất lần đầu tiên bởi Ponte và Croft.

Mô hình ngôn ngữ thống kê

- Gọi truy vấn $\mathbf{q} = q_1 q_2 \dots q_m$ là chuỗi từ, và tập hợp tài liệu $D = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N\}$.
- $\Pr(\mathbf{q} | d_j)$ là xác suất để truy vấn \mathbf{q} được phát sinh bởi mô hình ngôn ngữ của tài liệu \mathbf{d}_j .
- Để xếp hạng tài liệu, **luật Bayes** được dùng để ước lượng

$$\Pr(\mathbf{d}_j | \mathbf{q}) = \frac{\Pr(\mathbf{q} | \mathbf{d}_j) \Pr(\mathbf{d}_j)}{\Pr(\mathbf{q})}$$

- $\Pr(\mathbf{d}_j)$ thường được xem là đồng nhất và do đó không ảnh hưởng việc xếp hạng.

Mô hình ngôn ngữ thống kê

- Mỗi từ được giả sử phát sinh độc lập (**unigram**), về bản chất là một phân phối đa thức trên các từ.

$$\Pr(\mathbf{q} = \mathbf{q}_1 \mathbf{q}_2 \dots \mathbf{q}_m | \mathbf{d}_j) = \prod_{i=1}^m \Pr(\mathbf{q}_i | \mathbf{d}_j) = \prod_{i=1}^{|\mathbf{V}|} \Pr(\mathbf{t}_i | \mathbf{d}_j)^{f_{i\mathbf{q}}}$$

- $f_{i\mathbf{q}}$ là số lần xuất hiện của từ \mathbf{t}_i trong \mathbf{q} và $\sum_{i=1}^{|\mathbf{V}|} \Pr(\mathbf{t}_i | \mathbf{d}_j) = 1$.
- Bài toán truy vấn được giảm nhẹ thành bài toán ước lượng tần số tương đối $\Pr(\mathbf{t}_i | \mathbf{d}_j) = \frac{f_{i\mathbf{d}_j}}{|\mathbf{d}_j|}$
 - Trong đó $f_{i\mathbf{d}_j}$ là số lần xuất hiện của từ \mathbf{t}_i trong \mathbf{d}_j và $|\mathbf{d}_j|$ chỉ tổng số từ trong \mathbf{d}_j .

Phân phối đa thức: Ví dụ

- Cho bảng xác suất ước lượng $\Pr(t_i|\mathbf{d}_1)$ và $\Pr(t_i|\mathbf{d}_2)$

| t_i | $\Pr(t_i \mathbf{d}_1)$ | $\Pr(t_i \mathbf{d}_2)$ |
|-------|-------------------------|-------------------------|
| 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 |
| ... | ... | ... |

- Truy vấn \mathbf{q} = world share
- $\Pr(\mathbf{q}|\mathbf{d}_1) = \Pr(world|\mathbf{d}_1) \times \Pr(share|\mathbf{d}_1) = 0.2 \times 0.3 = 0.06$
- $\Pr(\mathbf{q}|\mathbf{d}_2) = \Pr(world|\mathbf{d}_2) \times \Pr(share|\mathbf{d}_2) = 0.1 \times 0.2 = 0.02$

Mô hình ngôn ngữ thống kê

- Từ không xuất hiện trong d_j có xác suất 0 \rightarrow đánh giá thấp xác suất của từ không thấy trong tài liệu.
- **Smoothing:** Một giá trị xác suất khác 0 được gán cho mỗi từ không xuất hiện trong tài liệu.
- Việc smoothing thường được triển khai theo công thức sau

$$Pr_{add}(t_i | \mathbf{d}_j) = \frac{\lambda + f_{ij}}{\lambda |V| + |\mathbf{d}_j|}$$

- $\lambda = 1$: Laplace smoothing, $0 < \lambda < 1$: Lidstone smoothing.

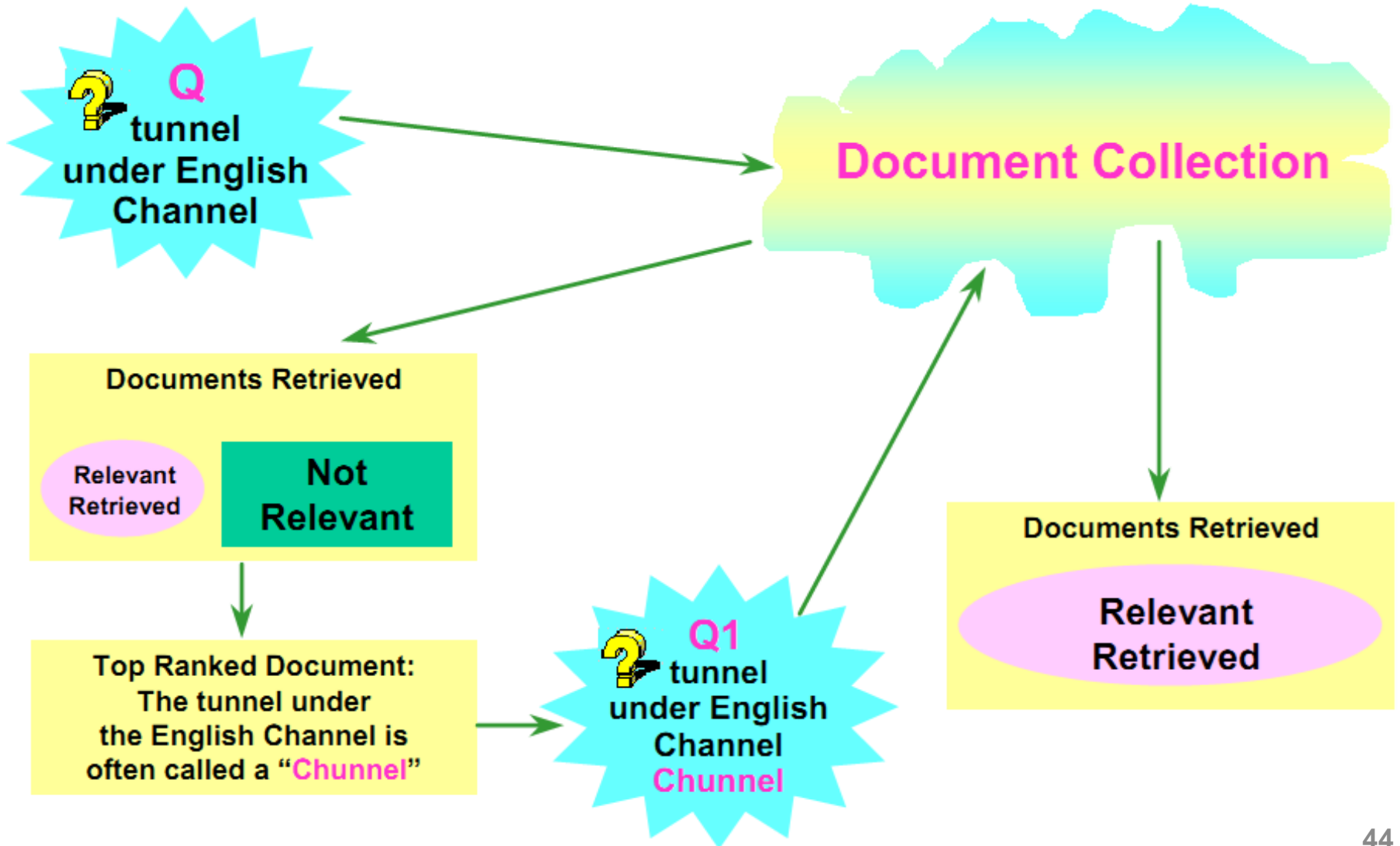


Phản hồi liên quan

Phản hồi liên quan

- Relevance feedback (RF)
- Người dùng xác định tài liệu liên quan/không liên quan trong danh sách tài liệu trả về ban đầu.
- Một số từ trong tài liệu trả về được trích ra để tạo truy vấn mở rộng cho vòng lặp kế tiếp.
- Tiến trình được lặp lại đến khi người dùng thỏa mãn với kết quả nhận được.

Phản hồi liên quan: Ví dụ



Phương pháp Rocchio

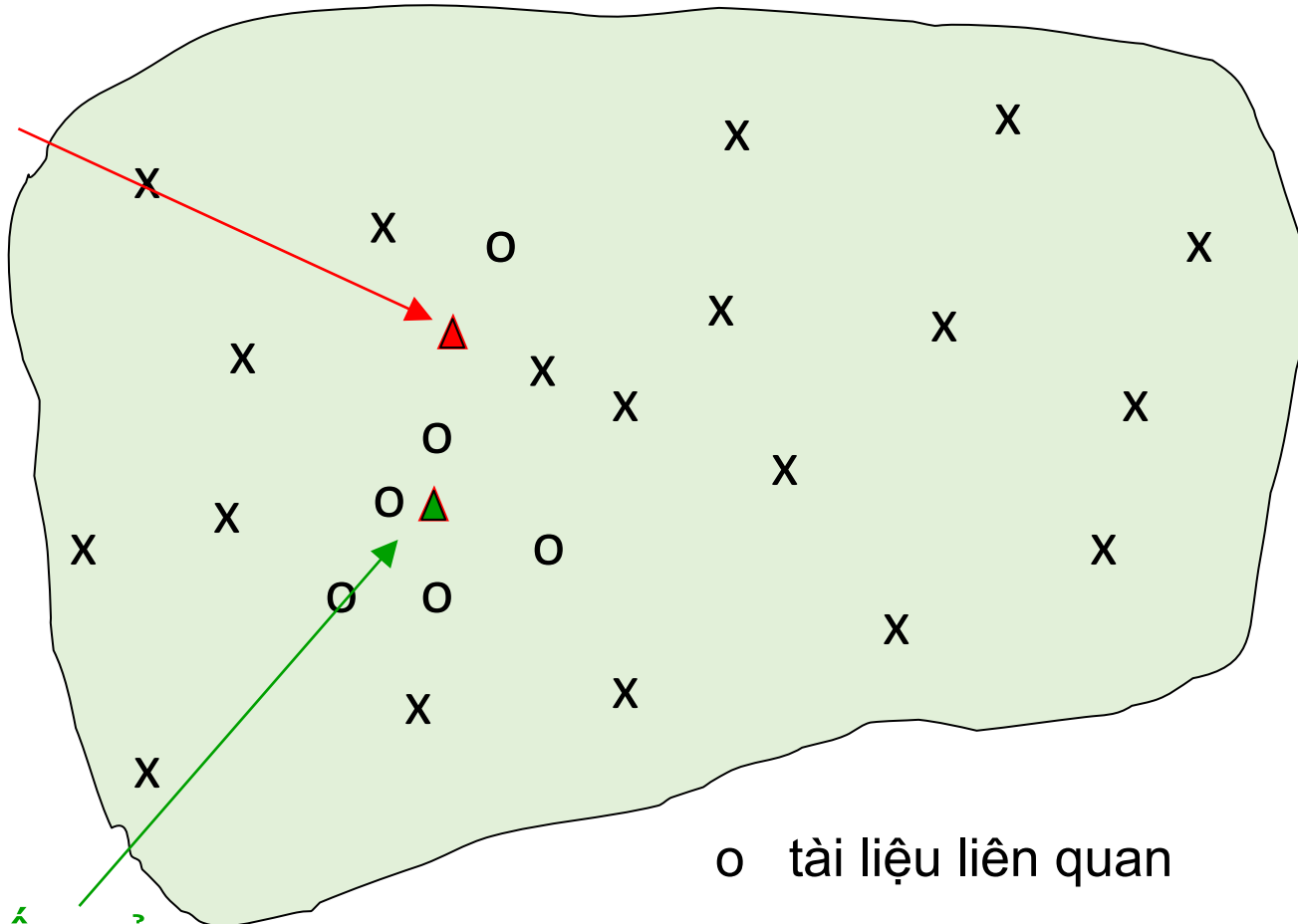
- Một trong những giải thuật phản hồi liên quan sớm nhất và có hiệu quả
- Gọi \mathbf{q} là vector truy vấn gốc, D_r là tập hợp tài liệu liên quan do người dùng chọn, và D_{ir} là tập tài liệu không liên quan.
- Truy vấn mở rộng \mathbf{q}_e được tính như sau

$$\mathbf{q}_e = \alpha \mathbf{q} + \frac{\beta}{|D_r|} \sum_{\mathbf{d}_r \in D_r} \mathbf{d}_r - \frac{\gamma}{|D_{ir}|} \sum_{\mathbf{d}_{ir} \in D_{ir}} \mathbf{d}_{ir}$$

- Trong đó α, β và γ là các tham số được thiết lập từ thực nghiệm.

Phương pháp Rocchio: Ví dụ

Truy vấn ban đầu



Truy vấn mở rộng

Mô hình học máy

- Xây dựng mô hình phân lớp từ tập tài liệu liên quan và không liên quan → bài toán học máy
- Có thể sử dụng bất kì phương pháp học có giám sát nào, ví dụ Naïve Bayesian, SVM, ...
 - Không cần so sánh độ tương tự với truy vấn gốc.



Mô hình học máy

- Học từ mẫu gán nhãn và chưa gán nhãn
 - Tập tài liệu liên quan và tài liệu không liên quan do người dùng chọn, tạo thành tập huấn luyện có gán nhãn với quy mô nhỏ.
 - Những tài liệu không được người dùng chọn cũng được tận dụng để tăng chất lượng học và do đó tạo ra bộ phân lớp chính xác hơn.
- Học từ mẫu dương và mẫu chưa gán nhãn
 - Người dùng chỉ chọn những tài liệu liên quan theo tiêu đề hoặc nội dung tóm tắt (ví dụ, snippets trong tìm kiếm Web) → mẫu dương, nhưng không chỉ định tài liệu không liên quan → mẫu chưa gán nhãn.
- Sử dụng Ranking SVM và các mô hình ngôn ngữ

Phương pháp phân lớp Rocchio

- Xây dựng một prototype vector \mathbf{c}_i cho mỗi lớp i

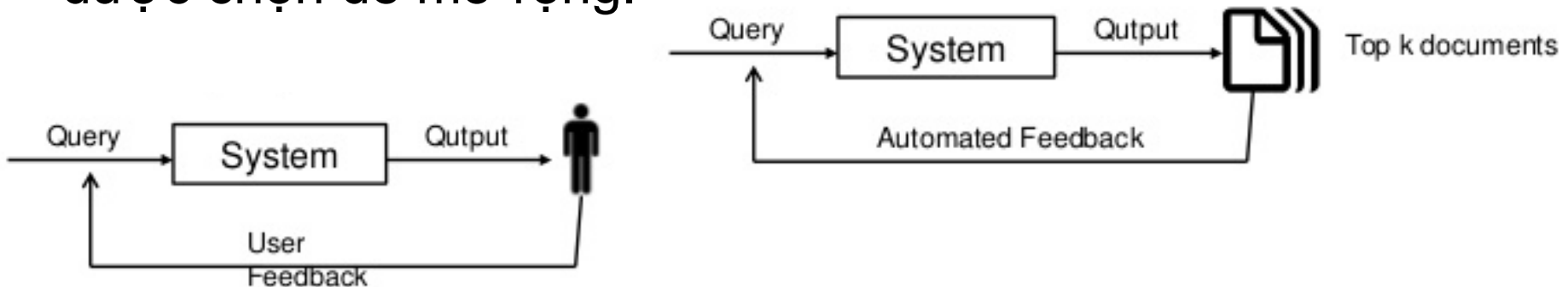
$$\mathbf{c}_i = \frac{\alpha}{|D_i|} \sum_{\mathbf{d} \in D_i} \frac{\mathbf{d}}{\|\mathbf{d}\|} - \frac{\beta}{|D - D_i|} \sum_{\mathbf{d} \in D - D_i} \frac{\mathbf{d}}{\|\mathbf{d}\|}$$

- Trong đó D_i là tập tài liệu thuộc phân lớp i , α và β là tham số ($\alpha = 16$ và $\beta = 4$ theo lược đồ TF-IDF).
- Thành phần âm của \mathbf{c}_i thường được thiết lập về 0

```
1  for each class  $i$  do
2      construct its prototype vector  $\mathbf{c}_i$  using Equation (15)
3  endfor
4  for each test document  $\mathbf{d}_t$  do
5      the class of  $\mathbf{d}_t$  is  $\arg \max_i \text{cosine}(\mathbf{d}_t, \mathbf{c}_i)$ 
6  endfor
```

Phản hồi liên quan giả

- Pseudo-Relevance feedback
- Một số từ (thường là từ phổ biến) được trích từ những tài liệu xếp hạng cao nhất và thêm vào truy vấn gốc để tạo thành truy vấn mới cho lượt tìm kiếm tiếp theo.
 - Tài liệu xếp hạng cao thường được cho là liên quan.
- Người dùng không can thiệp vào tiến trình.
- Hiệu quả phụ thuộc chủ yếu vào chất lượng của các từ được chọn để mở rộng.





Độ đo đánh giá

Xếp hạng tài liệu

- Các hệ thống IR và tìm kiếm Web cung cấp cho người dùng danh sách tài liệu kết quả có xếp hạng, thay vì chỉ phân loại liên quan.
- Gọi D là tập hợp gồm N tài liệu và q là truy vấn.
- Giải thuật truy vấn tính điểm số liên quan cho mọi tài liệu trong D để tạo thành danh sách xếp hạng tài liệu R_q .

$$R_q: \langle \mathbf{d}_1^q, \mathbf{d}_2^q, \dots, \mathbf{d}_N^q \rangle$$

- trong đó d_1^q và d_N^q lần lượt là những tài liệu trong D có liên quan nhất và ít liên quan nhất đến q .

Độ chính xác – Độ phủ

- Gọi $D_q \subseteq D$ là tập tài liệu trong D thật sự liên quan đến q .
- Gọi s_i là số tài liệu được cho là liên quan từ d_1^q đến d_i^q trong R_q ($s_i \leq |D_q|$, $|D_q|$ là kích thước của D_q).
- **Độ phủ** (recall) ở hạng thứ i (hay tài liệu d_i^q) là tỉ lệ tài liệu thật sự liên quan từ d_1^q đến d_i^q trong R_q , so với $|D_q|$.

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

- **Độ chính xác** (precision) ở hạng thứ i (hay tài liệu d_i^q) là tỉ lệ tài liệu thật sự liên quan từ d_1^q đến d_i^q trong R_q , so với i .

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

Độ chính xác – Độ phủ: Ví dụ

- Tập D có 20 tài liệu.
- Truy vấn 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% |

Độ chính xác trung bình

- Ta cần có một độ chính xác đơn để dễ dàng so sánh các giải thuật truy vấn khác nhau trên một truy vấn q .
- **Độ chính xác trung bình** (average precision) dựa trên độ chính xác ở mỗi tài liệu liên quan trong danh sách xếp hạng.

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

- Ví dụ, từ bảng tính precision và recall ở slide trước, ta có

$$p_{avg} = \frac{100\% + 100\% + 100\% + 80\% + 71\% + 67\% + 70\% + 62\%}{8} = 81\%$$

Đường cong độ chính xác – độ phủ

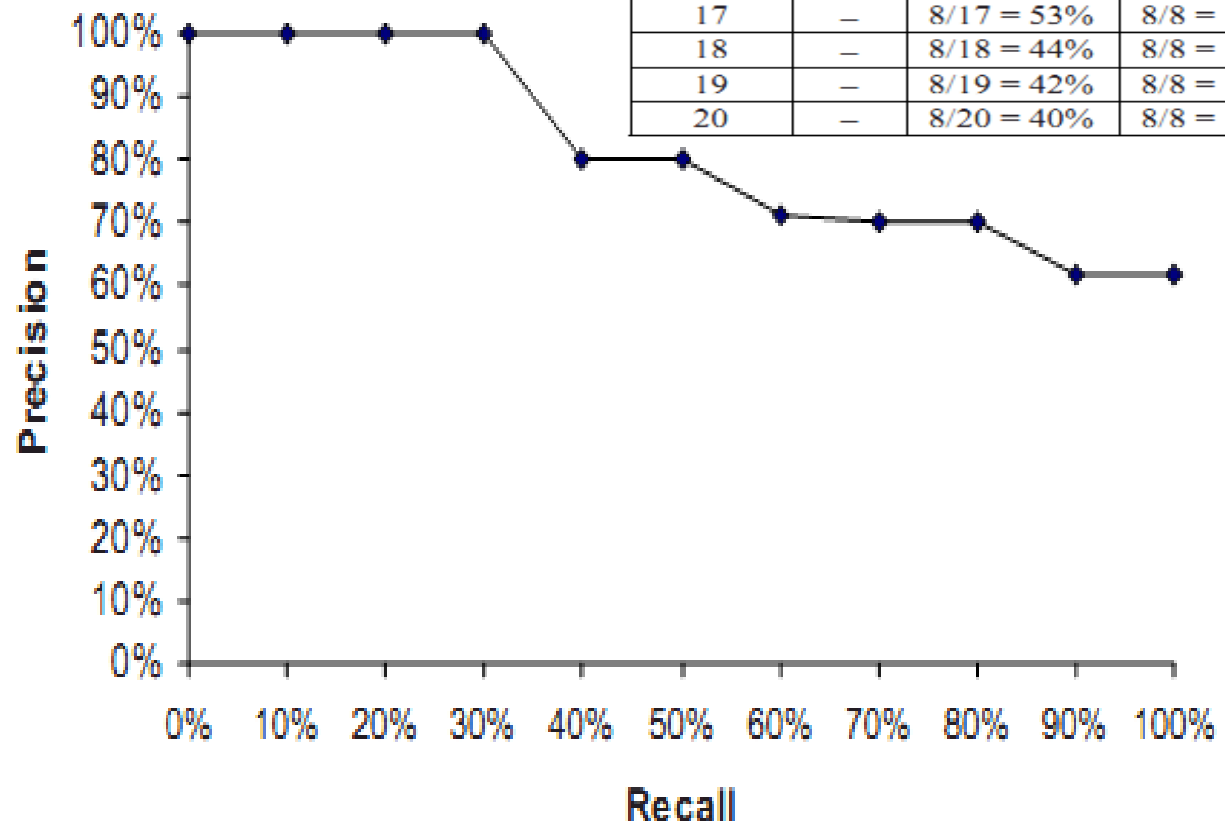
- **Precision–recall curve**

- Thể hiện giá trị độ phủ ở trục x và giá trị độ chính xác tương ứng ở trục y .
 - Trục x được chia thành 11 mức phủ chuẩn: 0%, 10%, 20%, ..., 100%.
- Gọi r_i là mức phủ, $i \in \{0, 1, 2, \dots, 10\}$.
- Độ chính xác tại mức phủ r_i được nội suy như sau

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

Đường cong độ chính xác – độ phủ

| 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% |



| 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% |

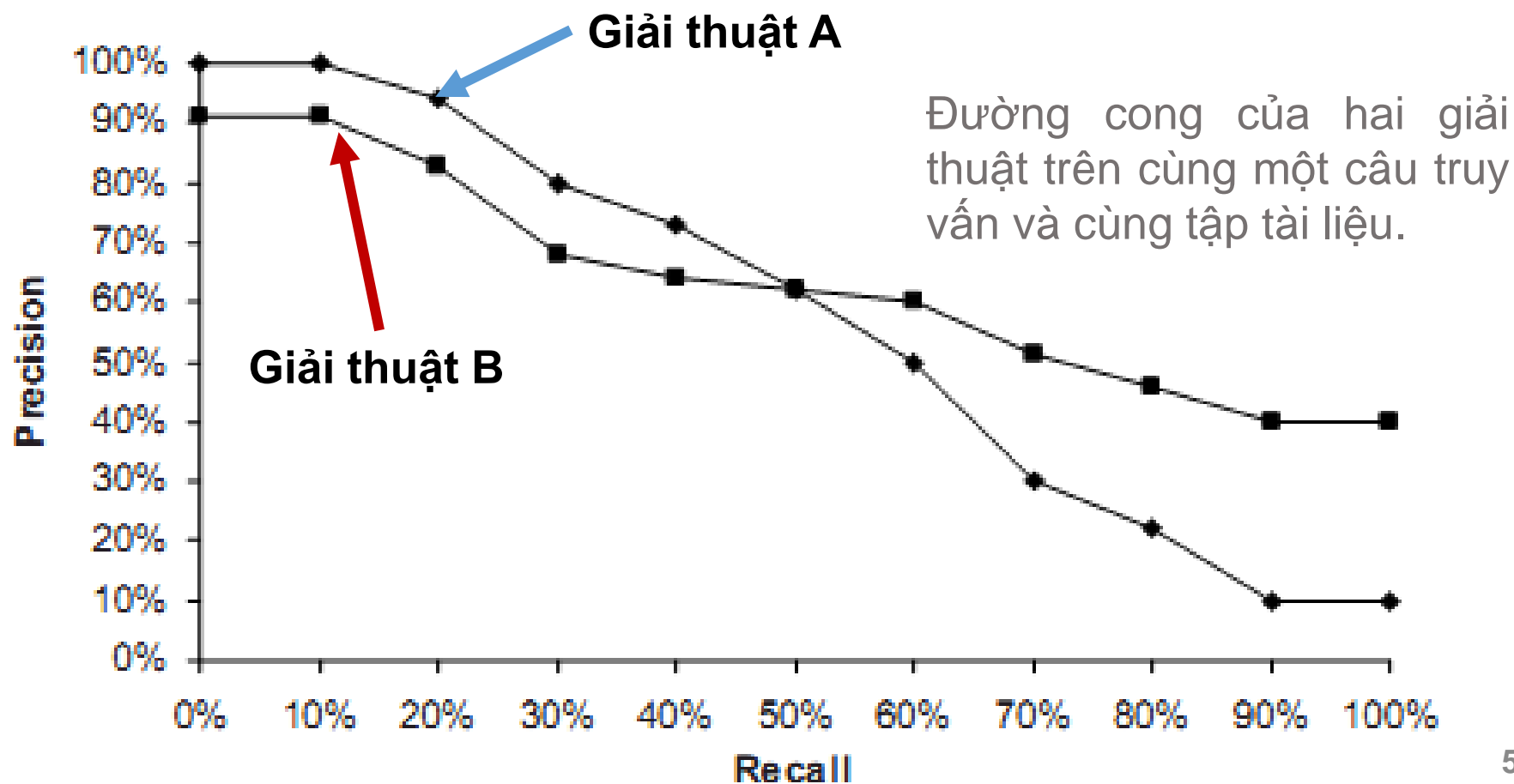
Bài tập 2: Đường cong precision-recall

- D chứa 9 tài liệu.
- Giả sử có 4 tài liệu thực sự liên quan đến q (kí hiệu dấu '+')
- Tính giá trị độ chính xác và độ phủ tại mỗi vị trí xếp hạng.
- Vẽ đường cong độ chính xác – độ phủ

| Rank | +/- |
|------|-----|
| 1 | + |
| 2 | - |
| 3 | + |
| 4 | + |
| 5 | - |
| 6 | - |
| 7 | - |
| 8 | + |
| 9 | - |

So sánh các giải thuật

- Độ chính xác của giải thuật truy vấn A tốt hơn của giải thuật truy vấn B ở mức phủ thấp nhưng lại kém hơn trong mức phủ cao.



Đánh giá trên nhiều truy vấn

- Hiệu quả của giải thuật truy vấn thường được đánh giá trên một lượng lớn câu truy vấn.
- **Độ chính xác toàn thể** (overall precision) tại mỗi mức phủ r_i là trung bình các giá trị độ chính xác tại mức phủ đó.

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

- trong đó Q là tập hợp truy vấn và $p_j(r_i)$ là độ chính xác của truy vấn j tại mức phủ r_i .

Nhận xét về độ chính xác – độ phủ

- Độ chính xác và độ phủ có sự đánh đổi (trade-off) lẫn nhau.
 - Ta thường đạt độ chính xác cao ở mức phủ thấp và khi tìm thấy càng nhiều tài liệu thật sự liên quan (độ phủ tăng) thì khả năng lẫn vào những tài liệu không liên quan cũng tăng (độ chính xác giảm).
- Hầu như không thể xác định D_q trên Web bởi vì có quá nhiều trang → không thể tính độ phủ
- Độ phủ không có ý nghĩa nhiều trong Tìm kiếm Web vì người dùng hiếm khi đến các trang xếp hạng dưới 30.
- Tuy nhiên, độ chính xác quan trọng hơn và do đó thường được tính cho các tài liệu xếp hạng đầu.

Độ chính xác xếp hạng

- Tính độ chính xác tại một vài vị trí xếp hạng chọn trước
- Đối với cỗ máy tìm kiếm Web, độ chính xác thường được tính cho 5, 10, 20, 25 và 30 trang đầu tiên được trả về.
 - Giả sử số trang liên quan lớn hơn 30.
- Ví dụ, từ bảng độ chính xác trước đó, ta có $p(5) = 80\%$, $p(10) = 70\%$, $p(15) = 53\%$ và $p(20) = 40\%$
- Độ chính xác không phải là độ đo duy nhất để xếp hạng kết quả tìm kiếm.
- Danh tiếng (reputation) hay chất lượng của trang được xếp hạng cũng là nhân tố quan trọng.

F-score và Breakeven point

- **F-score** tại vị trí xếp hạng i là trung bình điều hòa của độ chính xác và độ phủ.

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

- **Breakeven point** là điểm có độ chính xác bằng độ phủ.
 - Ví dụ, cho danh sách xếp hạng gồm 20 tài liệu, $p = r = 70\%$

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| + | + | + | - | + | - | + | - | + | + | - | - | + | - | - | - | + | - | - | + |

| | | |
|------------|-------------------|-------------------|
| At rank 1: | $p = 1/1 = 100\%$ | $r = 1/10 = 10\%$ |
|------------|-------------------|-------------------|

| | | |
|------------|-------------------|-------------------|
| At rank 2: | $p = 2/2 = 100\%$ | $r = 2/10 = 20\%$ |
|------------|-------------------|-------------------|

...

...

...

| | | |
|------------|--------------------|-------------------|
| At rank 9: | $p = 6/9 = 66.7\%$ | $r = 6/10 = 60\%$ |
|------------|--------------------|-------------------|

| | | |
|-------------|-------------------|-------------------|
| At rank 10: | $p = 7/10 = 70\%$ | $r = 7/10 = 70\%$ |
|-------------|-------------------|-------------------|

Bài tập 3: Các độ đo khác

- Dựa vào các giá trị độ phủ và độ chính xác đã tính ở Bài tập 2, tính các độ đo đánh giá dưới đây
 - Độ chính xác trung bình
 - F-score tại mỗi thứ hạng

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**.