



**ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN**

TRUY XUẤT THÔNG TIN

**Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn**

MỤC TIÊU MÔN HỌC

- 1) Hiểu các khái niệm, các vấn đề trong truy xuất thông tin
- 2) Áp dụng các kiến thức liên quan để đề xuất phương pháp phân tích văn bản.
- 3) Cài đặt thử nghiệm các mô hình truy xuất thông tin cơ bản
- 4) Đánh giá và so sánh các mô hình truy xuất thông tin
- 5) Xây dựng một Search Engine đơn giản.

NỘI DUNG MÔN HỌC

- ❖ CHƯƠNG I: DẪN NHẬP
- ❖ CHƯƠNG II: MÔ HÌNH KHÔNG GIAN VECTOR
- ❖ CHƯƠNG III: ĐÁNH GIÁ MÔ HÌNH TRUY XUẤT THÔNG TIN
- ❖ CHƯƠNG IV: XÂY DỰNG SEARCH ENGINE
- ❖ CHƯƠNG V: MỘT SỐ MÔ HÌNH TRUY XUẤT THÔNG TIN KHÁC
- ❖ CHƯƠNG VI: PHÂN LỚP VĂN BẢN

ĐÁNH GIÁ MÔN HỌC

- ❖ **ĐỒ ÁN:** 50% điểm tổng kết.
 - Thuyết trình theo nhóm: 50% điểm đồ án
 - Báo cáo + phần mềm: 50% điểm đồ án
- ❖ **THI CUỐI KỲ:** 50% điểm tổng kết

QUY ĐỊNH:

- Tham dự trên 80% số buổi học lý thuyết.
- Không thuyết trình thì không chấm điểm đồ án.
- Mỗi nhóm chỉ 2 đến 3 sinh viên. Nếu không tìm được nhóm có thể làm một mình nhưng không khuyến khích.

TÀI LIỆU

Christopher D. Manning, Prabhakar Raghavan and
Hinrich Schütze, *Introduction to Information Retrieval*,
Cambridge University Press, 2008.

THẢO LUẬN

- ❖ Trên lớp
- ❖ Website môn học trên moodle của trường:
 - Địa chỉ: courses.uit.edu.vn
 - Đăng nhập bằng tài khoản của nhà trường
 - Thảo luận tất cả các vấn đề về môn học
- ❖ Email cá nhân (không khuyến khích)
 - Địa chỉ: chinhht@uit.edu.vn
 - Chỉ sử dụng khi cần phải trao đổi khi không thể dùng website môn học.



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG I - DẪN NHẬP

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ GIỚI THIỆU
- ❖ CÁC MÔ HÌNH TRUY XUẤT THÔNG TIN CĂN BẢN
- ❖ LẬP CHỈ MỤC
- ❖ TRUY VÂN CHỈ MỤC

GIỚI THIỆU

❖ KHÁI NIỆM

Lập trình java



HỆ
THỐNG
TRUY
XUẤT
THÔNG
TIN

GIỚI THIỆU

❖ KHÁI NIỆM

Truy xuất thông tin là tìm kiếm

- Vật liệu chứa thông tin (tài liệu – Document).
- Phi cấu trúc hoặc bán cấu trúc: Free Text, XML
- Từ các tập lưu trữ lớn (Collections).
- Thỏa yêu cầu.

GIỚI THIỆU

❖ KHÁI NIỆM

Các dạng tài liệu:

- Email
- Tập tin trên máy tính cá nhân
- Hệ thống văn bản pháp lý
- Các cơ sở tri thức
- Hình ảnh
- Âm thanh
- Video
-

GIỚI THIỆU

❖ KHÁI NIỆM

Một số hệ thống truy xuất thông tin:

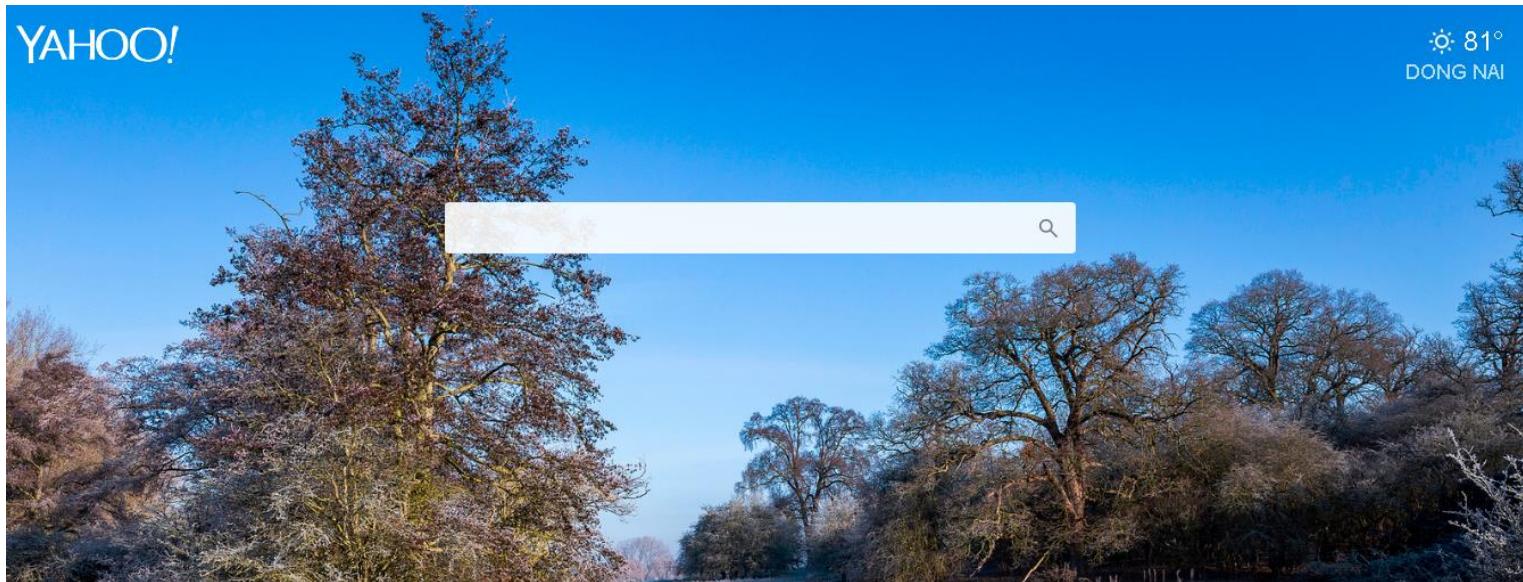
- Internet search engines:
 - Google

The screenshot shows a Google search results page. At the top, the search bar contains the word "google". Below the search bar, there are tabs for "All", "Videos", "Images", "Books", and "More", with "All" being the selected tab. To the right of these tabs are "Settings" and "Tools" buttons. A status message indicates "About 15,460,000,000 results (0.48 seconds)". The first result is a link to the Google homepage, labeled "Google" and "https://www.google.com/". Below the link, a snippet of text from the page describes Google Translate as "Google's free service" that "instantly translates words, phrases, and ...". At the bottom of the snippet is a blue link "More results from google.com »".

GIỚI THIỆU

❖ KHÁI NIỆM

- Internet search engines:
 - Yahoo! Web search



GIỚI THIỆU

❖ KHÁI NIỆM

- Internet search engines:
 - Bing



GIỚI THIỆU

❖ KHÁI NIỆM

- Danh bạ thư viện số: Yahoo!Directory

The screenshot shows the Yahoo! Directory homepage. At the top, there's a navigation bar with links for "Make Yahoo My Homepage", "Mail | My Yahoo | Yah", and a search bar. Below the bar is a large search input field with a magnifying glass icon and a yellow "Search Web" button. The main content area has a purple header bar with "Yahoo! Directory" on the left and "Advanced Search Suggest a Si" on the right. The main body contains a grid of categories:

Arts & Humanities Photography, History,	News & Media Newspapers, Radio,
Business & Economy B2B, Finance, Shopping,	Recreation & Sports Sports, Travel, Autos,
Computer & Internet Hardware, Software, Web,	Reference Phone Numbers, Dictionaries,
Education Colleges, K-12, Distance	Regional Countries, Regions, U.S.
Entertainment Movies, TV Shows, Music,	Science Animals, Astronomy, Earth
Government Elections, Military, Law,	Social Science Languages, Archaeology,
Health Disease, Drugs, Fitness,	Society & Culture Sexuality, Religion, Food &
New Additions 12/18, 12/17, 12/16, 12/15, 12/14...	

GIỚI THIỆU

❖ KHÁI NIỆM

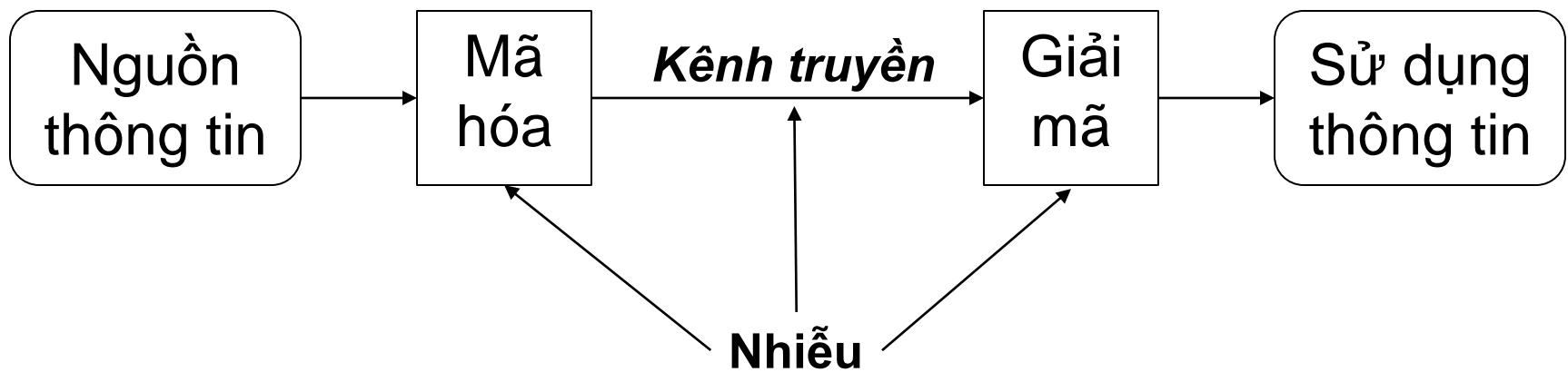
Một số ứng dụng của truy xuất thông tin:

- Truy xuất thông tin xuyên ngôn ngữ (Cross-lingual IR)
- Truy xuất giọng nói, bản tin của đài phát thanh.
- Phân loại văn bản
- Tóm tắt văn bản
- Truy xuất thông tin có cấu trúc (XML)
- Truy xuất thông tin địa lý (Geographic IR)

GIỚI THIỆU

❖ KHÁI NIỆM

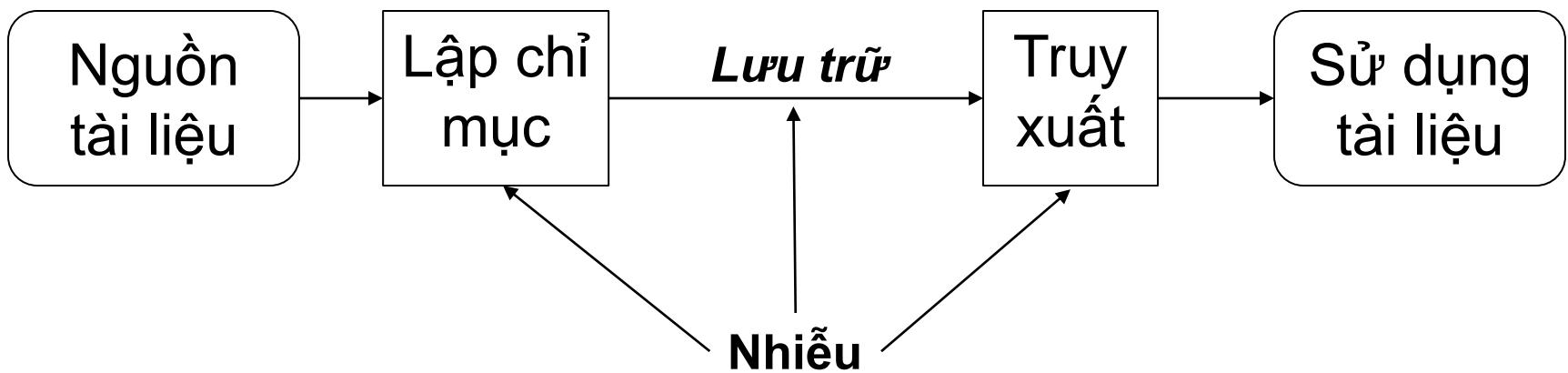
Mô hình truyền tin trong Lý thuyết thông tin:



GIỚI THIỆU

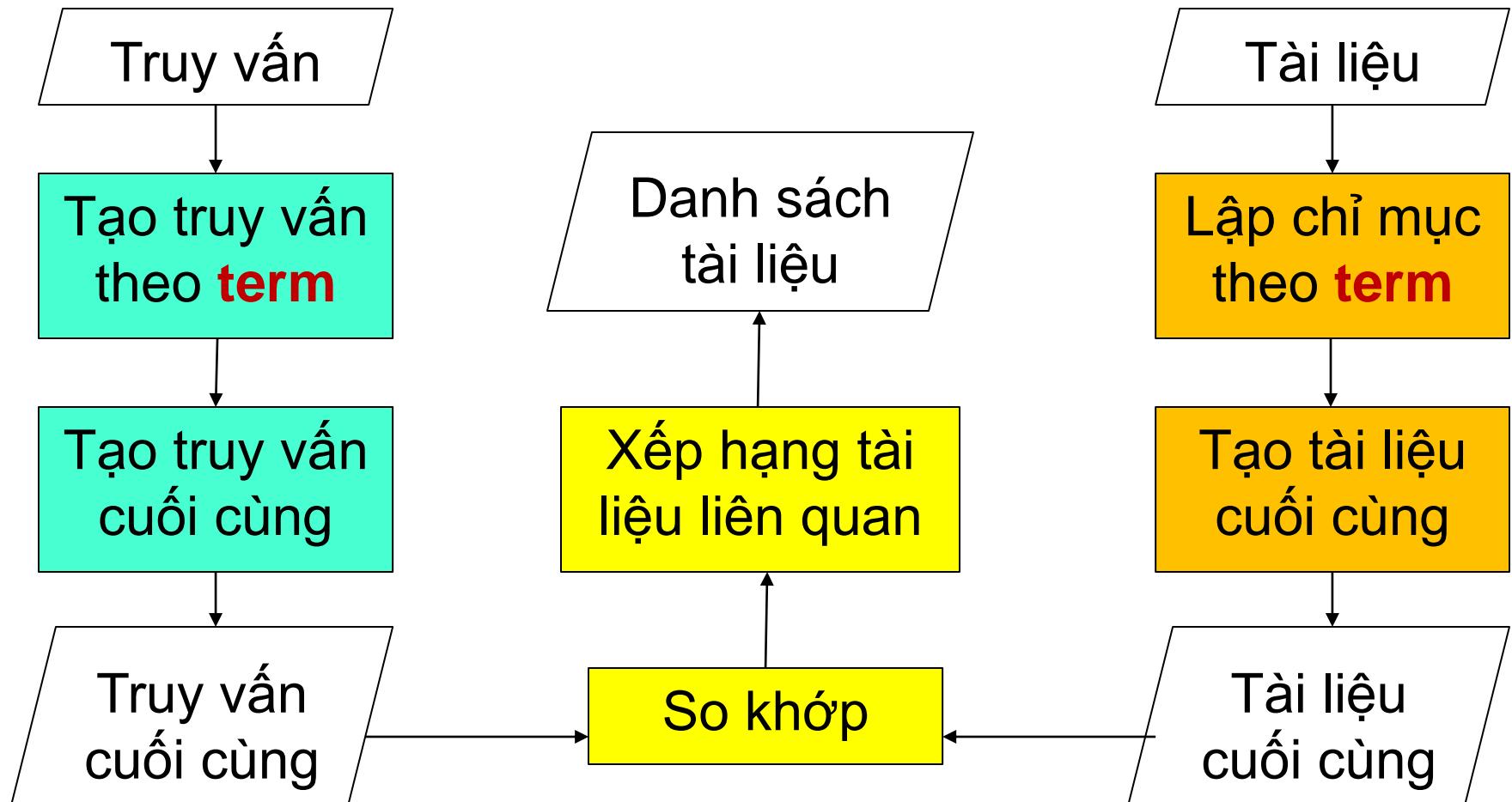
❖ KHÁI NIỆM

Truy xuất thông tin bắt nguồn từ lý thuyết thông tin:
Trong đó, chữ viết là dạng mã hóa của thông tin.



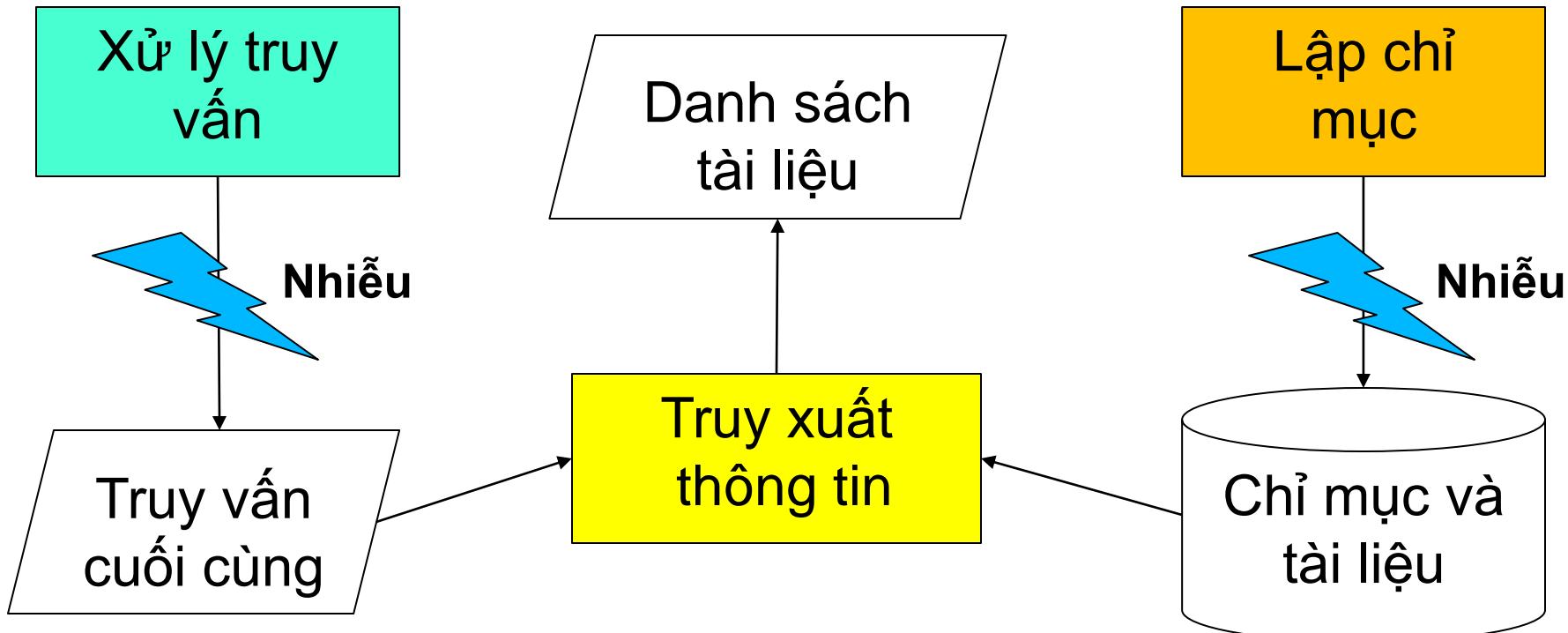
GIỚI THIỆU

❖ MÔ HÌNH ĐƠN GIẢN



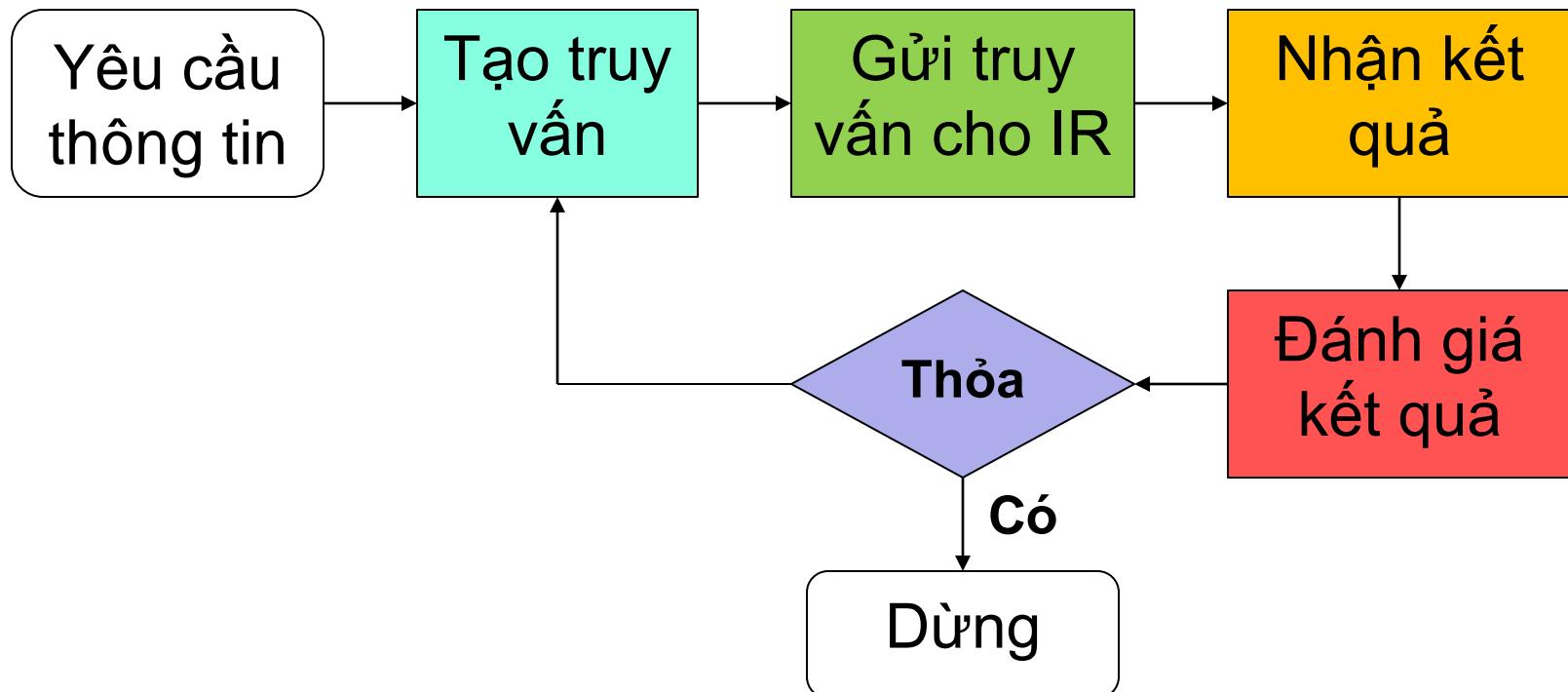
GIỚI THIỆU

❖ KIẾN TRÚC HỆ THỐNG ĐƠN GIẢN



GIỚI THIỆU

❖ SƠ ĐỒ TƯƠNG TÁC VỚI HỆ THỐNG IR



GIỚI THIỆU

❖ MÔ HÌNH CHUẨN CỦA IR

Mô hình chuẩn của IR được xây dựng dựa trên các giả thiết sau:

- Mục tiêu: nhằm cực đại hóa đồng thời độ phủ (recall) và độ chính xác (precision) của kết quả tìm kiếm
- Nhu cầu thông tin không thay đổi trong quá trình thực hiện
- Giá trị của mô hình là tập tài liệu được trả về.

GIỚI THIỆU

❖ MÔ HÌNH CHUẨN CỦA IR

Một số nhược điểm của mô hình chuẩn:

- Người sử dụng phải thực hiện các việc sau:
 - Đọc tựa của các tài liệu tìm được
 - Đọc tài liệu tìm được
 - Xem lại danh sách các chủ đề hoặc thuật ngữ liên quan
 - Phải duyệt các liên kết.
- Nhiều người sử dụng chỉ muốn trả về một vài tài liệu thực sự cần thiết.

GIỚI THIỆU

❖ LỊCH SỬ TRUY XUẤT THÔNG TIN

- Xuất phát từ danh mục các diển tích trong kinh thánh (được thực hiện bởi 500 tu sĩ vào năm 1247)
- Các hệ thống chỉ mục của các bài báo, bài viết từ thế kỷ 17.
- Sau chiến tranh thế giới thứ 2, Cranfield có nghiên cứu đầu tiên về ngôn ngữ chỉ mục và truy xuất thông tin.
- Năm 1951, kết quả nghiên cứu của Bagley cho thấy thời gian tìm kiếm trong tập 50 triệu tài liệu được lập chỉ mục với 30 từ vựng cần 41700 giờ.

GIỚI THIỆU

❖ LỊCH SỬ TRUY XUẤT THÔNG TIN

- Năm 1960, Mô hình truy xuất thông tin Boolean được đề xuất.
- Năm 1961, dịch vụ tìm kiếm abstract trong lĩnh vực hóa học trên máy tính được công bố. Hệ thống chỉ mục Medicus của Thư viện y khoa quốc gia được công bố dưới dạng cơ sở dữ liệu có tên MEDLARS
- Năm 1970, tất cả sản phẩm phụ, như hệ thống chỉ mục, các abstract, được xuất bản đều được làm từ máy tính nhờ cơ sở dữ liệu.

GIỚI THIỆU

❖ LỊCH SỬ TRUY XUẤT THÔNG TIN

- Năm 1970, hệ thống truy xuất thông tin SMART theo mô hình không gian vector được Salton công bố.
- Năm 1972, Hệ thống truy xuất thông tin pháp lý toàn văn theo mô hình Boolean được công bố với tên LEXIS.
- Những năm 70, truy vấn được phân tích theo xác suất.
- Những năm 80, các vấn đề về tập mờ, logic mờ và suy diễn được áp dụng trong truy xuất thông tin.

GIỚI THIỆU

❖ LỊCH SỬ TRUY XUẤT THÔNG TIN

- Những năm 90, các vấn đề về mạng neural, mạng suy diễn, chỉ mục ngữ nghĩa tiềm ẩn (Latent semantic index) được nghiên cứu và áp dụng vào truy xuất thông tin.

GIỚI THIỆU

❖ LỊCH SỬ TRUY XUẤT THÔNG TIN

Minh họa: thẻ được dùng để tìm tài liệu trong thư viện

EXCURSION										43821
90	241	52	63	34	25	66	17	58	49	
130	281	92	83	44	75	86	57	88	119	
640		122	93	104	115	146	97	158	139	
LUNAR										12457
110	181	12	73	44	15	46	7	28	39	
430	241	42	113	74	85	76	17	78	79	
820	761	602	233	134	95	136	37	118	109	
901	982		194	165			127	198	179	
							377	288		
							407			

Uniterm (Casey, Perry, Berry, Kent – 1958)

GIỚI THIỆU

❖ MỘT SỐ TẠP CHÍ, KỶ YẾU HỘI NGHỊ

- ACM Transaction on Information Systems
- Am. Society for Information Science Journal
- Document Analysis and IR Proceedings (Las Vegas)
- Information Processing and Management (Pergamon)
- Journal of Document
- SIGIR Conference Proceedings
- TREC Conference Proceedings



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG I - DẪN NHẬP

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ TRUY XUẤT THÔNG TIN
- ❖ CÁC MÔ HÌNH TRUY XUẤT THÔNG TIN CĂN BẢN
- ❖ LẬP CHỈ MỤC
- ❖ TẬP TỪ VỰNG VÀ DANH SÁCH “POSTING”
- ❖ TRUY VĂN CHỈ MỤC

CÁC MÔ HÌNH IR CĂN BẢN

❖ MỘT SỐ KHÁI NIỆM

- Term:
 - Là yếu tố cấu tạo thành tài liệu.
 - Được xác định tùy theo quan điểm phân tích tài liệu

Ví dụ:

Tài liệu là văn bản, quan điểm phân tích tài liệu là:

- Từ → term là từ. Ví dụ: *computer*, *science*
- Khái niệm → term là khái niệm. Ví dụ: *computer science*

CÁC MÔ HÌNH IR CĂN BẢN

❖ MỘT SỐ KHÁI NIỆM

- Biểu diễn (Representation):
 - Cấu trúc của tài liệu.
 - Được xác định tùy theo quan điểm phân tích tài liệu

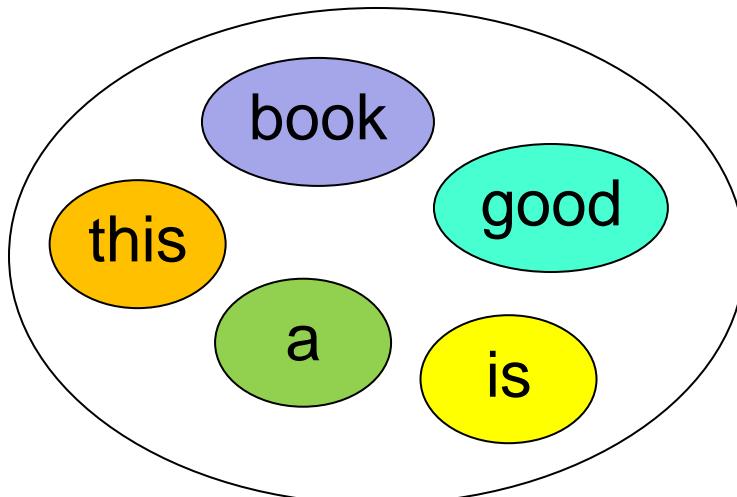
CÁC MÔ HÌNH IR CĂN BẢN

❖ MỘT SỐ KHÁI NIỆM

- Biểu diễn (Representation):

Ví dụ: Term là từ, Biểu diễn dạng tập hợp của tài liệu:

“this is a book. This book is good.”



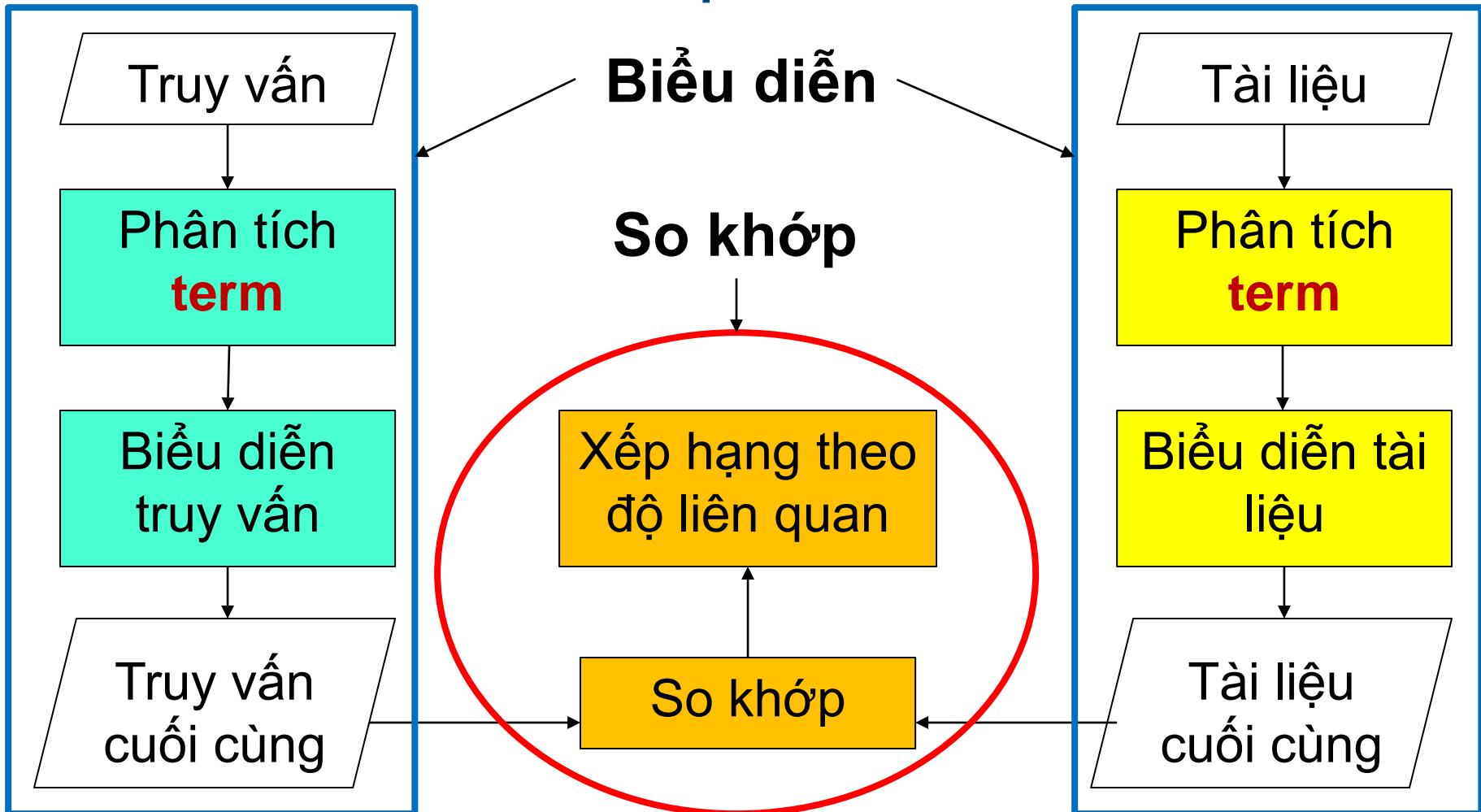
CÁC MÔ HÌNH IR CĂN BẢN

❖ MỘT SỐ KHÁI NIỆM

- So khớp (Matching):
 - Tính toán mức độ tương đồng giữa hai đối tượng.
 - Tùy thuộc vào độ đo sự tương đồng.

CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR



CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

Mô hình IR có hai đặc trưng:

- Biểu diễn tài liệu
- So khớp tài liệu

Trong đó

Biểu diễn tài liệu và phương pháp so khớp có mối liên hệ chặt chẽ với nhau để xác định mô hình IR

CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

- **Mô hình theo lý thuyết tập hợp:**
 - Biểu diễn: Tài liệu và truy vấn được biểu diễn dưới dạng tập hợp, là tập hợp các yếu tố cấu tạo nên chúng.
 - So khớp:
 - Tập hợp → mô hình tập hợp
 - Logic → mô hình Boolean
 - Logic có trọng số → mô hình Extended Boolean
 - Logic mờ → mô hình Fuzzy

CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

- **Mô hình đại số** (mô hình Vector)
 - Biểu diễn: tài liệu và truy vấn được biểu diễn bằng một vector trong không gian n chiều
 - So khớp: dựa vào các metric được định nghĩa trên không gian tài liệu:
 - Khoảng cách: chuẩn Euclidean
 - Góc giữa hai vector: chuẩn Cauchy
 - ...

CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

• Mô hình xác suất

- Biểu diễn: tài liệu được biểu diễn bằng đặc trưng phân phối của các chủ đề (topic).
- So khớp: sự khác biệt giữa phân phối xác suất của các chủ đề trong tài liệu và trong truy vấn.

CÁC MÔ HÌNH IR CĂN BẢN

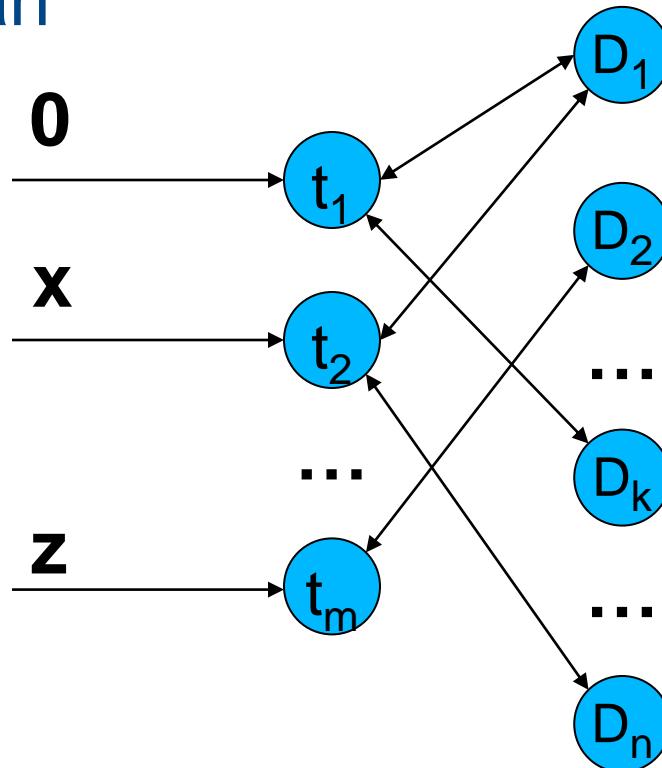
❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

- **Mô hình mạng neural**
- Biểu diễn: tài liệu được biểu diễn bằng một mạng hai lớp trong đó lớp input là các term và lớp output là các tài liệu.
- So khớp: dựa trên quá trình tính toán, tổng hợp giá trị tại các nút trên mạng theo cơ chế lan truyền.

CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

Giá trị node D_i là mức độ tương đồng giữa tài liệu D_i và truy vấn



CÁC MÔ HÌNH IR CĂN BẢN

❖ CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

Các mô hình IR căn bản gồm:

- Mô hình tập hợp
- Mô hình Boolean
- Mô hình Extended Boolean

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH TẬP HỢP

➤ Truy vấn là từng mô tả đơn lẻ.

Ví dụ: Các truy vấn

1) *toán lớp 12*

2) *hình vẽ*

Truy vấn 2 sau truy vấn 1 không có nghĩa “*toán hình lớp 12*”

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH TẬP HỢP

- Quy tắc truy hồi: có hai dạng
 - Truy hồi theo phép toán chứa trong (\subseteq):
 - Biểu diễn của truy vấn và tài liệu lần lượt là tập hợp P và Q, khi đó Q thỏa P nếu $Q \subseteq P$.
 - Là mô hình đơn giản nhất, thường được sử dụng trong các thư viện cho phép chọn lựa sách theo từng nội dung đơn lẻ.
 - Tài liệu được xác định là thỏa hay không thỏa truy vấn \rightarrow không xếp hạng được.

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH TẬP HỢP

Ví dụ:

Truy vấn: *hình học*

Tài liệu 1: *hình học, vuông góc, mặt phẳng*

Tài liệu 2: *mặt phẳng, vuông góc, lực*

→ trả về tài liệu 1

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH TẬP HỢP

- Truy hồi theo phép toán giao (\cap):
 - Biểu diễn của truy vấn tài liệu lần lượt là tập hợp P và Q, khi đó Q thỏa P nếu:
 - ✓ $P \cap Q = R$, và
 - ✓ $|R| > C$, với C là một hằng số nguyên
 - Có thể dựa vào số phần tử của R để xếp hạng.

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH TẬP HỢP

Ví dụ: với $C = 2$,

Truy vấn: *sinh học phân tử*

Tài liệu 1: *học sinh, giáo viên, sách giáo khoa*

Tài liệu 2: *hóa học, đồng phân, nguyên tử*

→ trả về tài liệu 2 với $|R| = 3 > 2$.

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH TẬP HỢP

Cho tập tài liệu:

d_1 : *máy tính, bộ nhớ, bàn phím*

d_2 : *bộ nhớ, con trỏ, phép tính*

d_3 : *con mèo, nhà bếp, bàn ăn*

phân tích tài liệu theo từng tiếng trong tiếng Việt.

Tìm tài liệu cho truy vấn: “*phép tính bộ nhớ*” với:

- Mô hình tập hợp theo phép chứa trong
- Mô hình tập hợp theo phép giao với $C = 2$

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH BOOLEAN

Dựa trên một trong các dạng logic sau:

- Logic mệnh đề
- Logic vị từ
- Logic mô tả
- Logic mờ
-

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH BOOLEAN

- Tri thức: tất cả hạng tử có thể có trong tập tài liệu
- Truy vấn là những mô tả đơn lẻ gồm một tổ hợp các hạng tử và các phép toán logic AND, OR, NOT.

Ví dụ: Truy vấn được biểu diễn theo logic mệnh đề toán **AND** phô **AND** thông

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH BOOLEAN

➤ Tài liệu:

- Tập các hạng tử, có được từ việc phân tích term của tài liệu, liên kết với nhau bằng toán tử AND
- Không chứa hạng tử $t \Leftrightarrow$ chứa hạng tử $\neg t$

Ví dụ: Tài liệu được biểu diễn theo logic mệnh đề

Tài liệu: máy tính, bộ nhớ, bàn phím

→ máy AND tính AND bộ AND nhớ AND bàn AND phím AND \neg con AND \neg trở AND ...

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH BOOLEAN

➤ Quy tắc truy hồi tài liệu

- Nếu truy vấn có dạng $t_1 \wedge t_2$: trả về tài liệu có cả t_1 và t_2
- Nếu truy vấn có dạng $t_1 \vee t_2$: trả về tài liệu có một trong hai term t_1 và t_2
- Nếu truy vấn có dạng $\neg t_1$: trả về tài liệu không chứa t_1

→ Tài liệu D thỏa Q nếu $D \models Q$

→ Không thể xếp hạng tài liệu trả về

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH BOOLEAN

Ví dụ:

Giả sử có tri thức: $\{t_1, t_2, t_3, t_4, t_5, t_6\}$

Truy vấn: $Q_1 = t_1 \wedge t_2,$

$$Q_2 = (t_1 \wedge t_2 \vee t_3) \wedge (t_4 \vee \neg(t_5 \wedge t_6))$$

Tài liệu: $D_1 = \{t_1, t_2\},$

$$D_2 = \{t_1, t_2, t_3, t_4\}$$

Tính truy kết quả các truy vấn.

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH FUZZY LOGIC

- Là mô hình mở rộng của mô hình Boolean,
 - Tính toán được mức độ liên quan của từng tài liệu dựa trên giá trị chân lý của hạng tử
- Biểu diễn tài liệu:
- Tương tự mô hình Boolean
 - Có thêm trọng số của w_t từng hạng tử t cho biết giá trị chân lý của t

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH FUZZY LOGIC

- Quy tắc so khớp:
- Dựa trên logic mờ

- $\text{Sim}_D(t_1 \vee t_2) = \max(w^D_{t1}, w^D_{t2})$
- $\text{Sim}_D(t_1 \wedge t_2) = \min(w^D_{t1}, w^D_{t2})$
- $\text{Sim}_D(\neg t_1) = 1 - w^D_{t1}$

→ Nhược điểm: không sử dụng giá trị của tất cả hạng tử của truy vấn.

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH EXTENDED BOOLEAN

- Là mô hình mở rộng của mô hình Boolean,
 - Tính toán được mức độ liên quan của từng tài liệu
- Biểu diễn tài liệu:
- Tương tự mô hình Boolean
 - Có thêm trọng số của w_t từng hạng tử t

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH EXTENDED BOOLEAN

- Quy tắc so khớp: dựa trên độ tương đồng có sử dụng giá trị của tất cả hạng tử của truy vấn theo các công thức sau

- $\text{Sim}_D(t_1 \vee t_2) = \sqrt{\frac{w_{t_1}^2 + w_{t_2}^2}{2}}$

- $\text{Sim}_D(t_1 \wedge t_2) = 1 - \sqrt{\frac{(1-w_{t_1})^2 + (1-w_{t_2})^2}{2}}$

- $\text{Sim}_D(\neg t_1) = 1 - w_{t_1}$

CÁC MÔ HÌNH IR CĂN BẢN

❖ MÔ HÌNH EXTENDED BOOLEAN

Giả sử có tri thức $K = \{t_1, t_2\}$

Các tài liệu theo K :

$$D_1 = \{t_1, t_2\}, D_2 = \{t_1\}, D_3 = \{t_2\}, D_4 = \emptyset$$

Tính mức độ liên quan giữa các tài liệu này với truy vấn

$$Q_1 = t_1 \wedge t_2$$

$$Q_1 = t_1 \vee t_2$$

Theo các mô hình Boolean và Extended Boolean (với trọng số của hạng tử xuất hiện là 1, không xuất hiện là 0)



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG I - DẪN NHẬP

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

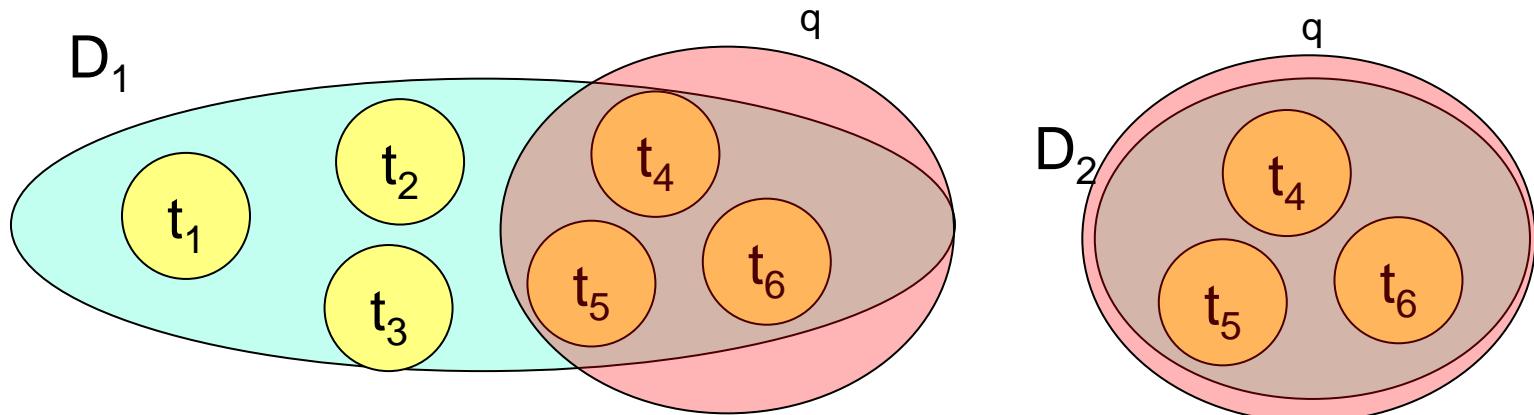
- ❖ TRUY XUẤT THÔNG TIN
- ❖ CÁC MÔ HÌNH TRUY XUẤT THÔNG TIN CĂN BẢN
- ❖ LẬP CHỈ MỤC
- ❖ TẬP TỪ VỰNG VÀ DANH SÁCH “POSTING”
- ❖ TRUY VĂN CHỈ MỤC

LẬP CHỈ MỤC

❖ MỤC ĐÍCH CỦA VIỆC LẬP CHỈ MỤC

Việc tính toán trong các mô hình

- Tập hợp: phải biểu diễn tài liệu và truy vấn thành tập hợp các term



LẬP CHỈ MỤC

❖ MỤC ĐÍCH CỦA VIỆC LẬP CHỈ MỤC

Việc tính toán trong các mô hình

- Boolean: phải biểu diễn tài liệu và truy vấn thành biểu thức logic

$$D_1: t_1 \wedge t_2 \wedge t_3 \wedge t_4 \wedge t_5 \wedge t_6$$

$$D_2: \neg t_1 \wedge \neg t_2 \wedge \neg t_3 \wedge t_4 \wedge t_5 \wedge t_6$$

$$q: t_4 \wedge t_5 \wedge t_6$$

LẬP CHỈ MỤC

❖ MỤC ĐÍCH CỦA VIỆC LẬP CHỈ MỤC

Việc tính toán trong các mô hình

Thao tác chung của cả hai mô hình: kiểm tra một phần tử có xuất hiện trong một danh sách (các phần tử hoặc hạng tử).

→ Một số vấn đề:

- Phải duyệt nội dung của từng tài liệu
 - Phải lưu trữ nội dung của từng tài liệu
- Không hiệu quả

LẬP CHỈ MỤC

❖ BIỂU DIỄN TÀI LIỆU

Dùng bảng biểu diễn Doc theo Term (Ma trận tài liệu Doc-Term):

Term Doc	t_1	t_2	t_3	t_4
d_1	1	1	1	1
d_2	1	1	0	0
d_3	1	0	0	0
d_4	0	0	1	0
d_5	0	1	0	1

LẬP CHỈ MỤC

❖ BIỂU DIỄN TÀI LIỆU

Giả sử dùng mô hình tập hợp với phép chúa
trong, truy vấn $q=\{t_1, t_2\}$

d	1	1	1	1
q	1	1	0	0
$q \cap d = q_t \times d_t$	1	1	0	0
Sim(d₁, q)	1			

LẬP CHỈ MỤC

❖ BIỂU DIỄN TÀI LIỆU

Để tăng hiệu quả truy xuất: đảo ngược bảng với dòng và cột → chỉ mục đảo ngược (Inverted Index)

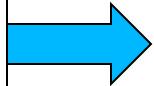
Doc Term \ Doc Term	d_1	d_2	d_3	d_4	d_5
t_1	1	1	1	0	0
t_2	1	1	0	0	1
t_3	1	0	0	1	0
t_4	1	0	0	0	1

LẬP CHỈ MỤC

❖ QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

Với mỗi tài liệu, tách các từ khóa để tạo thành danh sách gồm từ khóa và chỉ số tài liệu. Chỉ số tài liệu là thứ tự mà tài liệu đó được xử lý.

mục tài
liệu lập
chỉ mục



Từ khóa	Chỉ số tài liệu
mục	1
tài	1
liệu	1
lập	1
chỉ	1
mục	1

chỉ mục
từ khóa



Từ khóa	Chỉ số tài liệu
chỉ	2
mục	2
từ	2
khóa	2

LẬP CHỈ MỤC

❖ QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

Nối tất cả danh sách và sắp xếp theo từ khóa, chỉ số

Từ khóa	Chỉ số tài liệu
mục	1
tài	1
liệu	1
lập	1
chỉ	1
mục	1
chỉ	2
mục	2
từ	2
khóa	2



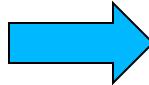
Từ khóa	Chỉ số tài liệu
chỉ	1
chỉ	2
khóa	2
lập	1
liệu	1
mục	1
mục	1
mục	2
tài	1
từ	2

LẬP CHỈ MỤC

❖ QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

Gom từ khóa có cùng chỉ số tài liệu và thêm tần số

Từ khóa	Chỉ số tài liệu
chỉ	1
chỉ	2
khóa	2
lập	1
liệu	1
mục	1
mục	1
mục	2
tài	1
từ	2



Từ khóa	Chỉ số tài liệu	Tần số
chỉ	1	1
chỉ	2	1
khóa	2	1
lập	1	1
liệu	1	1
mục	1	2
mục	2	1
tài	1	1
từ	2	1

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

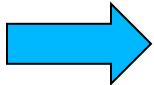
❖ KHÁI NIỆM

- Tập từ vựng (còn gọi là từ điển - Dictionary) gồm các thông tin: từ vựng, số lượng tài liệu chứa từ vựng và số lần xuất hiện của từ vựng đó trong toàn bộ tập lưu trữ. Mục đích để tra cứu dễ dàng.
- Danh sách Posting: chứa chỉ số tài liệu và số lần xuất hiện của một từ khóa trong tài liệu đó. Những dòng trong danh sách posting được trả tới bởi những mục trong tập từ vựng

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ TẠO TẬP TỪ VỰNG VÀ DS POSTING

Từ khóa	Chỉ số tài liệu	Tân số
chỉ	1	1
chỉ	2	1
khóa	2	1
lập	1	1
liệu	1	1
mục	1	2
mục	2	1
tài	1	1
từ	2	1



Tập từ vựng

Từ khóa	số tài liệu	Tân số
chỉ	2	2
khóa	1	1
lập	1	1
liệu	1	1
mục	2	3
tài	1	1
từ	1	1

DS Posting

Chỉ số tài liệu	Tân số
1	1
2	1
2	1
1	1
1	1
1	1
1	2
2	1
1	1
2	1

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ TẠO TẬP TỪ VỰNG VÀ DS POSTING

Tập từ vựng và danh sách posting có thể được lưu trữ theo nhiều cách khác nhau như danh sách liên kết, bảng băm, Btree.

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ VĂN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Từ vựng là đơn vị cơ bản cấu tạo thành tài liệu. Văn đề xác định tập từ vựng ảnh hưởng đến khả năng tìm kiếm tài liệu liên quan đến truy vấn.

VD: Cho các tài liệu sau:

d1: sun flowers

d2: a rose is a flower

d3: a lady in rose

Cho biết kết quả của truy vấn sau:

q1: a flower

q2: rose

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ VĂN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Trường hợp: sử dụng mô hình tập hợp:

- Phép toán chứa trong.
- Từ vựng được cách nhau bằng khoảng trắng

$$d_1 = \{\text{sun, flowers}\} \quad d_2 = \{\text{a, rose, is, flower}\}$$

$$d_3 = \{\text{a, lady, in, rose}\}$$

$$q_1 = \{\text{a, flower}\} \quad q_2 = \{\text{rose}\}$$

- $q_1 \not\subset d_1, q_1 \subset d_2, q_1 \not\subset d_3 \rightarrow d_2$
- $q_2 \not\subset d_1, q_2 \subset d_2, q_1 \subset d_3 \rightarrow d_2, d_3$

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Trường hợp: sử dụng mô hình tập hợp:

- Phép toán giao với $c = 1$.
- Từ vựng được cách nhau bằng khoảng trắng

$$d_1 = \{\text{sun, flowers}\} \quad d_2 = \{\text{a, rose, is, flower}\}$$

$$d_3 = \{\text{a, lady, in, rose}\}$$

$$q_1 = \{\text{a, flower}\} \quad q_2 = \{\text{rose}\}$$

- $|q_1 \cap d_1| = 0$, $|q_1 \cap d_2| = 2$, $|q_1 \cap d_3| = 1 \rightarrow d_2, d_3$
- $|q_2 \cap d_1| = 0$, $|q_2 \cap d_2| = 1$, $|q_2 \cap d_3| = 1 \rightarrow d_2, d_3$

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ VĂN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Trường hợp: sử dụng mô hình tập hợp:

- Phép toán giao với $c = 1$.
- Từ vựng là từ gốc, cách nhau bằng khoảng trắng.

$$d_1 = \{\text{sun, flower}\} \quad d_2 = \{\text{a, rose, be, flower}\}$$

$$d_3 = \{\text{a, lady, in, rose}\}$$

$$q_1 = \{\text{a, flower}\} \quad q_2 = \{\text{rose}\}$$

- $|q_1 \cap d_1| = 1$, $|q_1 \cap d_2| = 2$, $|q_1 \cap d_3| = 1 \rightarrow d_1, d_2, d_3$
- $|q_2 \cap d_1| = 0$, $|q_2 \cap d_2| = 1$, $|q_2 \cap d_3| = 1 \rightarrow d_2, d_3$

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Vấn đề:

- d_1 và d_2 đều thỏa thông tin của q_1
- Chỉ có d_2 thỏa thông tin của q_2

Nguyên nhân: chọn từ vựng chưa phù hợp.

TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

❖ VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

→ Chú ý các vấn đề về ngôn ngữ khi xác định tập từ vựng:

- Hình thái của từ (số, thì, thể, ..., ranh giới từ) – **stemming / lemmatizing**
- Những từ chủ yếu giữ chức năng ngữ pháp (mạo từ, định từ, giới từ, tình thái, trợ từ, ...) – **stopword removal**
- Ngữ nghĩa của từ (từ đồng âm, từ đồng nghĩa) – **query expansion.**

(Các vấn đề này cần được trình bày khi thuyết trình)

TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH TẬP HỢP

- Bước xây dựng tập từ vựng và danh sách posting: không cần thông tin tần số.
- Bước truy vấn chỉ mục: tích lũy số lần xuất hiện tài liệu theo từng từ vựng trong truy vấn

TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH TẬP HỢP

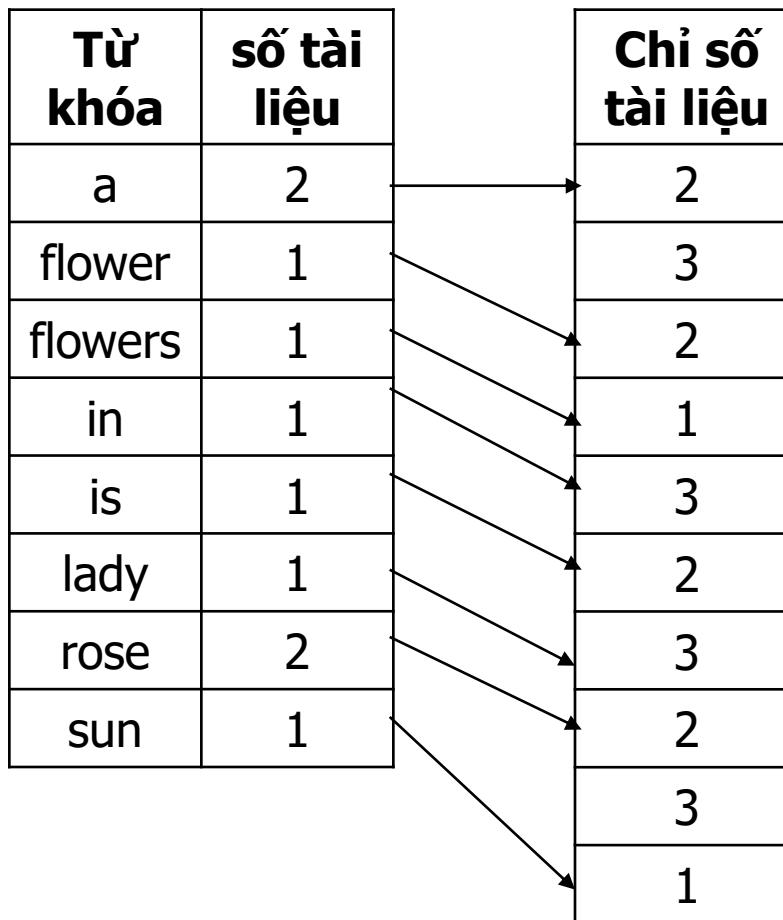
Các bước truy vấn chỉ mục:

- Danh sách kết quả ban đầu rỗng. Mỗi phần tử trong danh sách gồm chỉ số tài liệu và số lần xuất hiện.
- Với từng từ vựng q_i trong truy vấn q :
 - Xác định danh sách tài liệu
 - Cộng 1 số lần xuất hiện tài liệu tương ứng
- Chọn những tài liệu có số lần xuất hiện lớn hơn hoặc bằng c (trường hợp phép toán giao) hoặc $|q|$ (trường hợp phép toán chứa trong)

TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH TẬP HỢP

Ví dụ:



TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH TẬP HỢP

Ví dụ:

truy vấn $q_1 = \{a, \text{flower}\}$, $R = \{\}$

- Truy vấn "a" trên chỉ mục:

$$\begin{array}{ll} R \rightarrow \{\} & \\ "a" \rightarrow \{2,3\} & \end{array} \quad \left. \right\} \quad R \rightarrow \{(2,1), (3,1)\}$$

- Truy vấn "flower" trên chỉ mục:

$$\begin{array}{ll} R \rightarrow \{(2,1), (3,1)\} & \\ "flower" \rightarrow \{2\} & \end{array} \quad \left. \right\} \quad R \rightarrow \{(2,2), (3,1)\}$$

TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH BOOLEAN

- Bước xây dựng tập từ vựng và danh sách posting: không cần thông tin tần số.
- Bước truy vấn chỉ mục: xác định phép toán $d \rightarrow q$ theo cách:
 - Nếu $q = t$ thì $d \rightarrow q$ nếu $t \in d$
 - Nếu $q = t_1 \wedge t_2$ thì $d \rightarrow q$ nếu $t_1 \in d \wedge t_2 \in d$
 - Nếu $q = t_1 \vee t_2$ thì $d \rightarrow q$ nếu $t_1 \in d \vee t_2 \in d$
 - Nếu $q = \neg t$ thì $d \rightarrow q$ nếu $\neg(t \in d)$

TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH BOOLEAN

Các bước truy vấn chỉ mục:

- Áp dụng luật De Morgan chuyển biểu thức logic e của q thành dạng OR các thành phần AND.
- Phân tích biểu thức e dưới dạng cây:
 - Nút lá là các term
 - Nút trong là các phép toán: có các phép toán OR, AND, AND NOT (không có OR NOT).

TRUY VẤN CHỈ MỤC

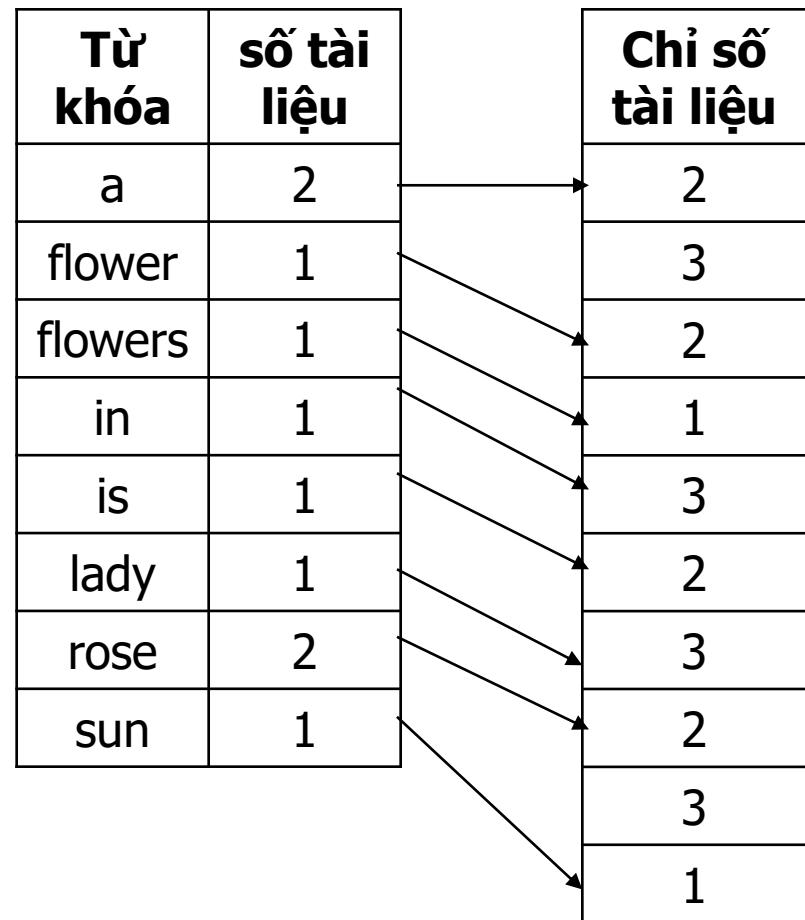
❖ MÔ HÌNH BOOLEAN

- Truy vấn chỉ mục cho các term để xác định danh sách tài liệu tương ứng
- Thực hiện các phép toán \cap , \cup trên các danh sách tương ứng với các phép toán \wedge , \vee

TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH BOOLEAN

Ví dụ: truy vấn:
sun AND NOT (a OR flower)



TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH BOOLEAN

Ví dụ: truy vấn:

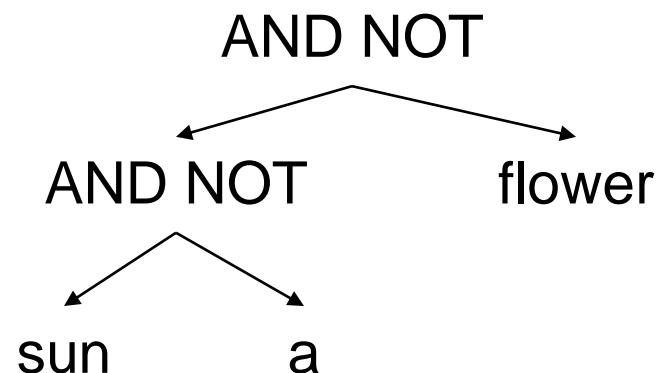
sun AND NOT (a OR flower)

- Chuyển về dạng OR:

sun AND (NOT a AND NOT flower)

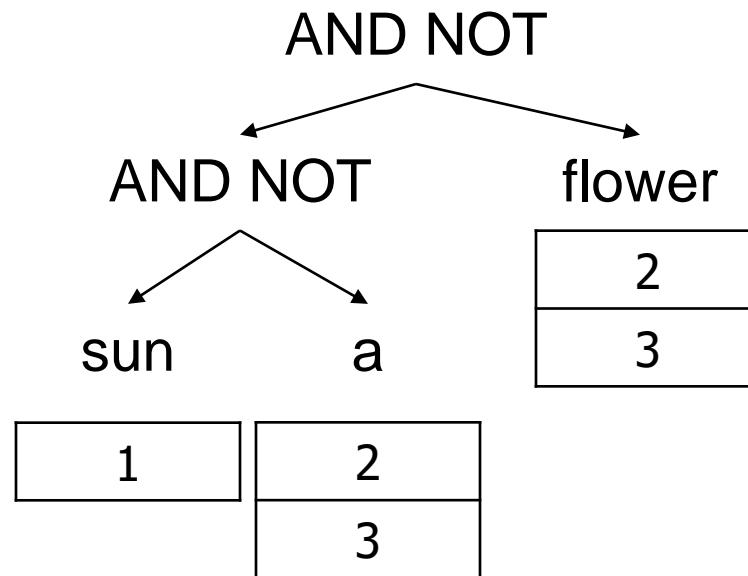
= sun AND NOT a AND NOT flower

- Chuyển thành dạng cây



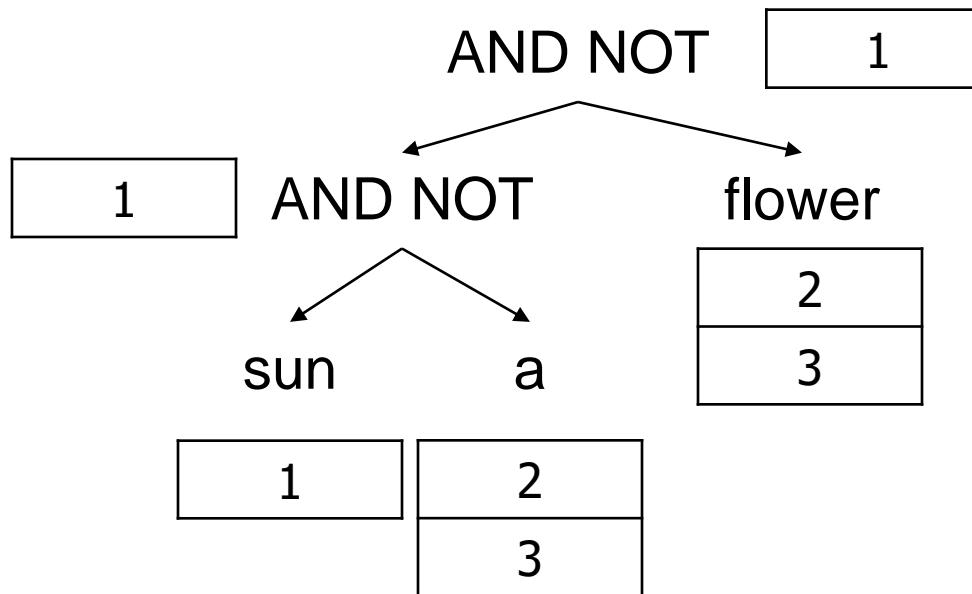
TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH BOOLEAN



TRUY VẤN CHỈ MỤC

❖ MÔ HÌNH BOOLEAN



BÀI TẬP

Cho tập tài liệu như sau:

- d1: sự thực hiện nay còn nhiều khó khăn
- d2: thực hiện quyết tâm vượt khó
- d3: hiện nay lượng khăn còn rất ít

Cho truy vấn sau:

q: lượng khăn hiện nay

Yêu cầu:

- 1) Xác định từ vựng cần phân tích
- 2) Xây dựng chỉ mục đảo ngược cho tập tài liệu
- 3) Xác định kết quả truy vấn. Cho biết kết quả truy vấn có phù hợp với mục đích truy vấn hay không?

BÀI TẬP

- 4) Nếu kết quả xác định được ở câu 3 chưa thỏa thì làm cách nào để cải thiện kết quả?



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG II – MÔ HÌNH KHÔNG GIAN VECTOR

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ VĂN ĐỀ TRONG MÔ HÌNH BOOLEAN
- ❖ MÔ HÌNH VECTOR
- ❖ TRỌNG SỐ CỦA TERM
- ❖ LẬP CHỈ MỤC
- ❖ TRUY VĂN TRÊN MÔ HÌNH KHÔNG GIAN VECTOR

VĂN ĐỀ TRONG MÔ HÌNH BOOLEAN

❖ MỨC ĐỘ QUAN TRỌNG CỦA TERM

- Mức độ quan trọng của term trong tài liệu chưa được thể hiện.
- Các giá trị chỉ gồm
 - + 0: không xuất hiện
 - + 1: có xuất hiện
- Việc sắp xếp thứ tự theo độ liên quan của các tài liệu tìm được không rõ ràng.

Ma trận tài liệu (Doc-Term)

DOC	t_1	t_2	t_3	t_4
d_1	1	1	1	1
d_2	1	1	0	0
d_3	1	0	0	0
d_4	0	0	1	0
d_5	0	1	0	1

VẤN ĐỀ TRONG MÔ HÌNH BOOLEAN

❖ KẾT QUẢ TÌM KIẾM

- Trong trường hợp không thỏa biểu thức logic, không đưa ra được tài liệu nào có thể dùng được

Vd: computer AND science AND programming

Kết quả Những tài liệu chỉ chứa computer và programming nhưng không có science sẽ bị loại.

- Cần có phương pháp tính toán sự tương đồng giữa truy vấn và tài liệu

MÔ HÌNH VECTOR

❖ ĐẶC ĐIỂM CỦA MÔ HÌNH VECTOR

- Các tài liệu được biểu diễn dưới dạng “bag of words”.
- Tài liệu được tính toán như một vector với các đặc điểm của nó:
 - + Tài liệu là một mảng số thực giống như một vector nhiều chiều.
 - + Mỗi term là một chiều trong không gian. (các vector thường thừa)
 - + Tính toán dựa trên hướng và độ lớn.

MÔ HÌNH VECTOR

❖ PHƯƠNG PHÁP SO KHỚP

- Truy vấn được biểu diễn bằng một vector cùng không gian với tài liệu
- Tính toán giữa truy vấn và tài liệu dựa trên chiều dài và hướng của vector tương ứng của chúng.
- Khoảng cách giữa hai vector tài liệu và truy vấn được xem là độ tương đồng giữa tài liệu và truy vấn. Khoảng cách này được dùng để xếp hạng tài liệu.

MÔ HÌNH VECTOR

❖ ĐẶC ĐIỂM CỦA VECTOR TÀI LIỆU

- Vị trí các term tạo thành không gian không quan trọng
- Giá trị mỗi chiều của vector tài liệu hay truy vấn là trọng số của term trong tài liệu hay truy vấn đó

Ma trận tài liệu (Doc-Term)

DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
d_1	8	20	2	0
d_2	10	0	31	0
d_3	1	0	42	14
d_4	0	3	0	3
d_5	0	21	9	1

MÔ HÌNH VECTOR

❖ ĐẶC ĐIỂM CỦA VECTOR TÀI LIỆU

Ma trận tài liệu (Doc-Term)

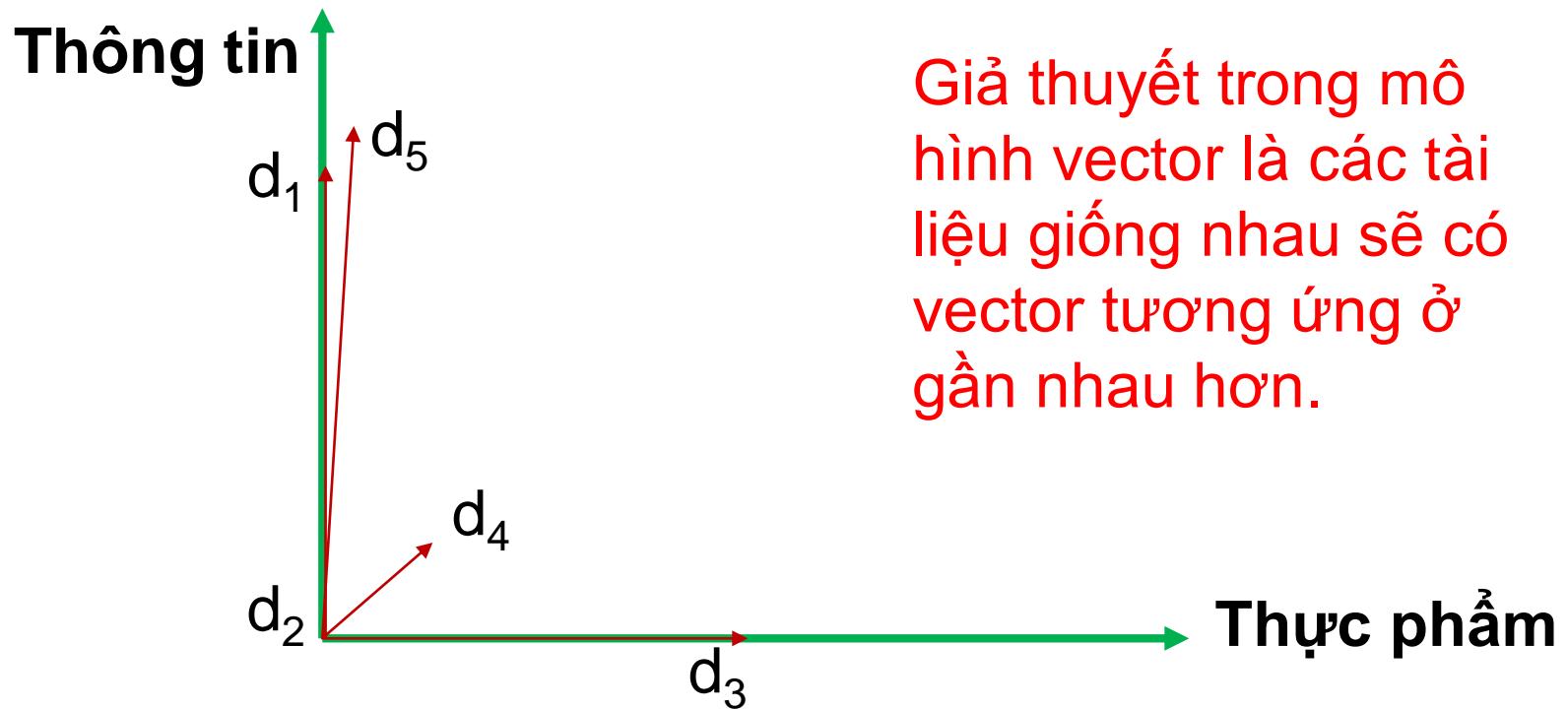
DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
d ₁	8	20	2	0

- Với cách tính trọng số là tần số xuất hiện của từ khóa trong tài liệu thì vector d1 cho biết:
 - + Từ “truy xuất” xuất hiện 8 lần
 - + Từ “thông tin” xuất hiện 20 lần
 - + Từ “công nghệ” xuất hiện 2 lần
 - + Từ “thực phẩm” không xuất hiện.

MÔ HÌNH VECTOR

❖ ĐẶC ĐIỂM CỦA VECTOR TÀI LIỆU

- Hình ảnh của các vector tài liệu trong mặt phẳng gồm 2 chiều “thông tin” và “thực phẩm”:



MÔ HÌNH VECTOR

❖ ĐỘ TƯƠNG ĐỒNG

- Độ tương đồng của hai vector được tính là cosine của góc tạo bởi hai vector

$$\cos(a, b) = \sum_t \frac{w_t^a \times w_t^b}{\sqrt{\sum_t w_t^{a^2}} \times \sqrt{\sum_t w_t^{b^2}}}$$

Trong đó: w_t^x là giá trị chiều thứ t của vector x

MÔ HÌNH VECTOR

❖ ĐỘ TƯƠNG ĐỒNG

Ví dụ: tính độ tương đồng giữa các cặp vector (d_1, d_3), (d_1, d_5)

DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
d_1	8	20	2	0
d_3	1	0	42	14
d_5	0	21	9	1

$$\cos(d_1, d_3) = 0.09$$

$$\cos(d_1, d_5) = 0.88$$

TRỌNG SỐ CỦA TERM

❖ KHÁI NIỆM

Trọng số của term:

- Là giá trị của chiều tương ứng trong vector tài liệu và vector truy vấn.
- Ảnh hưởng lớn đến sự tương đồng của những tài liệu.

TRỌNG SỐ CỦA TERM

❖ CÁC DẠNG TRỌNG SỐ

* Trọng số nhị phân: Giá trị mỗi chiều là 0 hoặc 1 tương ứng với có xuất hiện hay không xuất hiện từ khóa tương ứng

Ma trận tài liệu (Doc-Term)

DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
d_1	1	1	1	0
d_2	1	0	1	0
d_3	1	0	1	1
d_4	0	1	0	1
d_5	0	1	1	1

TRỌNG SỐ CỦA TERM

❖ CÁC DẠNG TRỌNG SỐ

* Tần số: Giá trị mỗi chiều của một vector là tần số của term tương ứng trong tài liệu gốc.

Ma trận tài liệu (Doc-Term)

DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
d_1	8	20	2	0
d_2	10	0	31	0
d_3	1	0	42	14
d_4	0	3	0	3
d_5	0	21	9	1

TRỌNG SỐ CỦA TERM

❖ CÁC DẠNG TRỌNG SỐ

Vấn đề của trọng số nhị phân và tần số là không thể hiện được tầm quan trọng của từng từ ngữ

Frequency of words

Upper
cut-off

Lower
cut-off

Resolving power of
significant words

Significant words

Words by rank order

Nghiên cứu của tác giả Rijbergen

TRỌNG SỐ CỦA TERM

❖ CÁC DẠNG TRỌNG SỐ

* Trọng số tf.idf: được tính dựa trên:

- Tần số của term trong tài liệu (tf – Term Frequency)
- Nghịch đảo tần suất tài liệu của term (idf – Inverse Document Frequency)
- Giá trị mỗi chiều trong một vector tài liệu hoặc truy vấn sẽ là giá trị tf.idf của term tương ứng.

TRỌNG SỐ CỦA TERM

❖ CÁC DẠNG TRỌNG SỐ

* Trọng số tf.idf:

Trọng số tf.idf của term thứ k trong tài liệu i được tính như sau:

$$w_{ik} = tf_{ik} * \log(N/n_k)$$

- tf_{ik} : tần số của term ở chiều thứ k tính trong tài liệu i
- N: tổng số tài liệu trong tập tài liệu
- n_k : số tài liệu chứa trong tập tài liệu chứa từ term ở chiều thứ k.

$$idf_k = \log(N/n_k)$$

TRỌNG SỐ CỦA TERM

❖ CÁC DẠNG TRỌNG SỐ

* Trọng số tf.idf:

Giả sử có một tập gồm 1000 tài liệu, tần số tài liệu chứa các term được cho như sau:

Term	truy xuất	thông tin	Công nghệ	Thực phẩm
Số TL	38	200	102	11

$$idf_{truy\ xuất} = \log(1000/38) = 1.42$$

$$idf_{thông\ tin} = \log(1000/200) = 0.7$$

$$idf_{công\ nghệ} = \log(1000/102) = 0.99$$

$$idf_{thực\ phẩm} = \log(1000/11) = 1.96$$

TRỌNG SỐ CỦA TERM

❖ CHUẨN HÓA TRỌNG SỐ

* Chuẩn hóa trọng số tf.idf:

- Nhằm khắc phục sự khác biệt giữa tài liệu dài và tài liệu ngắn.
- Đưa trọng số về miền giá trị [0, 1]
- Chuyển vector tài liệu thành vector cùng phương có độ dài bằng 1

$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^t (tf_{ik})^2 [\log(N / n_k)]^2}}$$

TRỌNG SỐ CỦA TERM

❖ CHUẨN HÓA TRỌNG SỐ

* Trọng số theo hệ thống SMART:

- SMART được Salton xây dựng ở ĐH Cornell
- Trọng số của term ở chiều thứ k của tài liệu d được tính dựa trên 3 thành tố freq_{kd}, collect_k và norm như sau:

$$w_{kd} = \frac{freq_{kd} * collect_k}{norm}$$

TRỌNG SỐ CỦA TERM

❖ CHUẨN HÓA TRỌNG SỐ

* Trọng số theo hệ thống SMART:

$$freq_{kd} = \begin{cases} \{0,1\} & \\ \frac{freq_{kd}}{\max(freq_{kd})} & \\ \frac{1}{2} + \frac{1}{2} \frac{freq_{kd}}{\max(freq_{kd})} & \\ \ln(freq_{kd}) + 1 & \end{cases}$$

Nhị phân

Chuẩn max

Tăng cường

Logarith

TRỌNG SỐ CỦA TERM

❖ CHUẨN HÓA TRỌNG SỐ

* Trọng số theo hệ thống SMART:

$$collect_k = \left\{ \begin{array}{l} \log \frac{NDoc}{Doc_k} \\ \left(\log \frac{NDoc}{Doc_k} \right)^2 \\ \log \frac{NDoc - Doc_k}{Doc_k} \\ \frac{1}{Doc_k} \end{array} \right\}$$

Nghịch đảo

Bình phương

Xác suất

Tần suất

TRỌNG SỐ CỦA TERM

❖ CHUẨN HÓA TRỌNG SỐ

* Trọng số theo hệ thống SMART:

$$norm = \begin{cases} \sum_{vector} w_j \\ \sqrt{\sum_{vector} w_j^2} \\ \sum_{vector} w_j^4 \\ \max_{vector} (w_j) \end{cases}$$

Chuẩn tổng

Chuẩn cosine

Chuẩn Bậc bốn

Chuẩn max

LẬP CHỈ MỤC

❖ VĂN ĐỀ

- Lưu trữ các vector có số chiều rất lớn
- Có nhiều trường hợp là dạng vector thừa.
- Truy xuất hiệu quả.

LẬP CHỈ MỤC

❖ QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

Với mỗi tài liệu, tách các từ khóa để tạo thành danh sách gồm từ khóa và chỉ số tài liệu. Chỉ số tài liệu là thứ tự mà tài liệu đó được xử lý.

mục tài
liệu lập
chỉ mục



Từ khóa	Chỉ số tài liệu
mục	1
tài	1
liệu	1
lập	1
chỉ	1
mục	1

chỉ mục
từ khóa



Từ khóa	Chỉ số tài liệu
chỉ	2
mục	2
từ	2
khóa	2

LẬP CHỈ MỤC

❖ QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

Nối tất cả danh sách và sắp xếp theo từ khóa, chỉ số

Từ khóa	Chỉ số tài liệu
mục	1
tài	1
liệu	1
lập	1
chỉ	1
mục	1
chỉ	2
mục	2
từ	2
khóa	2



Từ khóa	Chỉ số tài liệu
chỉ	1
chỉ	2
khóa	2
lập	1
liệu	1
mục	1
mục	1
mục	2
tài	1
từ	2

LẬP CHỈ MỤC

❖ QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

Gom từ khóa có cùng chỉ số tài liệu và thêm tần số

Từ khóa	Chỉ số tài liệu
chỉ	1
chỉ	2
khóa	2
lập	1
liệu	1
mục	1
mục	1
mục	2
tài	1
từ	2



Từ khóa	Chỉ số tài liệu	Tần số
chỉ	1	1
chỉ	2	1
khóa	2	1
lập	1	1
liệu	1	1
mục	1	2
mục	2	1
tài	1	1
từ	2	1

LẬP CHỈ MỤC

❖ QUÁ TRÌNH LẬP CHỈ MỤC

Từ khóa	Chỉ số tài liệu	Tân số
chỉ	1	1
chỉ	2	1
khóa	2	1
lập	1	1
liệu	1	1
mục	1	2
mục	2	1
tài	1	1
từ	2	1



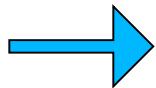
Tập từ vựng

Từ khóa	số tài liệu	Tân số
chỉ	2	2
khóa	1	1
lập	1	1
liệu	1	1
mục	2	3
tài	1	1
từ	1	1

DS Posting

Chỉ số tài liệu	Tân số
1	1
2	1
2	1
1	1
1	1
1	1
2	1
1	1
2	1

LẬP CHỈ MỤC



❖ QUÁ TRÌNH LẬP CHỈ MỤC

Tập từ vựng

Từ khóa	số tài liệu	Tân số	IDF
chỉ	2	2	0.5
khóa	1	1	1
lập	1	1	1
liệu	1	1	1
mục	2	3	0.5
tài	1	1	1
từ	1	1	1

DS Posting

Chỉ số tài liệu	Tân số	w
1	1	0.24
2	1	0.32
2	1	0.63
1	1	0.49
1	1	0.49
1	2	0.49
2	1	0.32
1	1	0.49
2	1	0.63

TRUY VẤN TRÊN MÔ HÌNH VECTOR

❖ XÁC ĐỊNH TÀI LIỆU LIÊN QUAN

* Đối với mỗi term của truy vấn:

- Xác định tần số của nó.
- Xác định vị trí của nó trong từ điển để xác định:
 - + n_k : số tài liệu chứa term này.
 - + tính trọng số w cho term
 - + Xác định vị trí trong danh sách posting.
 - + Xác định danh sách tài liệu từ danh sách posting.

TRUY VẤN TRÊN MÔ HÌNH VECTOR

❖ TÍNH TOÁN ĐỘ TƯƠNG ĐỒNG

* Thực hiện tương tự phép toán OR:

- Đưa lần lượt các tài liệu trong các danh sách tài liệu thu được vào danh sách kết quả theo cách trộn danh sách.
 - + Nếu tài liệu đưa vào chưa có trong danh sách kết quả thì Tổng trọng số WS là vector score của term đang xét giữa truy vấn và tài liệu.
 - + Nếu tài liệu đưa vào đã có thì cộng vector score của term đang xét vào WS.
- Sắp xếp danh sách theo WS.

TRUY XUẤT THÔNG TIN

❖ BÀI TẬP 1

Giả sử có ma trận doc-term với trọng số là tần số như sau:

Ma trận tài liệu (Doc-Term)

DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
d_1	8	20	2	0
d_2	10	0	31	0
d_3	1	0	42	14
d_4	0	3	0	3
d_5	0	21	9	1

TRUY XUẤT THÔNG TIN

❖ BÀI TẬP 1

Yêu cầu:

- 1) Xác định vector đã chuẩn hóa theo trọng số tf.idf của các tài liệu trong bảng.
- 2) Xác định vector của truy vấn q như sau:

Truy xuất thông tin truy xuất

- 3) Cho biết danh sách kết quả của truy vấn q.

TRUY XUẤT THÔNG TIN

❖ BÀI TẬP 2

Cho biết cách xử lý truy vấn như trên có những điểm nào chưa tốt về mặt tính toán và cách khắc phục những điểm này (nếu có).

TRUY XUẤT THÔNG TIN

❖ BÀI TẬP 3

Cho các tài liệu sau

1) ship ocean wood

2) boat ocean

3) ship

4)

wood tree

5) wood

6) tree

Yêu cầu:

- Lập chỉ mục cho tập tài liệu trên
- Xử lý truy vấn theo mô hình vector với trọng số tf.idf cho các truy vấn sau:

1) $Q_1 = \text{wood}$

2) $Q_2 =$



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG III – ĐÁNH GIÁ MÔ HÌNH TRUY XUẤT THÔNG TIN

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ ĐÁNH GIÁ HỆ THỐNG TRUY VĂN THÔNG TIN
- ❖ CÁC TẬP DỮ LIỆU KIỂM TRA CHUẨN
- ❖ ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XÉP HẠNG
- ❖ ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XÉP HẠNG

ĐÁNH GIÁ HỆ THỐNG TRUY VẤN THÔNG TIN

❖ MỤC TIÊU ĐÁNH GIÁ

- Đảm bảo hệ thống có khả năng đáp ứng yêu cầu.
- So sánh với các hệ thống khác.
- Thủ nghiệm và cải tiến thuật toán truy xuất thông tin.

ĐÁNH GIÁ HỆ THỐNG TRUY VẤN THÔNG TIN

❖ TIÊU CHÍ ĐÁNH GIÁ

- Khả năng đáp ứng yêu cầu thông tin.
- Khả năng hiểu của hệ thống.
 - Hiểu nội dung tập tài liệu
 - Hiểu nội dung câu truy vấn
- Sự thân thiện với người sử dụng của hệ thống

ĐÁNH GIÁ HỆ THỐNG TRUY VẤN THÔNG TIN

❖ SỰ LIÊN QUAN GIỮA TÀI LIỆU VÀ TRUY VẤN

- Tài liệu trả về chính xác đối với truy vấn.
- Tài liệu trả về đáp ứng một phần so với yêu cầu của truy vấn.
- Cung cấp nguồn thông tin liên quan để đáp ứng thông tin của truy vấn.
- Cung cấp thông tin cơ bản về vấn đề nêu trong truy vấn.
- Gợi ý một vấn đề khác có liên quan.

ĐÁNH GIÁ HỆ THỐNG TRUY VẤN THÔNG TIN

❖ CHỈ TIÊU ĐÁNH GIÁ

- Độ bao quát thông tin so với yêu cầu.
 - Hình thức trình bày kết quả.
 - Sử dụng phức tạp hay đơn giản.
 - Thời gian truy xuất thông tin
 - Kích thước dữ liệu mà hệ thống sử dụng.
 - Độ phủ (Recall)
 - Độ chính xác (Precision)
- Độ phủ và độ chính xác là 2 chỉ tiêu đánh giá tính hiệu quả của hệ thống

CÁC TẬP DỮ LIỆU KIỂM TRA CHUẨN

❖ CÁC TẬP DỮ LIỆU THỜI KỲ ĐẦU

- Số lượng tài liệu ít.
- Số lượng truy vấn ít.
→ Chỉ thích hợp cho thử nghiệm ban đầu.

Các bộ dữ liệu gồm:

- CACM (3204 tài liệu).
- CISI (1460 tài liệu)
- CRAN (1397 tài liệu)
- INSPEC (12684 tài liệu)
- MED (1033 tài liệu)

CÁC TẬP DỮ LIỆU KIỂM TRA CHUẨN

❖ CÁC TẬP DỮ LIỆU THỜI KỲ ĐẦU

- Truy vấn và đánh giá độ liên quan:
 - Thực hiện thủ công bằng cách đưa ra truy vấn theo ý định của các tác giả
 - Các tài liệu liên quan tương ứng được xác định trên toàn bộ tập tài liệu và được lập thành danh sách.

CÁC TẬP DỮ LIỆU KIỂM TRA CHUẨN

❖ TẬP DỮ LIỆU CỦA TREC

- Số lượng hơn 1,5 triệu tài liệu, chiếm dung lượng khoảng 5GB với các loại (2013): tin tức, tài liệu của chính phủ, bằng phát minh sang chép ở Hoa Kỳ, ...
- Truy vấn và đánh giá độ liên quan:
 - Do chuyên viên thông tin đưa ra và đánh giá.
 - Đánh giá độ liên quan dựa trên tài liệu trả về, không đánh giá trên toàn bộ tập tài liệu
- Chỉ tiêu đánh giá: Độ phủ (Recall) và độ chính xác (Precision)

CÁC TẬP DỮ LIỆU KIỂM TRA CHUẨN

❖ TẬP DỮ LIỆU CỦA TREC

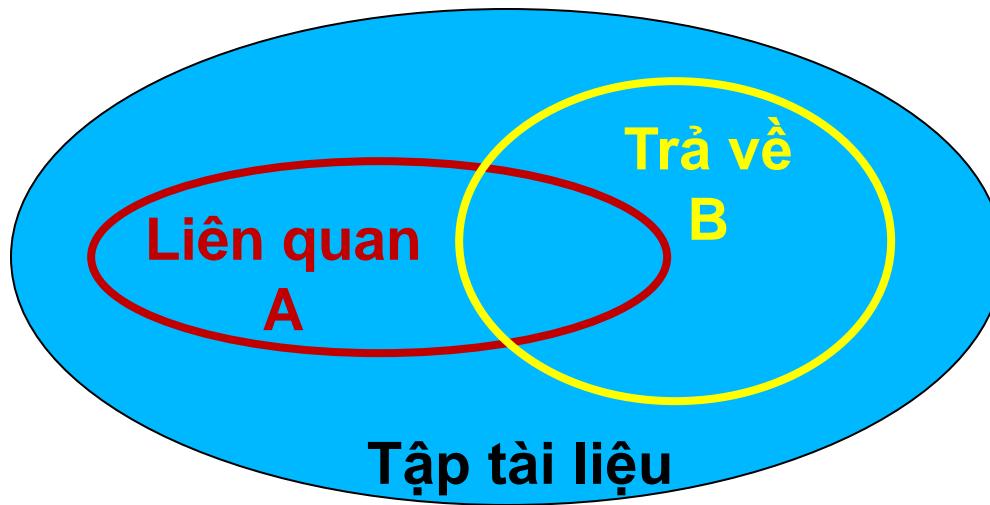
- Ngoài truy xuất thông tin dạng tài liệu, TREC còn tổ chức đánh giá truy xuất thông tin với các dạng:
 - Hệ thống hỏi đáp (QA System)
 - Tìm kiếm thông tin trên Web (Web Searching)
 - Tìm theo phim, âm thanh, hình ảnh.
 - Tìm xuyên ngôn ngữ (Cross-lingual).

ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

Độ phủ cho 1 truy vấn: $R = |A \cap B| / |A|$

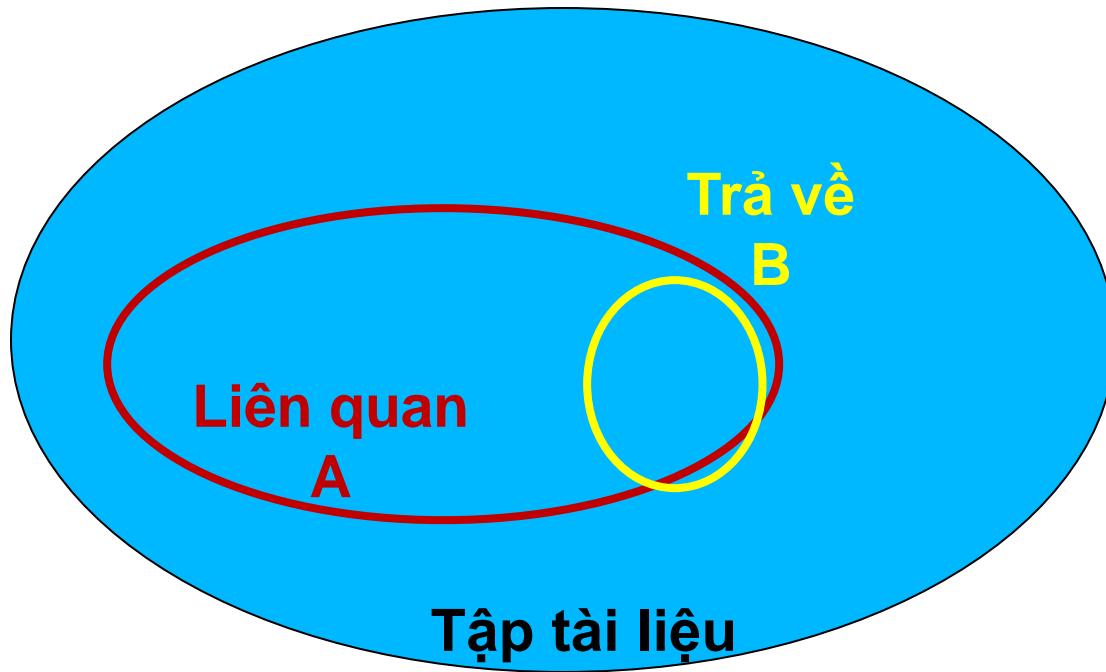
Độ chính xác cho 1 truy vấn: $P = |A \cap B| / |B|$



ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

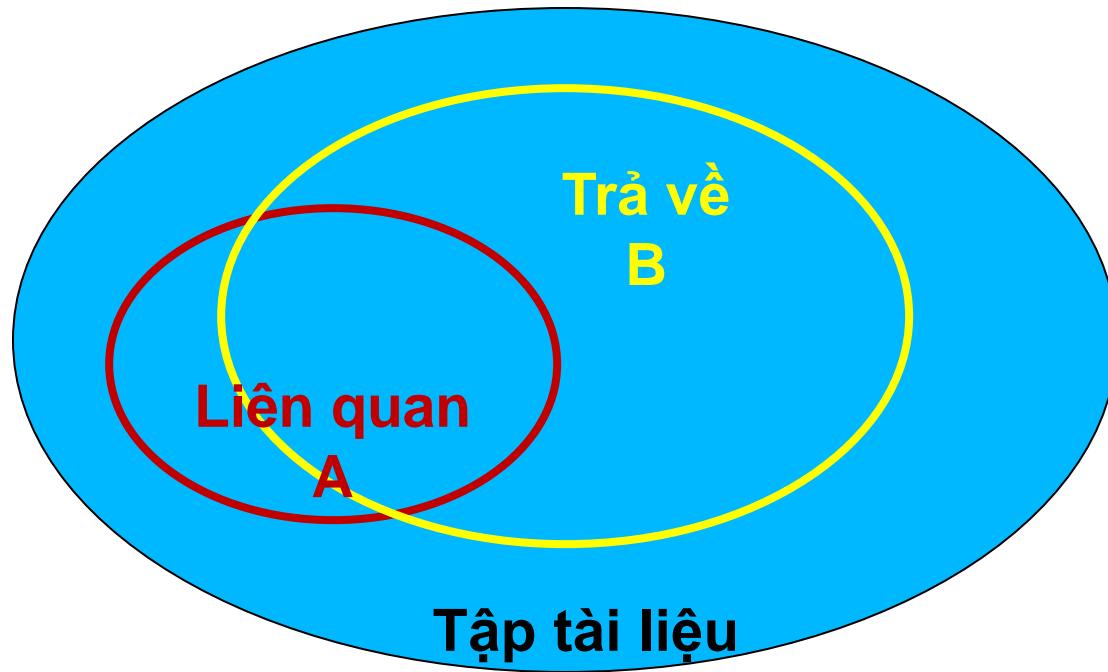
Trường hợp độ chính xác cao, độ phủ thấp



ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

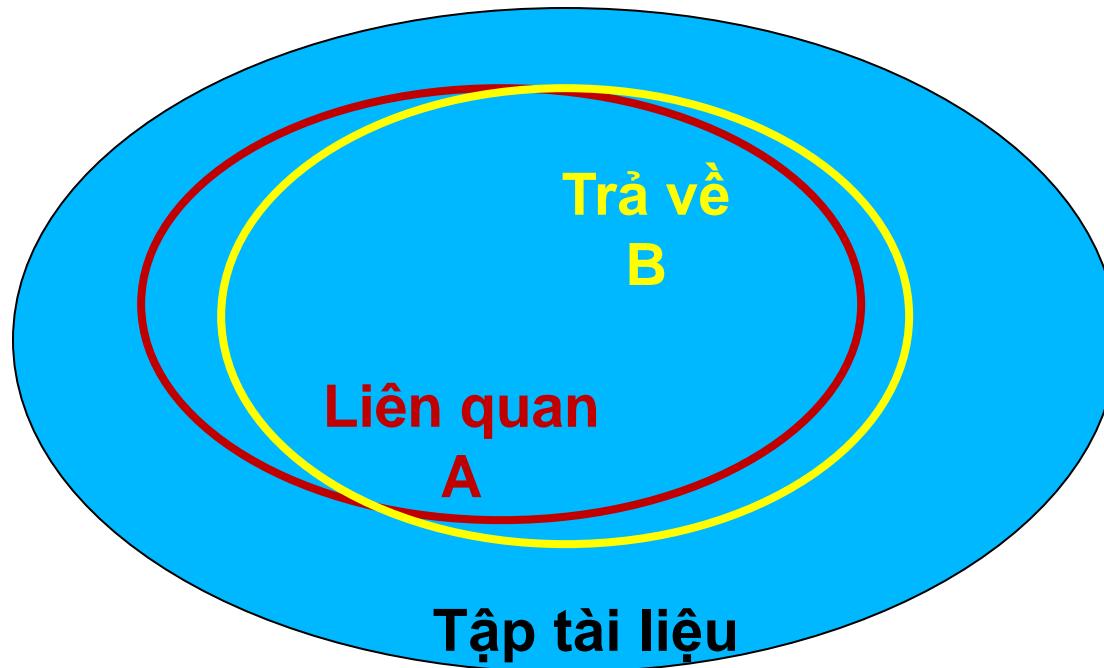
Trường hợp độ chính xác thấp, độ phủ cao



ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

Trường hợp độ chính xác cao, độ phủ cao



ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

Độ phủ và độ chính xác trung bình tương ứng với tập dữ liệu thử nghiệm được tính là trung bình cộng của độ phủ và độ chính xác tương ứng với từng truy vấn

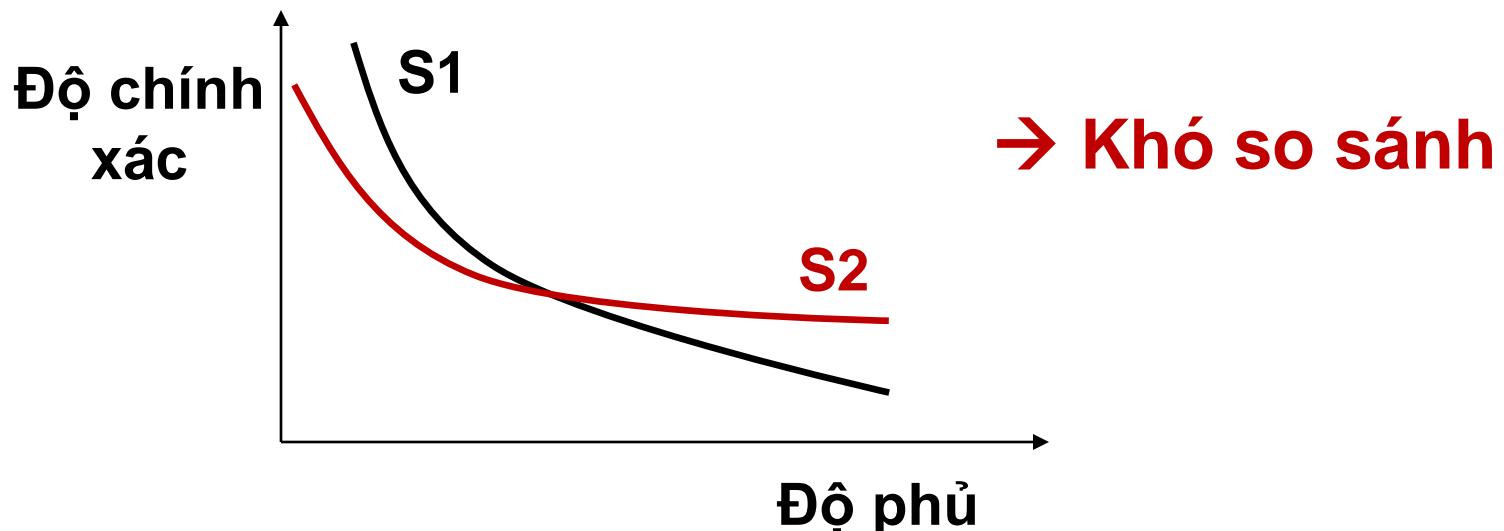
$$\bar{P} = \frac{1}{|Q|} \sum_q^{q \in Q} P_q$$

$$\bar{R} = \frac{1}{|Q|} \sum_q^{q \in Q} R_q$$

ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

- Có sự đánh đổi giữa độ phủ và độ chính xác
- Đồ thị thể hiện độ phủ và độ chính xác tính trung bình của 2 hệ thống IR S1 và S2 trên nhiều truy vấn có dạng:



ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ ĐO F (F-MEASURE)

Kết hợp độ phủ và độ chính xác tương ứng với 1 truy vấn thành một số tương tự như trung bình

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Với $\beta^2 = \frac{1-\alpha}{\alpha}$

ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ ĐỘ ĐO F (F-MEASURE)

Trường hợp cân bằng giữa độ phủ và độ chính xác ($\beta=1$)

$$F = \frac{2PR}{P + R}$$

Độ đo F trung bình của bộ dữ liệu thử nghiệm bằng trung bình cộng của độ đo F tương ứng với từng truy vấn

$$\bar{F} = \frac{1}{|Q|} \sum_q^{q \in Q} F_q$$

ĐÁNH GIÁ CÁC KẾT QUẢ CHƯA XẾP HẠNG

❖ VÍ DỤ

Cho tập tài liệu liên quan đến truy vấn là:

$$R = \{3, 5, 9, 25, 39, 44, 56, 71, 89, 123\}$$

Cho kết quả truy vấn gồm tập các tài liệu được xếp hạng như sau:

$$\{123, 84, \mathbf{56}, 6, 8, \mathbf{9}, 511, 129, 187, \mathbf{25}, 38, 48, 250, 113, 3\}$$

Độ phủ: $R = 5/10 = 0.5$

Độ chính xác $P = 5/15 = 0.33$

Độ F_1 $F_1 = 2PR / (P + R) = 0.398$

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ NHƯỢC ĐIỂM CỦA ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

- Độ phủ và độ chính xác có quan hệ với nhau
- Đối với tập tài liệu lớn thì không thể xác định độ phủ.

→ Cần phương pháp đánh giá trong trường hợp tập tài liệu lớn.

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ PHƯƠNG PHÁP CHIA CẮT

- Áp dụng đối với tập tài liệu lớn, không thể xác định tập tài liệu liên quan trên toàn bộ tập tài liệu.
- Phương pháp thực hiện:
 - Kết quả trả về của 1 truy vấn từ hệ thống phải được xếp hạng.
 - Chọn các tập kết quả là n tài liệu đầu tiên ($n = 5, 10, 20, 50, 100, 500$)
 - Tính độ chính xác tại mỗi tập kết quả $P@n$
 - Tính độ chính xác trung bình (có thể có trọng số) từ độ chính xác của mỗi tập kết quả.

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ VÍ DỤ:

Cho kết quả truy vấn gồm tập các tài liệu được xếp hạng như sau: (tài liệu tô đậm là có liên quan)

{123, 84, **56**, 6, 8, **9**, 511, 129, 187, **25**, 38, 48, 250, 113, **3**}

Độ chính xác tại mức cắt 5: $P@5 = 2/5 = 0.4$

Độ chính xác tại mức cắt 10: $P@10 = 4/10 = 0.4$

Độ chính xác trung bình: $P = 0.4*0.5+0.4*0.5=0.4$

(giả sử chọn trọng số của mỗi mức cắt là nhau)

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ PHƯƠNG PHÁP TƯƠNG QUAN GIỮA ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

- Áp dụng khi có thể xác định tập tài liệu liên quan trên toàn bộ tập tài liệu.
- Phương pháp thực hiện:
 - Kết quả trả về của 1 truy vấn từ hệ thống phải được xếp hạng.
 - Tại mỗi thứ hạng có tài liệu liên quan trong kết quả, tính độ chính xác và độ phủ.
 - Dựa trên những điểm tính được độ phủ và độ chính xác để nội suy những điểm cần so sánh.

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ PHƯƠNG PHÁP TƯƠNG QUAN GIỮA ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

- Nguyên tắc nội suy như sau:
 - Giả sử có các điểm tính được độ phủ là $i = 1, 2, \dots$, giá trị độ phủ và độ chính xác tương ứng là r_i và p_i
 - Khi đó, độ chính xác p tương ứng với độ phủ r cần nội suy được xác định:

$$p = \max(p_{r'}) \text{ với } r \leq r'$$

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ PHƯƠNG PHÁP TƯƠNG QUAN GIỮA ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

- Mục đích của phương pháp nhằm so sánh độ chính xác của các hệ thống tại một độ phủ r xác định.

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ VÍ DỤ

Cho tập tài liệu liên quan đến truy vấn là:

$$R = \{3, 5, 9, 25, 39, 44, 56, 71, 89, 123\}$$

Cho kết quả truy vấn gồm tập các tài liệu được xếp hạng như sau: **{123, 84, 56, 6, 8, 9, 511, 129, 187, 25, 38, 48, 250, 113, 3}**

Yêu cầu: Tính p tại điểm có $r = 0.15$.

- Tài liệu liên quan đầu tiên ở vị trí 1:

$$p_1 = 1/1 = 1.0; \quad r_1 = 1/10 = 0.1$$

- Tài liệu liên quan thứ hai ở vị trí 3:

$$p_2 = 2/3 = 0.66; \quad r_2 = 2/10 = 0.2$$

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ VÍ DỤ

- Tài liệu liên quan thứ 3 ở vị trí 6:

$$p_3 = 3/6 = 0.5; \quad r_3 = 3/10 = 0.3$$

- Tài liệu liên quan thứ 4 ở vị trí 10:

$$p_4 = 4/10 = 0.4; \quad r_4 = 4/10 = 0.4$$

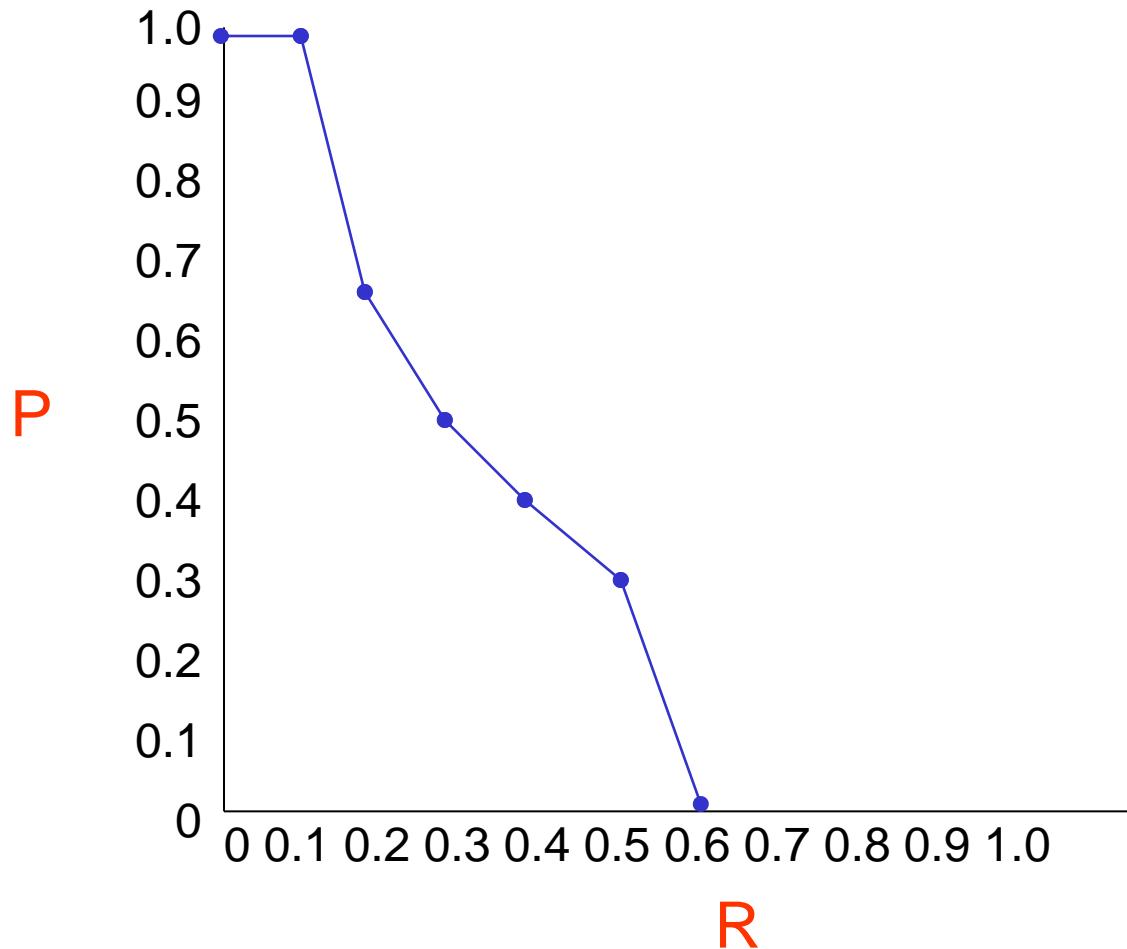
- Tài liệu liên quan thứ 5 ở vị trí 15:

$$p_5 = 5/15 = 0.33; \quad r_5 = 5/10 = 0.5$$

- Với $r = 0.15$, $p = p_2 = 0.66$

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ VÍ DỤ



ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ PHƯƠNG PHÁP TƯƠNG QUAN GIỮA ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

Độ chính xác trung bình không nội suy

- Average Precision
- Tính theo từng truy vấn
- Bằng độ chính xác trung bình tại các độ phủ khác nhau của truy vấn:

$$AP = 1/|r| * \sum_r p_r$$

ĐÁNH GIÁ CÁC KẾT QUẢ ĐÃ XẾP HẠNG

❖ PHƯƠNG PHÁP TƯƠNG QUAN GIỮA ĐỘ PHỦ VÀ ĐỘ CHÍNH XÁC

Độ chính xác trung bình kỳ vọng

- Mean Average Precision
- Tính trên tất cả truy vấn
- Bằng giá trị trung bình của độ chính xác trung bình của từng truy vấn:

$$\text{MAP} = 1/|Q| * \sum_Q p_q$$



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG IV – XÂY DỰNG SEARCH ENGINE

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ GIỚI THIỆU
- ❖ WEB CRAWLER
- ❖ TÁCH NỘI DUNG
- ❖ LẬP CHỈ MỤC VÀ TRUY XUẤT

GIỚI THIỆU

❖ SEARCH ENGINE

- Search Engine cho nội dung văn bản.
- Có chức năng như một thư viện số.
- Không cần phân loại tài liệu trước khi tìm kiếm.
- Tìm theo từ khóa do người sử dụng chọn.

GIỚI THIỆU

❖ CÁC DẠNG SEARCH ENGINE

- **Personal computer search:**
 - Tập tài liệu trên một máy tính cá nhân (file).
 - Khối lượng dữ liệu tương đối nhỏ.
- **Domain-specific search**
 - Tập tài liệu trên hệ thống mạng của một tổ chức (web page, file)
 - Nội dung xác định trước (có thể xử lý ngữ nghĩa)
 - Khối lượng dữ liệu không quá lớn

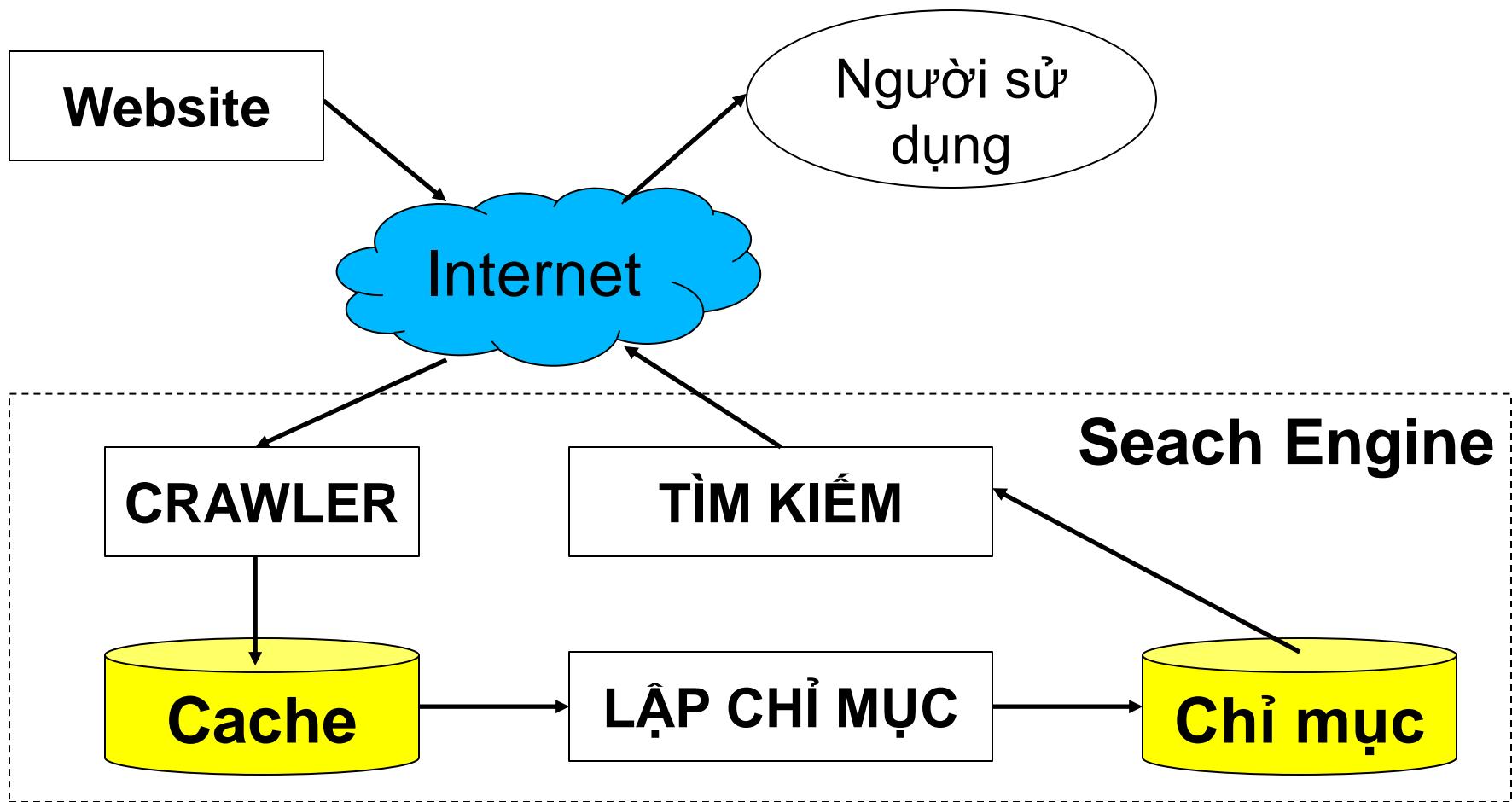
GIỚI THIỆU

❖ CÁC DẠNG SEARCH ENGINE

- **Web search: tương tự Domain-specific search**
 - Tài liệu trên các website (web page, file)
 - Nội dung đa dạng (khó đảm bảo ngũ nghĩa)
 - Khối lượng tài liệu rất lớn (vấn đề hiệu quả trong việc lập chỉ mục và tìm kiếm)

GIỚI THIỆU

❖ MÔ HÌNH CỦA WEB SEARCH ENGINE



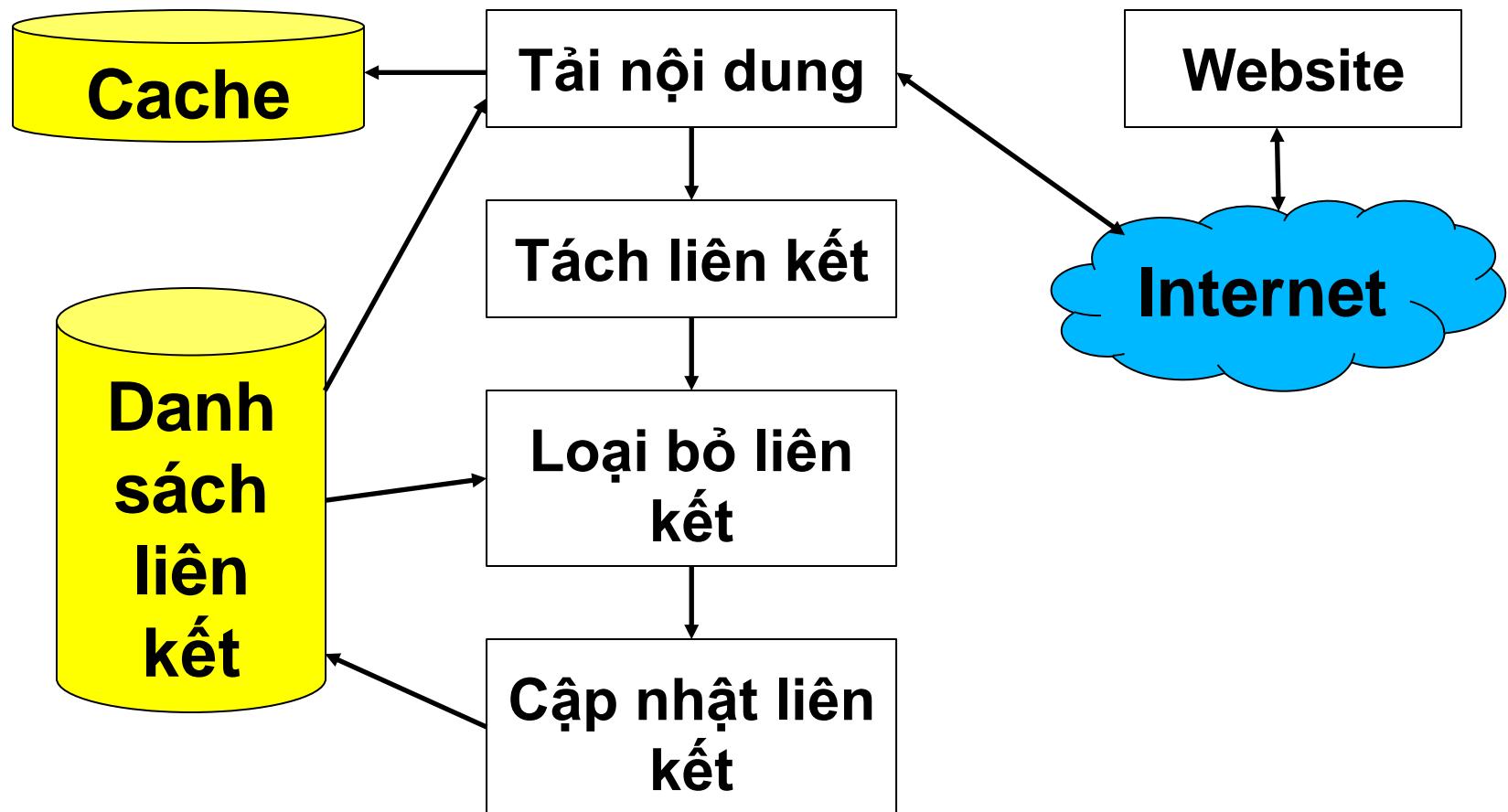
WEB CRAWLER

❖ GIỚI THIỆU

- Còn gọi là web spider hoặc bot.
- Duyệt các trang web trên internet để thu thập nội dung.
- Tự động duyệt theo các liên kết (hyper-link)
- Cần có một trang web ban đầu, gọi là seed để duyệt đến những trang khác. Thông thường seed được chọn là sitemap

WEB CRAWLER

❖ MÔ HÌNH CỦA CRAWLER



WEB CRAWLER

❖ MỘT SỐ VẤN ĐỀ CỦA WEB CRAWLER

- Trùng trang web.
- Bẫy crawler (crawler trap – không thể kết thúc duyệt).
- Có thể bị xác định là DDoS (Distributed Denial of Service) → kiểm tra file robots.txt của website (Robots Exclusion Protocol).
- Website sử dụng công nghệ AJAX.

WEB CRAWLER

❖ MỘT SỐ CRAWLER

- Crawler4j.
- Wget.
- Apache Nutch.

TÁCH NỘI DUNG

❖ MỤC ĐÍCH

- Lấy nội dung văn bản của các file (doc, xls, pdf, ...)
- Lấy nội dung chính của trang web. Loại bỏ:
 - Quảng cáo
 - Web Navigator (hệ thống menu của website)
 - Những liên kết chứa nội dung không liên quan đến nội dung chính của trang web.

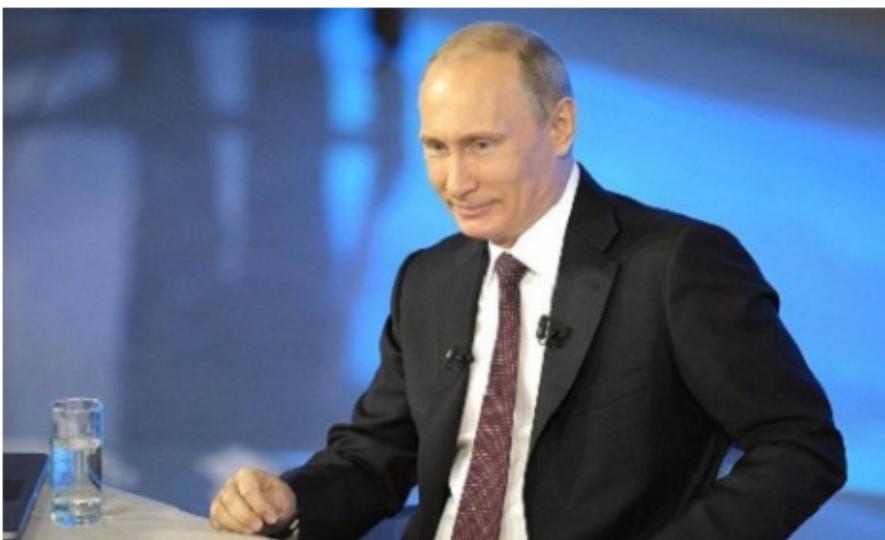
Thứ năm, 6/4/2017 | 21:35 GMT+7



Putin tiết lộ thói quen ăn uống

Tổng thống Nga Vladimir Putin tiết lộ với báo giới sở thích và thói quen trong cuộc sống hàng ngày của ông.

• Putin trích dẫn nhằm lời cựu tổng thống Mỹ



Tổng thống Nga Vladimir Putin. Ảnh: *Sputnik*.

Trong buổi họp báo tại Diễn đàn Truyền thông ONF ngày 5/4, Tổng thống Nga Vladimir Putin tiết lộ món ăn ưa thích nhất của ông là cháo Nga truyền thống. Ông có thể ăn bất cứ loại cháo nào, trừ ngũ cốc, theo *Sputnik*.



In Hanoi, train tracks are just your front yard
Tài trợ



SAIGON SOUTH RESIDENCES

Tháng 4.2017

Cơ hội cuối cùng

Tòa nhà hướng sông
ngay trung tâm dự án

Xem thêm »



088 801 1977 - 098 1357 444 - 0936 090 566

AD BY ECLICK

Tư liệu

Xung đột của ông Trump và ông Tập



Toàn cầu hóa, EU, Trung Đông và Triều Tiên là những vấn đề toàn cầu được dự báo tạo ra đối thoại lớn giữa ông Trump và ông Tập trong ...

• Người dàn xếp cuộc gặp đầu tiên Trump - Tập

TÁCH NỘI DUNG

Dân trí > Văn hóa >

Thứ Năm, 06/04/2017 - 09:48

Vợ nhạc sĩ Châu Kỳ bất ngờ lên tiếng về việc "Con đường xưa em đi" bị cấm

Chia sẻ

Thích <232

G+1



Gửi

Liên quan đến việc 5 ca khúc sáng tác trước năm 1975, trong đó có "Con đường xưa em đi" của nhạc sĩ Châu Kỳ bị Cục NTBD cấm lưu hành vô thời hạn do cho rằng đã sửa lời và có nội dung không đúng với bản gốc, vợ của nhạc sĩ Châu Kỳ đã lần đầu tiên lên tiếng chia sẻ về việc này.

- >> Sao không công bố bản gốc "Con đường xưa em đi"?
- >> Sửa lời khác bản gốc, "Con đường xưa em đi" bị cấm vĩnh viễn
- >> "Con đường xưa em đi" bị tạm dừng lưu hành vì ca từ không đúng?

Theo đó, bà Kha Thị Đặng là vợ của nhạc sĩ Châu Kỳ hiện đang sinh sống cùng gia đình ở TP.HCM. Bà Đặng cho biết, từ khi ca khúc "Con đường xưa em đi" bị cơ quan quản lý văn hóa tạm dừng lưu hành cho đến khi có thông tin bài hát này sẽ bị cấm vĩnh viễn, bà và mọi người trong nhà đều biết nhưng không hiểu rõ lý do của việc này.

Quảng cáo bởi Admicro

GÓC TƯ VẤN BỆNH VIÊM THANH QUẢN



Hết khốn khổ vì đau họng, khản giọng mất tiếng

Trở trời là họng lại sưng đau, mất tiếng khiến cô không ăn uống được

- Khản tiếng có khó thuyên giảm như bạn nghĩ ?
- Bí quyết cải thiện bệnh viêm thanh quản hành hạ suốt 8 năm

[khantieng.vn](#) tài trợ thông tin

HỖ TRỢ ĐIỀU TRỊ BỆNH VÂY NẾN



Tránh được bệnh vây nến sau 15 năm nhờ bí quyết

TÁCH NỘI DUNG

❖ PHƯƠNG PHÁP

- Phân tích cấu trúc tài liệu HTML thành dạng cây. Công cụ phân tích là các HTML Parser, chẳng hạn jSoup.
- Dựa vào quy ước hoặc đặc điểm trình bày trang web để xác định node chứa nội dung chính.
- Lấy phần text của node chứa nội dung chính.

TÁCH NỘI DUNG

❖ PHƯƠNG PHÁP

- **Dựa vào quy ước:** Các thẻ **<DIV>**, **<P>**, **<TABLE>**, ... có thuộc tính (thường là **id**) đánh dấu nội dung chính.

```
496         </div>
497     </div>
498 <div class="clearfix adm-mainsection">
499     <div class="ads-sponsor type-2 adm-hidden"><div id="admsection2"></div></div>
500     <div class="ads-sponsor type-2 adm-hidden"><div id="admsection3"></div></div>
501     <div class="clearfix wrapper">
502         <div class="container">
503             <div class="fl wid470 adm-leftsection">
504                 <div id="ctl00_IDContent_Tin_Chi_Tiet">
505                     <div id="ctl00_IDContent_ctl00_divContent" class="clearfix">
506
507             <div class="box26 clearfix">
508
509             <!--VuLV - edit 11/06-->
510
511                 <ol class="breadcrumb clearfix inline" itemscope itemtype="http://schema.org/BreadcrumbList">
512                     <li class="fl" itemprop="itemListElement" itemscope itemtype="http://schema.org/ListItem">
513                         <a class="breadcrumbitem1" itemprop="item" href="/">
514                             <span itemprop="name">DÂN TRÍ </span> <span>Danh mục </span> <span>Danh mục </span> <span>Danh mục </span>
```

TÁCH NỘI DUNG

❖ PHƯƠNG PHÁP

- **Dựa vào quy ước:**
 - **Ưu điểm:**
 - ❖ Phương pháp đơn giản.
 - ❖ Dễ thực hiện.
 - **Nhược điểm:**
 - ❖ Website khác nhau có quy ước khác nhau.
 - ❖ Một website cũng thay đổi theo thời gian.

TÁCH NỘI DUNG

❖ PHƯƠNG PHÁP

- Dựa vào hình thức trình bày vùng nội dung chính của trang web:
 - Chiều ngang lớn nhất.
 - Tỉ lệ liên kết so với văn bản thấp.
 - Cấu trúc không lặp với phần khác trong trang web.
 -

TÁCH NỘI DUNG

❖ PHƯƠNG PHÁP

- Dựa vào hình thức trình bày vùng nội dung chính của trang web:
 - Ưu điểm.
 - ❖ Có thể áp dụng cho nhiều website khác nhau.
 - ❖ Có thể áp dụng khi một website thay đổi không đáng kể.
 - Nhược điểm:
 - ❖ Phức tạp
 - ❖ Không đảm bảo việc áp dụng được cho tất cả Website

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖ TIỀN XỬ LÝ

- Chuyển bộ mã.
- Loại bỏ các ký tự không sử dụng.
- Sửa lỗi chính tả.

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖ PHÂN TÍCH TERM

- Tách từ.
- Xử lý ngũ nghĩa:
 - Stemming hoặc Lemmatizing (Vd: tiếng Anh).
 - Xác định ranh giới từ (Vd: tiếng Việt, tiếng Hoa)
 - Xác định khái niệm hoặc thực thể.
 - Xác định ý nghĩa của câu (cần nhiều nghiên cứu)
- Kết quả phân tích term ảnh hưởng đến độ phủ của hệ thống.

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖ LẬP CHỈ MỤC VÀ TÌM KIẾM

- Chọn mô hình: Boolean, Vector Space, Xác suất,..
- Chọn phương pháp tính trọng số cho term.
- Chọn công thức tính độ tương đồng giữa truy vấn và tài liệu.

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖ MỘT SỐ THƯ VIỆN HỖ TRỢ LẬP CHỈ MỤC VÀ TÌM KIẾM

- Apache Lucene
- Apache Solr



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG V – MỘT SỐ MÔ HÌNH TRUY XUẤT THÔNG TIN KHÁC

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR
- ❖ MÔ HÌNH LSI
- ❖ MÔ HÌNH XÁC SUẤT

NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR

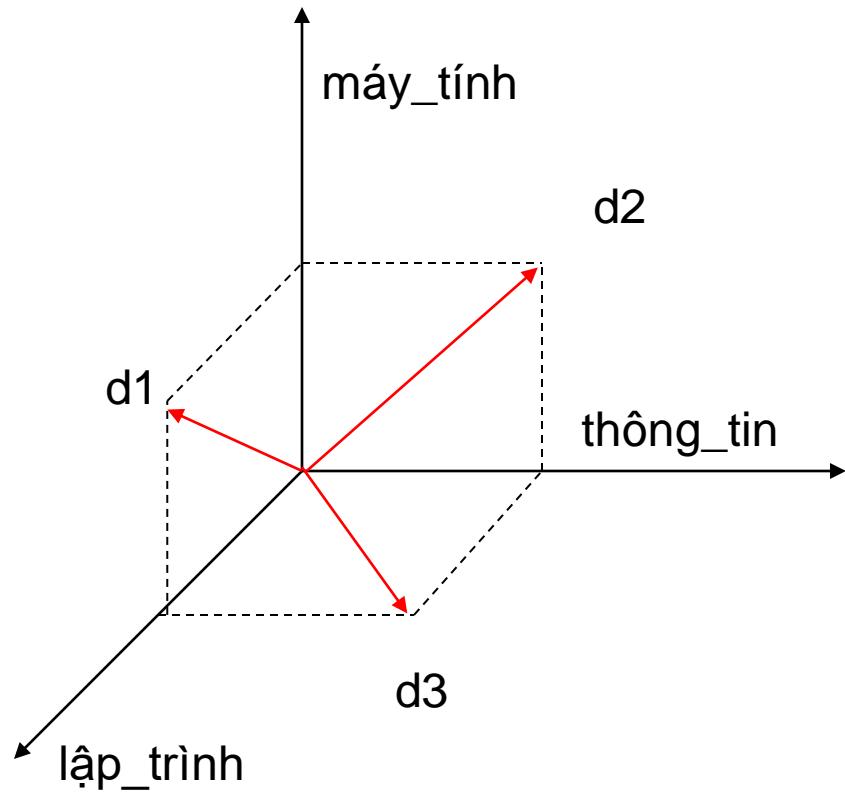
❖ NHƯỢC ĐIỂM TRONG BIỂU DIỄN

Giả sử có các tài liệu:

d1: máy_tính lập_trình

d2: máy_tính thông_tin

d3: lập_trình thông tin



NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR

❖ NHƯỢC ĐIỂM TRONG BIỂU DIỄN

- Các chiều trong không gian là mỗi từ khóa, không đảm bảo độc lập về ngữ nghĩa.
- Các vector chỉ nằm trong phần dương của không gian.
- Số lượng chiều rất lớn, trong đó có những từ khóa có thể không cần thiết phải biểu diễn (những từ nhiễu)

NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR

❖ NHƯỢC ĐIỂM TRONG SO KHỚP

- So khớp dựa vào từ khóa, nếu tài liệu và truy vấn không có từ khóa chung thì độ tương đồng bằng 0
- Chưa có cơ chế so khớp những từ có nghĩa gần nhau, chẳng hạn: “máy_tính” và “lập_trình”

MÔ HÌNH LSI

❖ MỤC TIÊU CỦA MÔ HÌNH LSI

Mô hình LSI (Latent Semantic Index) được đề xuất nhằm:

- Đảm bảo các chiều trong không gian là độc lập.
- Các chiều được chọn mang ý nghĩa của những từ khoá dựa trên sự xuất hiện đồng thời của chúng. Không nhất thiết phải là tập từ khóa sử dụng trong các tài liệu (ngữ nghĩa tiềm ẩn).
- Có thể giảm số chiều trong không gian mà cho kết quả xấp xỉ.

MÔ HÌNH LSI

❖ CƠ SỞ TOÁN

Cho ma trận A, vector x được gọi là vector riêng của A nếu tồn tại một số λ , gọi là trị riêng sao cho:

$$Ax = \lambda x$$

- x là vector riêng của A thì x không thay đổi phương khi nhân với A.
- Giả sử x_1, x_2, \dots, x_n là vector riêng ứng với các trị riêng λ_i khác nhau của A, khi đó, x_i và x_j ($i \neq j$) độc lập tuyến tính.

MÔ HÌNH LSI

❖ CƠ SỞ TOÁN

Cho ma trận A, giả sử $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_n^2$ là các trị riêng của $A^T A$ tương ứng với các vector riêng x_1, x_2, \dots, x_n .

Đặt: $U = \{x_1, x_2, \dots, x_n\}$

$$y_i = (1/\sigma_i)A x_i \quad (i = 0..n)$$

$$S = \{y_1, y_2, \dots, y_n\}$$

Σ là ma trận đường chéo trong đó các giá trị trên đường chéo là $\sigma_1, \sigma_2, \dots, \sigma_n$

Khi đó
$$A = S \Sigma U^T$$

Được gọi là một phép tách ma trận SVD (Singular Value Decomposition)

MÔ HÌNH LSI

❖ CƠ SỞ TOÁN

Trong trường hợp chỉ chọn r giá trị riêng đầu tiên
($r < n$)

Đặt: $U_r = \{x_1, x_2, \dots, x_r\}$

$$y_i = (1/\sigma_i) A x_i \quad (i = 0..r)$$

$S_r = \{y_1, y_2, \dots, y_r\}$

Σ là matrận đường chéo trong đó các giá trị trên
đường chéo là $\sigma_n, \sigma_n, \dots, \sigma_r$

Khi đó

$$A \approx S_r \Sigma U_r^T$$

MÔ HÌNH LSI

❖ LATENT SEMANTIC INDEX

Áp dụng phép tách ma trận SVD cho ma trận A là ma trận Term-Document:

$$A = S \Sigma U^T$$

Khi đó, các vector từ khóa K và các vector tài liệu D sẽ được biểu diễn trong cùng không gian với các chiều là các vector riêng qua phép biến đổi:

$$K = S \Sigma$$

$$D = \Sigma U^T$$

MÔ HÌNH LSI

❖ LATENT SEMANTIC INDEX

Các vector riêng được xem là những nghĩa tiềm ẩn trong mối liên hệ cùng xuất hiện của các từ khóa

Trong trường hợp muốn giảm số chiều, sẽ chọn r trị riêng và vector riêng đầu tiên, khi đó

$$K = S_r \Sigma_r$$

$$D = \Sigma_r U_r^T$$

MÔ HÌNH LSI

❖ LATENT SEMANTIC INDEX

Ví dụ:

	d_1	d_2	d_3
máy_tính	1	1	0
lập_trình	1	0	1
thông_tin	0	1	1

$$S = \begin{pmatrix} 0.57 & -0.4 & 0.71 \\ 0.57 & -0.4 & -0.71 \\ 0.57 & 0.82 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0.57 & -0.81 & 0 \\ 0.57 & 0.41 & 0.71 \\ 0.57 & 0.41 & -0.71 \end{pmatrix}$$

MÔ HÌNH LSI

❖ LATENT SEMANTIC INDEX

$$\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Công cụ tách ma trận theo SVD trực tuyến:
<http://users.telenet.be/paul.larmuseau/SVD.htm>

MÔ HÌNH LSI

❖ TÍNH ĐỘ TƯƠNG ĐỒNG GIỮA TÀI LIỆU VÀ TRUY VẤN

- Chuyển truy vấn thành vector dựa trên các vector từ khóa.
- Tính độ đo cosine (hoặc độ đo nào khác tùy chọn) dựa trên vector truy vấn và vector tài liệu.

MÔ HÌNH LSI

❖ TÍNH ĐỘ TƯƠNG ĐỒNG GIỮA TÀI LIỆU VÀ TRUY VẤN

Ví dụ:

d₁ Romeo and Juliet

d₂ Juliet: Oh happy dagger

d₃ Romeo died by dagger

d₄ "live free or die" is from New-Hampshire

d₅ he is from New-Hampshire

Xếp hạng tài liệu theo truy vấn q: "die dagger"



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG V – MỘT SỐ MÔ HÌNH TRUY XUẤT THÔNG TIN KHÁC

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR
- ❖ MÔ HÌNH LSI
- ❖ MÔ HÌNH XÁC SUẤT

MÔ HÌNH XÁC SUẤT

❖ CÁCH TIẾP CẬN

- Truy vấn thông tin là một quá trình không chắc chắn
 - Ngữ nghĩa câu truy vấn
 - Tài liệu thỏa truy vấn.
- Lý thuyết xác suất
 - Cở sở của suy luận không chắc chắn
 - Ước lượng khả năng tài liệu liên quan đến truy vấn

MÔ HÌNH XÁC SUẤT

❖ CÁCH TIẾP CẬN

- Các mô hình xác suất
 - Các mô hình điển hình: BIM, Two Poisson, BM11, BM25
 - Mô hình ngôn ngữ
 - Mạng Bayes (Bayesian networks)

Các mô hình xác suất là những mô hình đã cũ nhưng có hiệu quả cao.

MÔ HÌNH XÁC SUẤT

❖ NGUYÊN LÝ XẾP HẠNG

- Bài toán: cho một truy vấn q và một tập tài liệu D , xác định thứ tự các tài liệu trong D theo truy vấn q .
- Sự liên quan giữa tài liệu và truy vấn được thể hiện bằng một biến ngẫu nhiên nhị phân R
 - $R_{d,q}=1$ nếu tài liệu d liên quan với q
 - $R_{d,q}=0$ nếu tài liệu d không liên quan với q
- Kết quả xếp hạng được sắp xếp giảm dần theo xác suất của sự liên quan $p(R_{d,q}=1)$ hay $p(R=1|d,q)$

MÔ HÌNH XÁC SUẤT

❖ NGUYÊN LÝ XẾP HẠNG

- Các mô hình vẫn thông tin theo xác suất cần giải quyết:
 - Xác định các giá trị ước lượng tốt nhất.
 - Phương pháp tính toán xác suất liên quan giữa tài liệu d và truy vấn q

MÔ HÌNH XÁC SUẤT

❖ PHÁT BIỂU BÀI TOÁN

- Tài liệu $d = \{td_1, td_2, \dots, td_m\}$
 - td_i là term thứ i của tài liệu d
 - $p(td_i = 'tù')$ là xác suất term thứ i của tài liệu d có giá trị là ' $tù$ '
- Truy vấn $q = \{tq_1, tq_2, \dots, tq_n\}$
 - tq_i là term thứ i của truy vấn q
 - $p(tq_i = 'tù')$ là xác suất term thứ i của truy vấn q có giá trị là ' $tù$ '
- Sự liên quan $R \in \{0,1\}$

MÔ HÌNH XÁC SUẤT

❖ PHÁT BIỂU BÀI TOÁN

- Xác suất tài liệu d có liên quan với truy vấn q là
 $p(R=1|d, q)$

MÔ HÌNH XÁC SUẤT

❖ MỘT SỐ CÔNG THỨC XÁC SUẤT

- Xác suất của hai sự kiện A, B xảy ra đồng thời là $p(A,B)$. Nếu A và B độc lập thì
$$p(A,B) = p(A) * p(B)$$
- Xác suất có điều kiện $P(A|B)$ là xác suất sự kiện A nếu trước đó có sự kiện B xảy ra.
- $p(A,B,C) = p(A) * p(B|A) * p(C|A, B)$
- $p(A) = p(A,B) + p(A,\neg B)$
- $p(A) = p(A,B=b_1) + p(A,B=b_2) + \dots + p(A,B=b_m)$

MÔ HÌNH XÁC SUẤT

❖ MỘT SỐ CÔNG THỨC XÁC SUẤT

- Công thức Bayes:

$$p(A|B) = p(A) * p(B|A) / p(B)$$

$$p(A|B) = p(A) * p(B|A) / [\sum_{X \in \{A, \neg A\}} p(B|X) * p(X)]$$

- Tỉ lệ Odds:

$$O(A) = p(A) / p(\neg A) = p(A) / (1-p(A))$$

- log-odds:

$$\log(O(A)) = \log(p(A)) - \log(1-p(A))$$

MÔ HÌNH XÁC SUẤT

❖ XẾP HẠNG THEO MÔ HÌNH XÁC SUẤT

- Giả thiết: sự liên quan của các tài liệu với một câu truy vấn là độc lập.
- Ước lượng xác suất của sự liên quan $p(R=1|d, q)$ theo:
 - Tần suất của term
 - Tần suất của tài liệu
 - Độ dài của văn bản
- Đặt r là $R=1$ và $\neg r$ là $R=0$: các tài liệu được xếp hạng theo thứ tự giá trị $p(r|d, q)$ giảm dần.
- Thay vì xếp hạng theo giá trị $p(r|d, q)$, có thể xếp hạng theo giá trị $\text{odds}(p(r|d, q))$

MÔ HÌNH XÁC SUẤT

❖ XẾP HẠNG THEO MÔ HÌNH XÁC SUẤT

- Thay vì xếp hạng theo giá trị $p(r|d,q)$, có thể xếp hạng theo giá trị $\text{odds}(p(r|d,q))$.

→ Xếp hạng theo giá trị

$$p(d|r,q)/p(d|\neg r,q)$$

Giá trị $p(d|r,q)$ và $p(d|\neg r,q)$ được ước lượng tùy theo mô hình.

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Tần số của mỗi term là 0 (không xuất hiện) và 1 (có xuất hiện)

1) Giả thiết độc lập:

- Các term trong một tài liệu và một truy vấn độc lập với nhau
- Xác suất của một term xuất hiện trong các tài liệu liên quan không ảnh hưởng đến xác suất của các term khác trong các tài liệu liên quan

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

- 1) Giả thiết độc lập:
 - Giả thiết này đơn giản hóa việc tính toán và khá hiệu quả.

$$p(d|q, r) = \prod_{i \in [1, m]} p(td_i|q, r)$$

$$p(d|q, \neg r) = \prod_{i \in [1, m]} p(td_i|q, \neg r)$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

- 2) Các term của câu truy vấn là yếu tố duy nhất xác định sự liên quan giữa tài liệu và truy vấn:
- Với term $td_i \notin q$ thì $p(td_i|q, r)$ không phụ thuộc vào r :

$$p(td_i|q, r) = p(td_i|q, \neg r)$$

→ Chỉ cần tính xác suất các term trong truy vấn q

$$p(d|q, r) = \prod_{td \in q} p(td|q, r)$$

$$p(d|q, \neg r) = \prod_{td \in q} p(td|q, \neg r)$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Dựa trên hai giả thiết trên, sự liên quan được thể hiện qua giá trị:

$$\prod_{td \in q} p(td|r, q) / p(td|\neg r, q)$$

Việc ước lượng giá trị $p(td|r, q)$ và $p(td|\neg r, q)$ được thực hiện theo hai trường hợp:

- Trường hợp không có ngũ liệu mẫu
- Trường hợp có ngũ liệu mẫu

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

- ❖ Trường hợp không có ngũ liệu mẫu

Sử dụng hai giả thiết:

- Term của truy vấn xuất hiện hay không xuất hiện trong tài liệu liên quan là nhữ nhau:
$$p(tq_i|r,q) = 0.5$$
- Xác suất term xuất hiện trong tài liệu không liên quan (N_{td} : số tài liệu chứa td, N: tổng số tài liệu)
$$p(td|\neg r,q) = N_{td}/N$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

- ❖ Trường hợp không có ngũ liệu mẫu
 - Độ liên quan giữa tài liệu và truy vấn:

$$\begin{aligned} \text{rel}(d,q) &= \prod_{\text{td} \in q} p(\text{td}|r,q)/p(\text{td}|\neg r,q) \\ &= \prod_{\text{td} \in q} 0.5 * N/N_{\text{td}} \end{aligned}$$

Sử dụng độ đo log-odds:

$$\text{rel}(d,q) = \sum_{\text{td} \in q} \log(0.5 * N/N_{\text{td}})$$

→ Trọng số của mỗi term là $w_{\text{td}} = \log(0.5 * N/N_{\text{td}})$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Ví dụ: Tính độ liên quan giữa truy vấn q và các tài liệu sau theo mô hình BIM

d_1 Romeo and Juliet

d_2 Juliet: Oh happy dagger

d_3 Romeo died by dagger

q : die dagger

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

	Romeo	Juliet	happy	dagger	die
$d(td r,q)$	0.5	0.5	0.5	0.5	0.5
$d(td \neg r,q)$	2/3	2/3	1/3	2/3	1/3
w_{td}	-0.41	-0.41	0.58	-0.41	0.58

$$rel(d_1, q) = ?$$

$$rel(d_2, q) = -0.41$$

$$rel(d_3, q) = 0.17$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

- ❖ Trường hợp có ngũ liệu mẫu
 - r_{td} là số tài liệu liên quan chứa term td
 - N_R là tổng số tài liệu liên quan
 - Ước lượng các xác suất

$$p(td|r,q) = r_{td}/N_R$$

$$p(td|\neg r,q) = (N_{td}-r_{td})/(N-N_R)$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

- ❖ Trường hợp có ngũ liệu mẫu
 - Để tránh trường hợp $r_{td}=0$ và $r_{td}=N_{td}$, thực hiện smoothing:

$$p(td|r,q) = (r_{td}+0.5)/(N_R+1)$$

$$p(td|\neg r,q) = (N_{td}-r_{td}+0.5)/(N-N_R+1)$$

- Độ liên quan:

$rel(d,q)$

$$= \sum_{td \in q} \log([(r_{td}+0.5)*(N-N_R+1)]/[(N_{td}-r_{td}+0.5)*(N_R+1)])$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

❖ Trường hợp có ngũ liệu mẫu

- Trọng số của mỗi term:

$$w_{td} = \log([(r_{td} + 0.5) * (N - N_R + 1)] / [(N_{td} - r_{td} + 0.5) * (N_R + 1)])$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Ví dụ: Tính độ liên quan giữa truy vấn q và các tài liệu sau theo mô hình BIM. Biết $N=30$, $N_R=6$, $r_{die}=3$, $r_{dagger}=4$, $N_{die}=15$, $N_{dagger}=16$.

d_1 Romeo and Juliet

d_2 Juliet: Oh happy dagger

d_3 Romeo died by dagger

q : die dagger

MÔ HÌNH XÁC SUẤT

❖ TWO POISSON MODEL

Tần số của term là số lần xuất hiện của term trong tài liệu

→ Xác suất term td xuất hiện k lần trong tài liệu d là $p(td=k|d)$

Để ước lượng xác suất $p(td=f|d,r)$, giả thiết td tuân theo quy luật phân phối Poisson, khi đó:

$$p(td=k|d) = \lambda^k * e^{-\lambda} / k!, \quad k=0,1,2,\dots$$

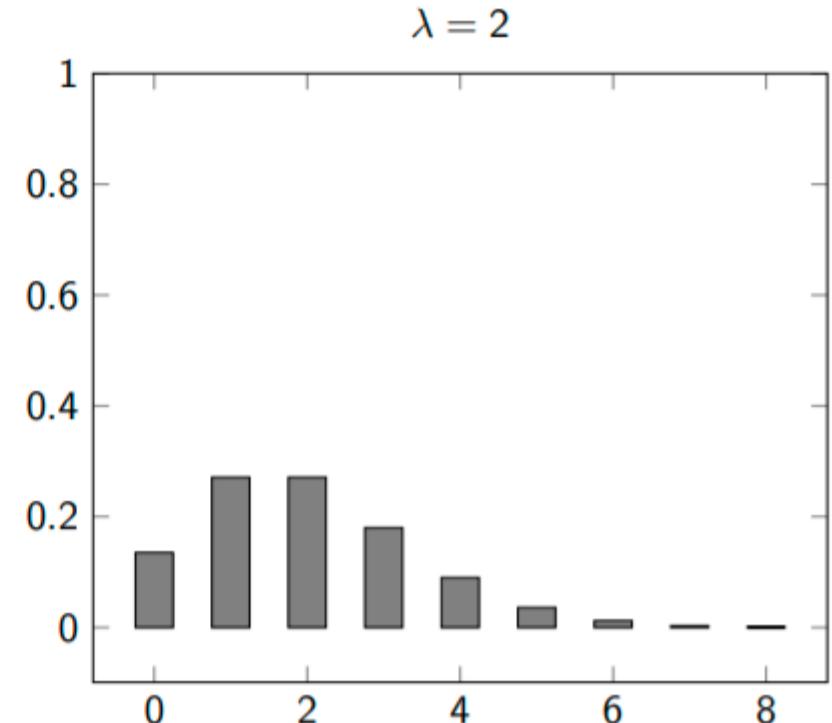
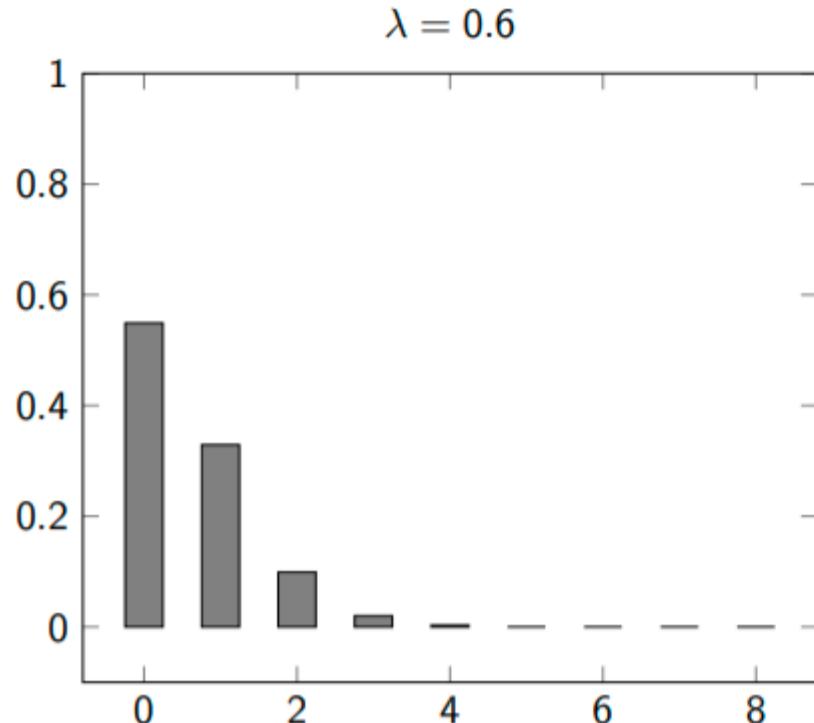
Giá trị λ được ước lượng như sau

$$\lambda = \text{count(td)} / N_{\text{term}}$$

MÔ HÌNH XÁC SUẤT

❖ TWO POISSON MODEL

Ví dụ phân phối xác suất Poisson



MÔ HÌNH XÁC SUẤT

❖ TWO POISSON MODEL

Tuy nhiên, một quy luật Poisson không thể hiện đúng thực tế nên sử dụng hai quy luật phân phối Poisson, gọi là Elite (E) và non-Elite($\neg E$):

$$p(td=k|r,d) = u^k \lambda^k e^{-\lambda} / k! + (1-u)^k \mu^k e^{-\mu} / k!, \quad k=0,1,2,\dots$$

$$p(td=k|\neg r,d) = v^k \lambda^k e^{-\lambda} / k! + (1-v)^k \mu^k e^{-\mu} / k!, \quad k=0,1,2,\dots$$

Trong đó:

- u là xác suất tài liệu là Elite và có liên quan
- v là xác suất tài liệu là Elite và không liên quan

MÔ HÌNH XÁC SUẤT

❖ TWO POISSON MODEL

Vì vấn đề ước lượng các tham số u, v, λ và μ , Robertson và Walker đề xuất một cách xếp xỉ theo hình dạng của hàm tính trọng số term sao cho:

- $w_{td} = 0$ nếu $k=0$
- w_{td} đồng biến với k
- w_{td} có dạng log-odds

Công thức đề nghị: $w'_{td} = [k/(k_1+k)]^*w_{td}$

Với w_{td} là trọng số được tính như mô hình BIM



ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG VI – PHÂN LỚP VĂN BẢN

Nguyễn Trọng Cảnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- ❖ VĂN ĐỀ PHÂN LỚP VĂN BẢN
- ❖ PHƯƠNG PHÁP GOM CỤM
- ❖ PHƯƠNG PHÁP PHÂN LỚP

VĂN ĐỀ PHÂN LỚP VĂN BẢN

❖ Nhu cầu phân lớp

Xuất phát từ yêu cầu cho phép tìm tài liệu theo chủ đề trong các thư viện:

- Nhóm các tài liệu theo các chủ đề chung
- Đặt tên cho nhóm, tên của mỗi nhóm được gọi là một lớp (class)

→ phân lớp văn bản (Text classification hoặc text categorization)

VĂN ĐỀ PHÂN LỚP VĂN BẢN

❖ Nhu cầu phân lớp

Phân lớp là một phương tiện để tổ chức thông tin:

- Quản lý số lượng tài liệu lớn thường xuyên được tạo
- Các tài liệu nếu được phân nhóm cẩn thận có thể sử dụng để hỗ trợ ra quyết định
- Là công nghệ quan trọng trong các tổ chức hiện đại.

VĂN ĐỀ PHÂN LỚP VĂN BẢN

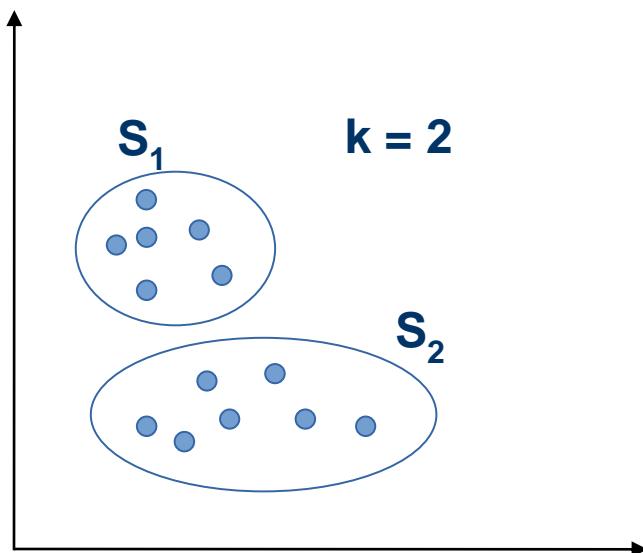
❖ Học máy (machine learning)

- Học các mẫu (pattern) từ dữ liệu
- Các mẫu này cho phép dự đoán trên dữ liệu mới
- Có ba dạng thuật toán học:
 1. Học có giám sát (supervised learning)
 2. Học không giám sát (unsupervised learning)
 3. Học nửa giám sát (semi-supervised learning)

PHƯƠNG PHÁP GOM CỤM

❖ Phát biểu bài toán

Cho một tập hợp các $S = \{x_1, x_2, \dots, x_n\}$. Chia tập hợp thành k tập hợp con $S_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$, $i=1,k$.



PHƯƠNG PHÁP GOM CỤM

❖ Phương pháp K-means

Cho một tập hợp các phần tử $S = \{x_1, x_2, \dots, x_n\}$.

- B1) Chọn k phần tử bất kỳ trong S làm k cụm ban đầu.
Thông thường chọn các phần tử có khoảng cách xa nhất. Gọi g_i là trọng tâm của cụm S_i
- B2) Với mỗi phần tử x_j còn lại, chọn cụm S_i có khoảng cách đến g_i nhỏ nhất.
- B3) Tính lại trọng tâm g_i của mỗi cụm S_i , bằng trung bình cộng của tất cả phần tử x_j trong cụm S_i

PHƯƠNG PHÁP GOM CỤM

❖ Phương pháp K-means

B4) Thực hiện lại bước B2 cho tới khi các phần tử x_j trong các cụm S_i không thay đổi, hoặc đến số lần thực hiện thứ N.

Ví dụ: gom cụm các điểm sau thành 2 cụm:

1 (1, 1); 2 (1.5, 2); 3 (3, 4); 4(5, 7); 5(3.5, 5);
6 (4.5, 5); 7(3.5, 4.5);

PHƯƠNG PHÁP GOM CỤM

Cụm $S_1 = \{1\}$, $g_1 = (1, 1)$; Cụm $S_2 = \{4\}$, $g_2 = (5, 7)$

Xét điểm 2 (1.5, 2):

- Khoảng cách đến S_1 : $d_1 = 1.11$

- Khoảng cách đến S_2 : $d_2 = 6.1$

$\rightarrow S_1 = \{1, 2\}$, $g_1 = (1.25, 1.5)$; $S_2 = \{4\}$, $g_2 = (5, 7)$

Xét điểm 3(3,4):

- Khoảng cách đến S_1 : $d_1 = 3.05$

- Khoảng cách đến S_2 : $d_2 = 3.6$

$\rightarrow S_1 = \{1, 2, 3\}$, $g_1 = (1.83, 2.33)$; $S_2 = \{4\}$, $g_2 = (5, 7)$

PHƯƠNG PHÁP GOM CỤM

Xét điểm 5 (3.5, 5):

- Khoảng cách đến S_1 : $d_1 = 3.14$

- Khoảng cách đến S_2 : $d_2 = 2.5$

$\rightarrow S_1 = \{1, 2, 3\}, g_1 = (1.83, 2.33); S_2 = \{4, 5\}, g_2 = (4.25, 6)$

Xét điểm 6(4.5, 5):

- Khoảng cách đến S_1 : $d_1 = 3.77$

- Khoảng cách đến S_2 : $d_2 = 1.03$

$\rightarrow S_1 = \{1, 2, 3\}, g_1 = (1.83, 2.33); S_2 = \{4, 5, 6\}, g_2 = (4.33, 5.66)$

PHƯƠNG PHÁP GOM CỤM

Xét điểm 7 (3.5, 4.5):

- Khoảng cách đến S_1 : $d_1 = 2.73$

- Khoảng cách đến S_2 : $d_2 = 1.43$

$\rightarrow S_1 = \{1, 2, 3\}$, $g_1 = (1.83, 2.33)$; $S_2 = \{4, 5, 6, 7\}$, $g_2 = (4.12, 5.37)$

PHƯƠNG PHÁP GOM CỤM

Với cụm $S_1 = \{1, 2, 3\}$, $g_1 = (1.83, 2.33)$; và

$S_2 = \{4, 5, 6, 7\}$, $g_2 = (4.12, 5.37)$

Xét lại khoảng cách các điểm.

Điểm	S_1	S_2
1	1.57	5.37
2	0.4	4.27
3	2.03	1.77
4	5.63	1.84
5	3.14	0.72
6	3.77	0.53
7	2.73	1.07

PHƯƠNG PHÁP GOM CỤM

Cập nhật lại cụm

$$S_1 = \{1, 2\}, g_1 = (1.25, 1.5); \text{ và}$$

$$S_2 = \{3, 4, 5, 6, 7\}, g_2 = (3.9, 5.1)$$

Xét lại khoảng cách các điểm.

Điểm	S_1	S_2
1	0.55	5.02
2	0.55	3.92
3	3.05	1.42
4	6.65	2.19
5	4.16	0.41
6	4.77	0.6
7	3.75	0.72

PHƯƠNG PHÁP GOM CỤM

Phần tử các cụm không đổi, vậy hai cụm được xác định là

$$S_1 = \{1, 2\}, g_1 = (1.25, 1.5); \text{ và}$$

$$S_2 = \{3, 4, 5, 6, 7\}, g_2 = (3.9, 5.1)$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Phát biểu bài toán

Cho một tập hợp các lớp $C = \{c_1, c_2, \dots, c_n\}$ và một phần tử d . Xác định lớp c_i chứa d . Biết một tập huấn luyện gồm m bộ được xác định thủ công

$$T = \{(d_1, c_{k1}), (d_2, c_{k2}), \dots, (d_m, c_{km})\}$$

Trong đó (d_i, c_{ki}) cho biết phần tử d_i thuộc lớp c_{ki}

Xác định hàm phân lớp $f: d \rightarrow c, c \in C$

Các tham số của hàm f được gọi là mô hình của bộ phân lớp

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

- Phân lớp dựa trên công thức xác suất có điều kiện (Bayes)
- Phân lớp dựa trên các đặc trưng của phần tử cần phân lớp. Đối với phần tử là một văn bản, đặc trưng có thể là các ký hiệu phân tách nhau bằng khoảng trắng.
- Số lượng các lớp là 2 (+, -).

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

- Cho một phần tử d và một lớp c, xác suất để d thuộc lớp c là:

$$P(c | d) = P(d | c) * P(c) / P(d)$$

- Lớp c_i sẽ là lớp chứa phần tử d nếu

$$P(c_i | d) \geq P(c_j | d)$$

$$\rightarrow P(d | c_i) * P(c_i) / P(d) \geq P(d | c_j) * P(c_j) / P(d)$$

$$\rightarrow P(d | c_i) * P(c_i) \geq P(d | c_j) * P(c_j)$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

Giả sử phần tử d có các đặc trưng x_k , $k = 1, n$

Khi đó,

$$P(d | c_i) = P(x_1, x_2, x_3, \dots, x_n | c_i)$$

Để tính được xác suất này một cách dễ dàng, cần có hai giả thiết:

- 1) Vị trí của x_k trong dãy các đặc trưng không ảnh hưởng đến xác suất.
- 2) Các đặc trưng x_k là độc lập với nhau.

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

Khi đó,

$$\begin{aligned} P(d | c_i) &= P(x_1, x_2, x_3, \dots, x_n | c_i) \\ &= P(x_1 | c_i) * P(x_2 | c_i) * \dots * P(x_n | c_i) \end{aligned}$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

* Giai đoạn học:

Giai đoạn học sẽ xác định các giá trị:

- $P(c_i) = (\text{số cặp chứa } c_i \text{ trong tập } T) / m$
- $P(x_k | c_i) = (\text{số lần xuất hiện } x_k \text{ trong } c_i) / (\text{số lần xuất hiện của tất cả đặc trưng trong } c_i)$

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

* Giai đoạn học:

Do có những đặc trưng x_u không xuất hiện trong các phần tử của lớp c_t nhưng lại có trong những phần tử của lớp khác. Khi đó $P(x_u | c_t) = 0$

$$\begin{aligned}\rightarrow P(d | c_i) &= P(x_1, x_2, x_3, \dots, x_n | c_i) \\ &= P(x_1 | c_i) * P(x_2 | c_i) * \dots * P(x_n | c_i) \\ &= 0\end{aligned}$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

* Giai đoạn học:

Vì thế, cần dùng ước lượng Laplace:

$$P(x_k | c_i) = ((\text{số lần xuất hiện } x_k \text{ trong } c_i) + 1) / ((\text{số lần xuất hiện của tất cả đặc trưng trong } c_i) + n_v)$$

Với n_v là số lượng các đặc trưng x trong tất cả các phần tử của tất cả các lớp. Tập các đặc trưng này ký hiệu là V .

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

* Giai đoạn áp dụng:

Sử dụng công thức xác suất có điều kiện dựa trên các ước lượng đã tính trong giai đoạn học để chọn lớp có xác suất cao nhất.

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

Ví dụ: Cho hai lớp là “China” và “Japan”, và tập huấn luyện gồm các đoạn văn bản với phân lớp của nó như sau:

- 1) Chinese Beijing Chinese (China)
- 2) Chinese Chinese Shanghai (China)
- 3) Chinese Macao (China)
- 4) Tokyo Japan Chinese (Japan)

Cho biết tài liệu sau thuộc lớp nào?

- 5) Chinese Chinese Chinese Tokyo Japan

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp Naive Bayes

Một số đặc điểm áp dụng phương pháp Naive Bayes

- Dễ dàng tính toán
- Xử lý được số đặc trưng rất lớn
- Với tập dữ liệu lớn, độ chính xác của phân lớp sẽ cao.

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp K - Nearest Neighboor (KNN)

- Dùng chính tập huấn luyện làm mô hình
- Xác định k phần tử trong tập huấn luyện gần với phần tử d nhất.
- Chọn phân lớp nào có nhiều thành viên được xác định nhất.

PHƯƠNG PHÁP PHÂN LỚP

❖ Phương pháp K - Nearest Neighboor

Ví dụ: Cho tập huấn luyện gồm tập S các điểm trong không gian hai chiều với phân lớp tương ứng của chúng gồm (+) và (-). Xác định phân lớp của điểm d với k=3

$$S = \{ [(1,2), +], [(2,3), +], [(2,1), -], [(3, 2), -] \}$$

$$d(1, 1.5)$$

PHƯƠNG PHÁP PHÂN LỚP

- Khoảng cách từ d đến (1,2): 0.5
 - Khoảng cách từ d đến (2,3): 1.802
 - Khoảng cách từ d đến (2,1): 1.118
 - Khoảng cách từ d đến (3,2): 2.061
- Phần tử thuộc lớp (+) được chọn là (1,2) và (2,3)
Phần tử thuộc lớp (-) được chọn là (2,1)
Vậy, điểm d(1,1.5) thuộc lớp (+)

PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

Cho hai tập tách rời nhau A và B chứa các điểm trong không gian n chiều R^n . Xác định siêu phẳng H có dạng $\langle W, X \rangle = b$ tách rời A và B. Có nghĩa là:

$$\langle W, X \rangle > b, \forall X \in A$$

$$\langle W, X \rangle < b, \forall X \in B$$

Trong đó:

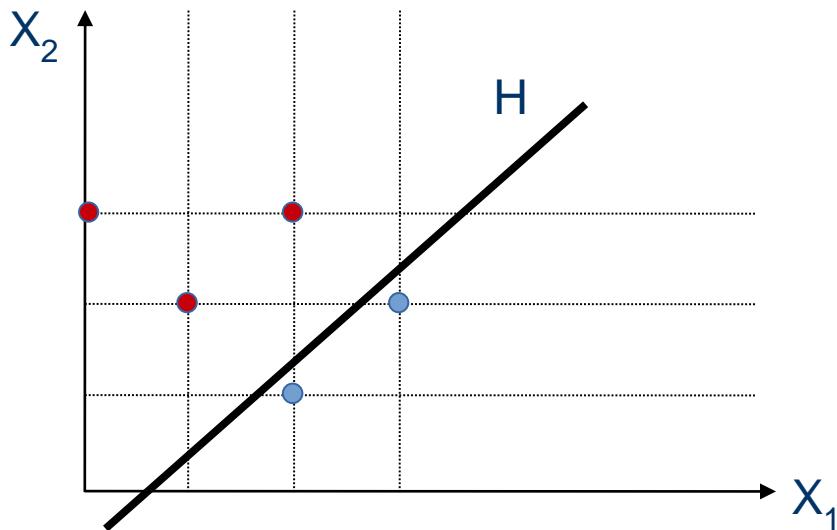
- W là bộ tham số,
- X là vector trong R^n
- b là giá trị ngưỡng.
- $\langle W, X \rangle$ là tích vô hướng giữa W và X

PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

Ví dụ: Trong không gian 2 chiều, cho $A = \{(0, 3), (1, 2), (2, 3)\}$ và $B = \{(2, 1), (3, 2)\}$. Đường thẳng H tách rời hai tập A và B có phương trình:

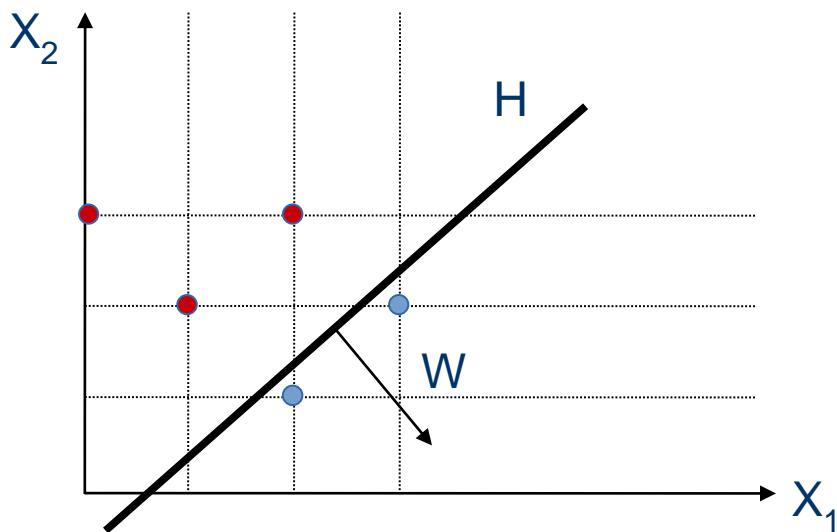
$$H: 1.5x_1 - 1.5x_2 = 1$$



PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

Bộ tham số W là vector pháp tuyến của (H). Để điều chỉnh độ dốc và vị trí của (H) cần điều chỉnh W và b sao cho H tách rời A và B.



PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

Phương pháp xác định W và b với hai tập A, B :

1) Từ phương trình $\langle W, X \rangle = b$, chuyển về dạng

$$\langle W', X' \rangle = 0$$

Điều kiện tách rời:

$$\langle W', X' \rangle > 0, \forall X \in A$$

$$\langle W', X' \rangle < 0, \forall X \in B$$

Trong đó:

- $W' = (W, -b) = (w_1, w_2, \dots, w_n, -b)$
- $X' = (X, 1) = (x_1, x_2, \dots, x_n, 1)$

PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

2) Xác định W' theo thuật toán sau:

Bước 1: Khởi tạo $W' = (0, 0, \dots, 0, 1)$.

Bước 2: lần lượt xét tất cả các điểm X_i :

- Nếu $X_i \in A$ và $\langle W', X_i \rangle \leq 0$ thì qua Bước 3, ngược lại tiếp tục Bước 2.
- Nếu $X_i \in B$ và $\langle W', X_i \rangle \geq 0$ thì qua Bước 3, ngược lại tiếp tục Bước 2.

PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

Bước 3:

- Nếu $X_i \in A$, thì $W' = W' + \delta X_i$, tiếp tục Bước 2.
- Nếu $X_i \in B$, thì $W' = W' - \delta X_i$, tiếp tục Bước 2.

Trong đó, $\delta > 0$ là một hằng số qui định tỉ lệ cập nhật

Thuật toán dừng khi

$$\langle W', X' \rangle > 0, \forall X \in A$$

$$\langle W', X' \rangle < 0, \forall X \in B$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Perceptron

Ví dụ: Trong không gian 2 chiều, cho $A = \{(0, 3), (1, 2), (2, 3)\}$ và $B = \{(2, 1), (3, 2)\}$. Xác định đường thẳng H tách rời hai tập A và B:

Có $W' = (0, 0, 1)$; chọn $\delta = 0.5$

- Xét điểm $(0, 3)$ thuộc A, có

$$0 * 0 + 0 * 3 + 1 * 1 = 1 > 0$$

- Xét điểm $(1, 2)$ thuộc A, có

$$0 * 1 + 0 * 2 + 1 * 1 = 1 > 0$$

PHƯƠNG PHÁP PHÂN LỚP

- Xét điểm $(2, 3)$ thuộc A, có

$$0 * 2 + 0 * 3 + 1 * 1 = 1 > 0$$

- Xét điểm $(2, 1)$ thuộc B, có

$$0 * 2 + 0 * 1 + 1 * 1 = 1 > 0$$

→ Cập nhật lại W'

$$W' = W' - 0.5 * (2, 1, 1) = (0, 0, 1) - (1, 0.5, 0.5)$$

$$= (-1, -0.5, 0.5)$$

- Xét điểm $(3, 2)$ thuộc B, có

$$-1 * 3 - 0.5 * 2 + 0.5 * 1 = -3.5 < 0$$

PHƯƠNG PHÁP PHÂN LỚP

Do có điểm chưa thỏa trong lần xét các tập A, B nên thực hiện lại.

- Xét điểm (0, 3) thuộc A, có

$$-1 * 0 - 0.5 * 3 + 0.5 * 1 = -1 < 0$$

→ Cập nhật lại W'

$$\begin{aligned}W' &= W' + 0.5 * (0, 3, 1) = (-1, -0.5, 0.5) + (0, 1.5, 0.5) \\&= (-1, 1, 1)\end{aligned}$$

PHƯƠNG PHÁP PHÂN LỚP

- Xét điểm $(1, 2)$ thuộc A, có

$$-1 * 1 + 1 * 2 + 1 * 1 = 2 > 0$$

- Xét điểm $(2, 3)$ thuộc A, có

$$-1 * 2 + 1 * 3 + 1 * 1 = 2 > 0$$

- Xét điểm $(2, 1)$ thuộc B, có

$$-1 * 2 + 1 * 1 + 1 * 1 = 0$$

→ Cập nhật lại W'

$$\begin{aligned} W' &= W' - 0.5 * (2, 1, 1) = (-1, 1, 1) - (1, 0.5, 0.5) \\ &= (-2, 0.5, 0.5) \end{aligned}$$

PHƯƠNG PHÁP PHÂN LỚP

- Xét điểm $(3, 2)$ thuộc B, có

$$-2 * 3 + 0.5 * 2 + 0.5 * 1 = -4.5 < 0$$

Do có điểm chưa thỏa trong lần xét các tập A, B nên thực hiện lại.

- Xét điểm $(0, 3)$ thuộc A, có

$$-2 * 0 + 0.5 * 3 + 0.5 * 1 = 2 > 0$$

PHƯƠNG PHÁP PHÂN LỚP

- Xét điểm $(1, 2)$ thuộc A, có

$$-2 * 1 + 0.5 * 2 + 0.5 * 1 = -0.5 < 0$$

→ Cập nhật lại W'

$$\begin{aligned} W' &= W' + 0.5 * (1, 2, 1) = (-2, 0.5, 0.5) + (0.5, 1, 0.5) \\ &= (-1.5, 1.5, 1) \end{aligned}$$

- Xét điểm $(2,3)$ thuộc A, có

$$-1.5 * 2 + 1.5 * 3 + 1 * 1 = 2.5 > 0$$

PHƯƠNG PHÁP PHÂN LỚP

Do có điểm chưa thỏa trong lần xét các tập A, B nên thực hiện lại.

- Xét điểm (0, 3) thuộc A, có

$$-1.5 * 0 + 1.5 * 3 + 1 * 1 = 5.5 > 0$$

- Xét điểm (1, 2) thuộc A, có

$$-1.5 * 1 + 1.5 * 2 + 1 * 1 = 2.5 > 0$$

- Xét điểm (2, 3) thuộc A, có

$$-1.5 * 2 + 1.5 * 3 + 1 * 1 = 2.5 > 0$$

- Xét điểm (2, 1) thuộc B, có

$$-1.5 * 2 + 1.5 * 1 + 1 * 1 = -0.5 < 0$$

PHƯƠNG PHÁP PHÂN LỚP

- Xét điểm (3, 2) thuộc B, có

$$-1.5 * 3 + 1.5 * 2 + 1 * 1 = -0.5 < 0$$

Do các điểm trong A và B đều thỏa mãn kết quả là:

$$W' = (-1.5, 1.5, 1)$$

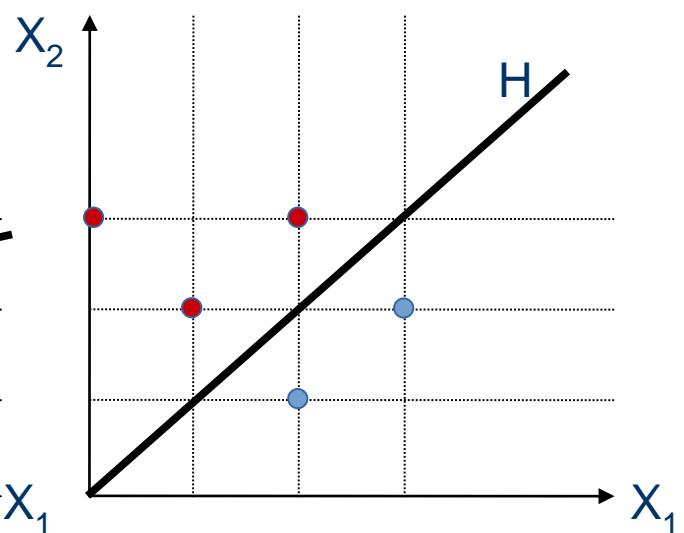
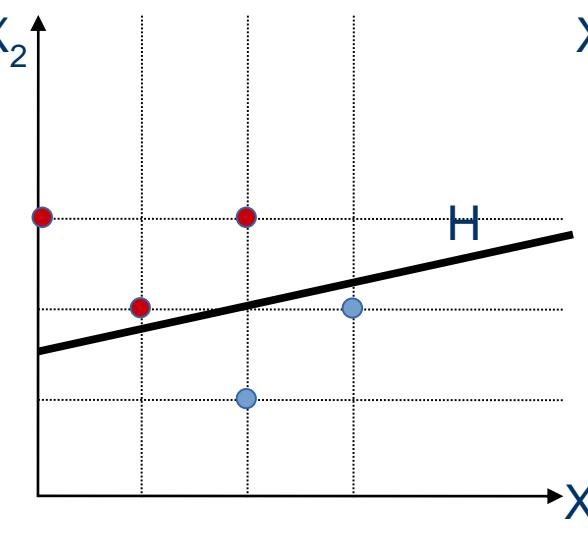
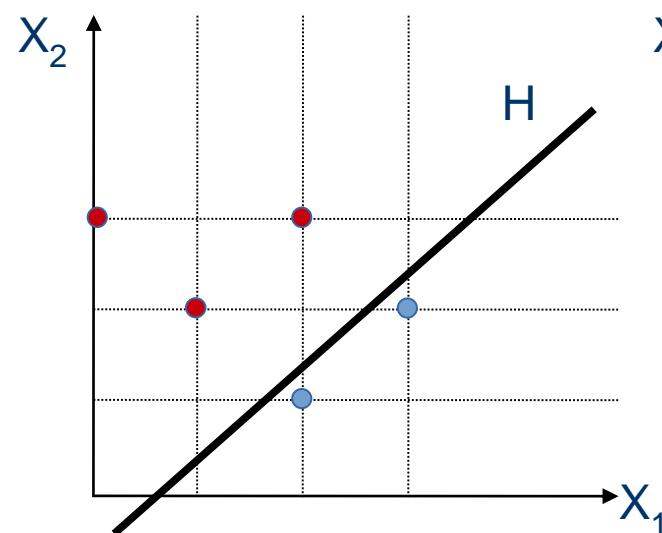
Hay đường thẳng H có dạng:

$$H: -1.5x_1 + 1.5x_2 + 1 = 0$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

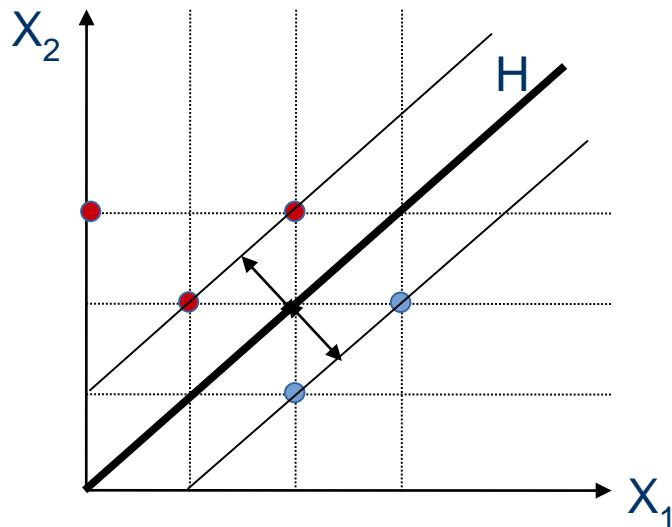
Ví dụ: Trong không gian 2 chiều, cho $A = \{(0, 3), (1, 2), (2, 3)\}$ và $B = \{(2, 1), (3, 2)\}$. Đường thẳng H tốt nhất tách rời hai tập A và B là đường nào



PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Phương pháp SVM sẽ xác định siêu phẳng H tách rời hai tập A và B sao cho khoảng cách từ biên của mỗi tập điểm đến H là bằng nhau và là lớn nhất.



PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Giả sử siêu phẳng cần xác định có dạng

$$H: \langle W, X \rangle = 0$$

Khi đó, hai siêu phẳng biên tương ứng với các tập A, B lần lượt là:

$$H_A: \langle W, X \rangle = 1$$

$$H_B: \langle W, X \rangle = -1$$

Khoảng cách giữa một điểm bất kỳ trên H_A đến H_B là:

$$|\langle W, X \rangle + 1| / \|W\| = 2 / \|W\|$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Giả sử để phân biệt điểm thuộc tập A và B, dùng biến $Y \in \{-1, 1\}$ trong đó những điểm a thuộc A có $Y_a = 1$ và b thuộc B có $Y_b = -1$. Khi đó, vector W cần thỏa điều kiện:

$$Y_i * \langle W, X_i \rangle \geq 1, Y_i \in A \cup B$$

Như vậy, bài toán xác định W là bài toán cực đại hóa với hàm mục tiêu và tập ràng buộc lần lượt là

$$\max_W f(W) = 2 / \|W\|$$

$$Y_i * \langle W, X_i \rangle \geq 1, Y_i \in A \cup B$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Bài toán tương đương là

$$\min_W f(W) = 1/2 \|W\|^2$$

$$g_i(W) = 1 - Y_i * \langle W, X_i \rangle \leq 0, Y_i \in A \cup B$$

Lời giải của bài toán W_0 được xác định theo hệ sau:

$$\begin{cases} \delta(f(W) + \sum \alpha_i * g_i(W)) / \delta_W = 0 \\ g_i(W) \leq 0 \end{cases}$$

Trong đó $C \geq \alpha_i \geq 0$ là các hệ số Lagrange

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

- Kết quả của hệ là W_0 và bộ các hệ số Lagrange α_i .
- Những điểm tương ứng với $\alpha_i > 0$ được gọi là Support Vector vì chỉ những điểm này quyết định kết quả phân lớp.
- Xác định điểm $D(d_1, d_2, \dots, d_n)$ thuộc tập A hay B, tính:

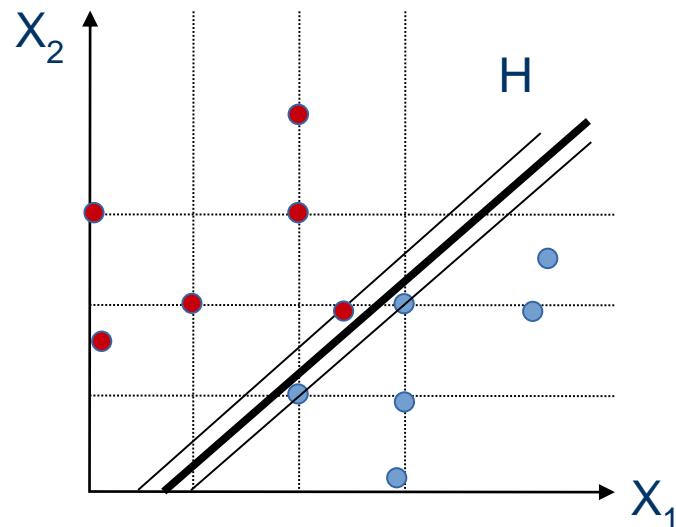
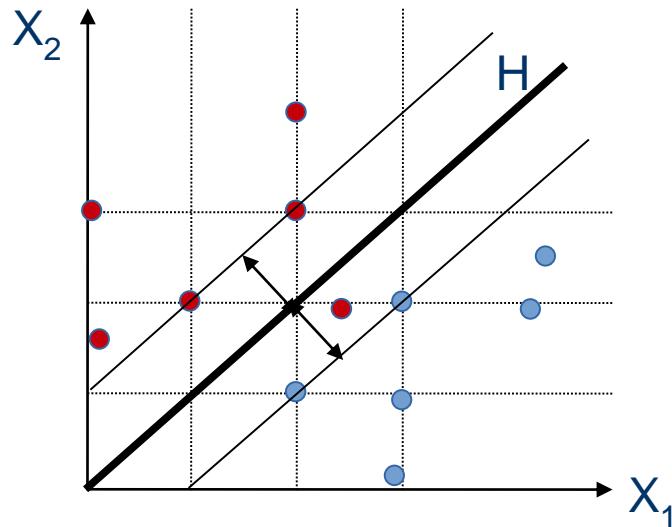
$$F = \sum \alpha_i * Y_i * \langle X_i, D \rangle + b$$

nếu $F \geq 0$ thì D thuộc tập A, ngược lại D thuộc tập B.

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Xét trường hợp các điểm trong tập A và B như bên dưới, đường thẳng H nào sẽ tốt?



PHƯƠNG PHÁP PHÂN LỚP

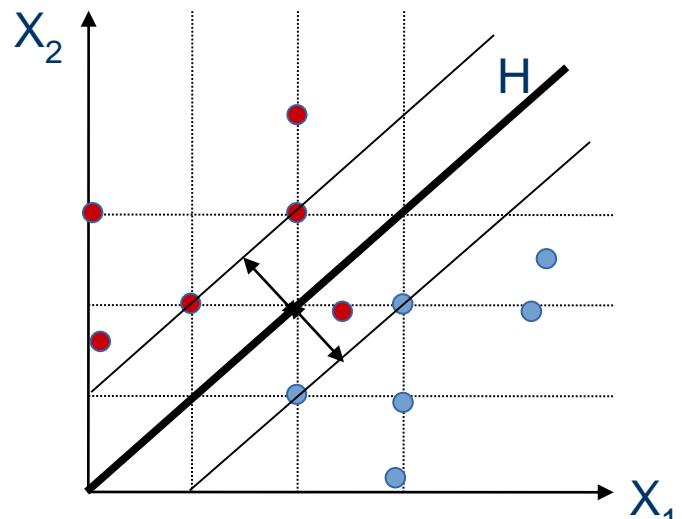
❖ Support Vector Machine (SVM)

Để tìm được siêu phẳng H như hình bên dưới, H_A và H_B cần có dạng như sau:

$$H_A: \langle W, X \rangle \geq 1 - \varepsilon$$

$$H_B: \langle W, X \rangle \leq -1 + \varepsilon$$

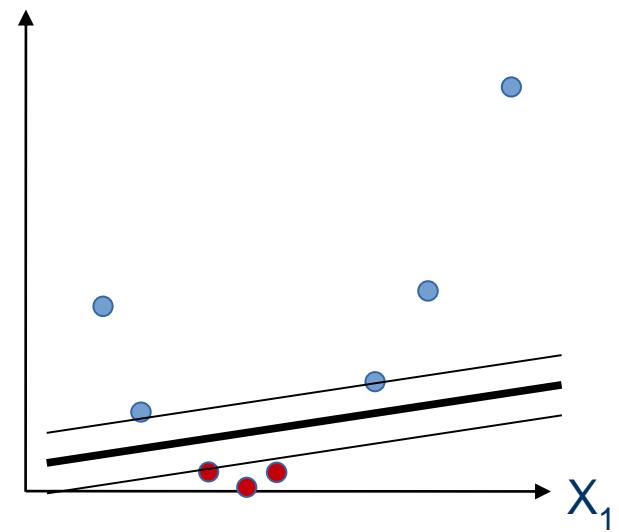
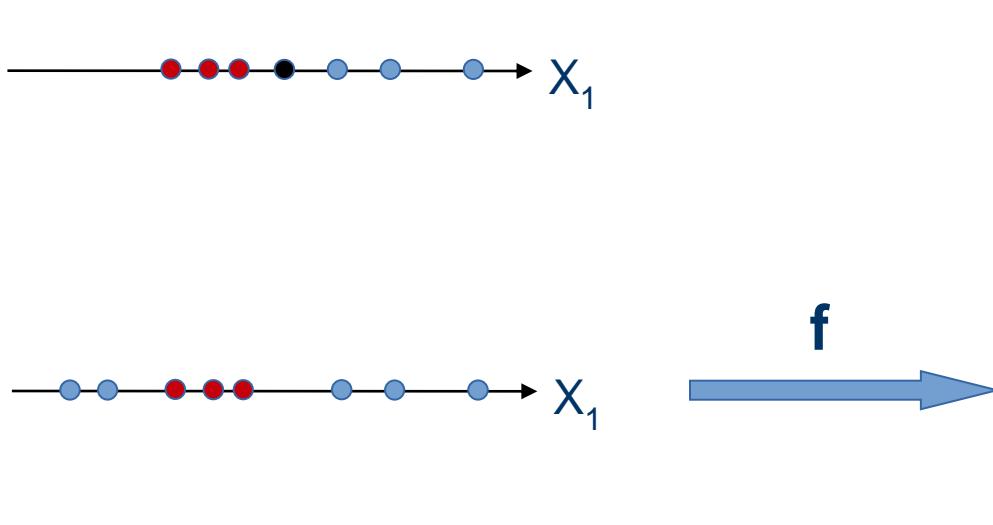
Trong đó ε được gọi là biến slack, khi đó, siêu phẳng H_A và H_B được gọi là siêu phẳng mềm (Soft Hyperlane)



PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Trường hợp hai tập A, B không tách rời tuyến tính như hình bên dưới, ngoài cách dùng biến slack, có thể chuyển các điểm trong A và B vào một không gian có số chiều lớn hơn gọi là không gian đặc trưng (Feature space).



PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Thay vì xác định không gian đặc trưng cụ thể, hàm kernel được sử dụng để:

- Tránh việc tính toán trên không gian nhiều chiều.
- Kết quả tính toán tương đương.

Hàm kernel được dùng thay thế cho $\langle X_i, D \rangle$, có dạng:

- + Polynominal: $K(X, Y) = (X^T Y + 1)^d$
- + Radial: $K(X, Y) = \exp(-||X - Y||^2 / (2 * \gamma^2))$

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Chẳng hạn, chọn hàm kernel là hàm polynominal:

$$K(X, Y) = K(X, Y) = (X^T Y + 1)^2$$

Khi đó, sau khi dùng hàm kernel để tìm bộ hệ số Lagrange α_i , để xác định điểm D (d_1, d_2, \dots, d_n) thuộc tập A hay B, xác định giá trị:

$$\begin{aligned} F &= \sum \alpha_i * Y_i * K(X_i, D) + b \\ &= \sum \alpha_i * Y_i * (X_i^T D + 1)^2 + b \end{aligned}$$

PHƯƠNG PHÁP PHÂN LỚP

❖ Support Vector Machine (SVM)

Lưu ý: khi sử dụng SVM cần xác định:

- C là hằng số chặn trên của α_i
- Hàm kernel (Polynomial, Radial, ...)
- ε là giá trị của biến slack cho phép dùng siêu phẳng mềm để xác định siêu phẳng H.