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

TRUY XUẤT THÔNG TIN

CHƯƠNG I - DẪN NHẬP

Nguyễn Trọng Chỉnh
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 VẤN 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

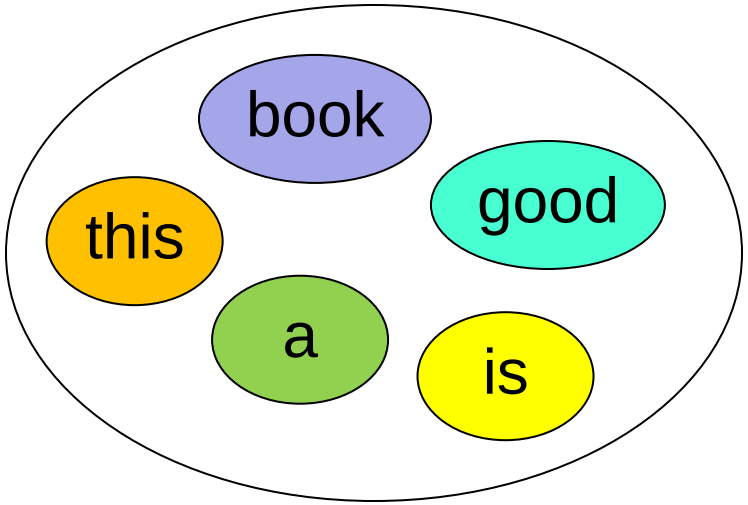
CÁC MÔ HÌNH IR CĂN BẢN

❖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.”



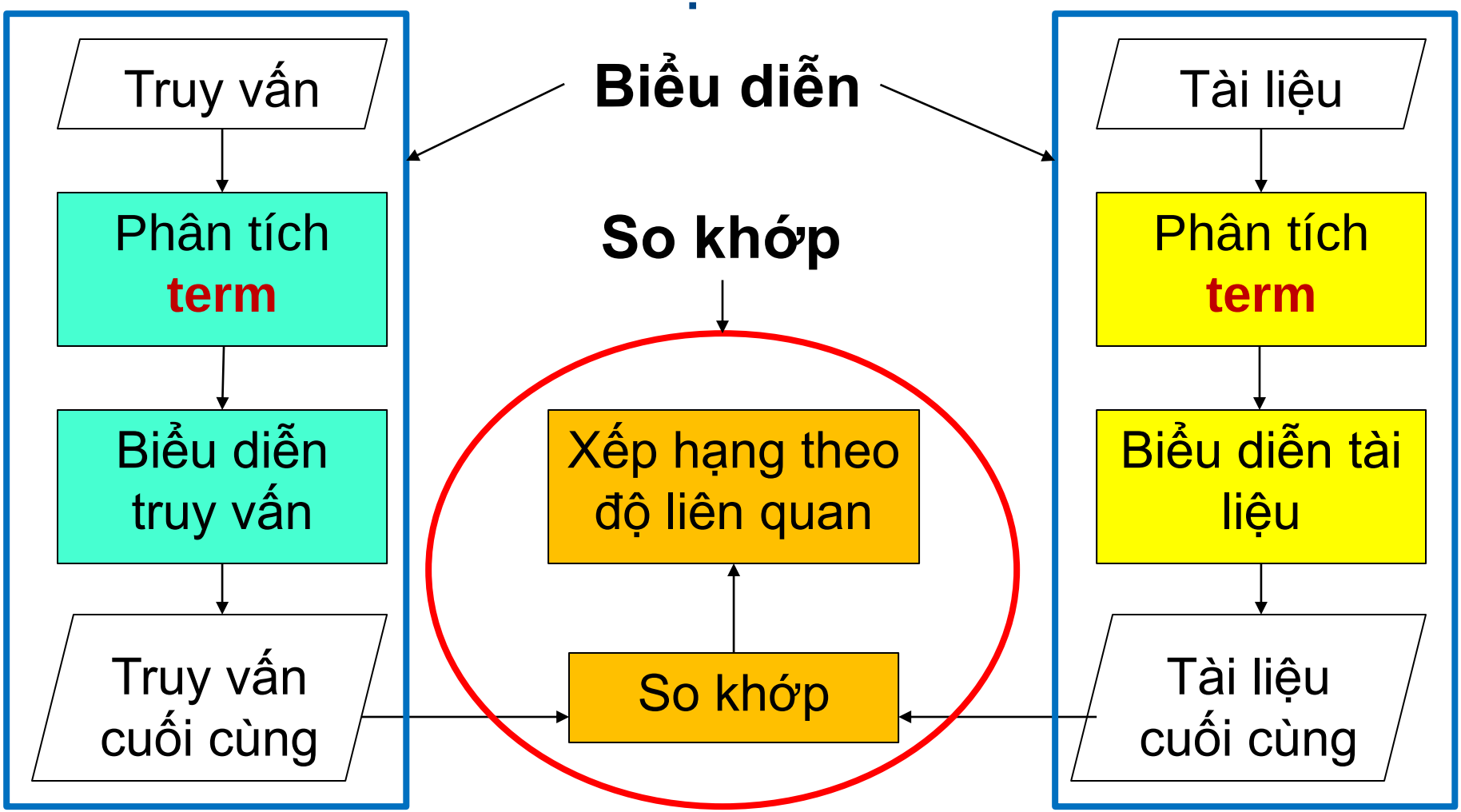
CÁC MÔ HÌNH IR CĂN BẢN

❖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 Euclide
  - 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

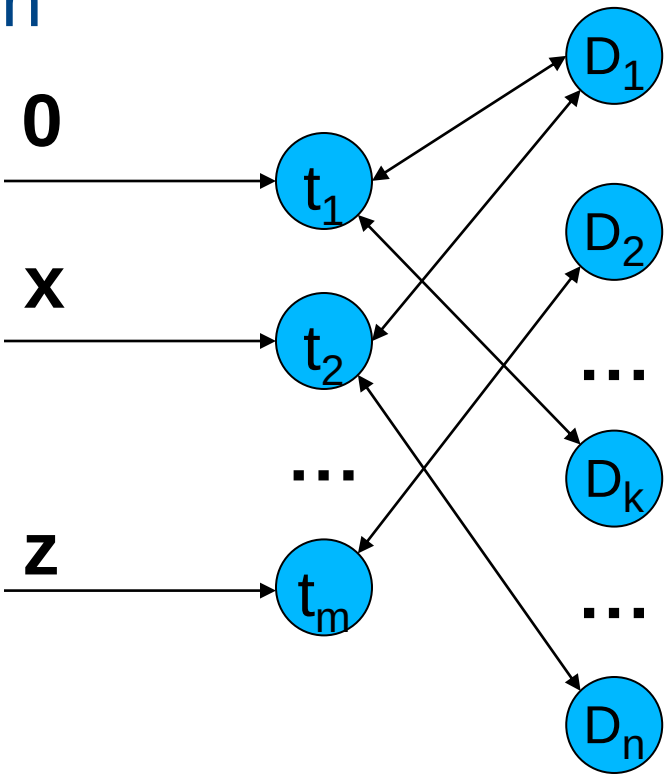
- **Mô hình mạng neural**

- *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 → 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 (∩):
  - 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 phổ 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 ⇔ chứa hạng từ ¬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 ¬con AND ¬trỏ 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: $Q_1 = t_1 \wedge t_2$,

$Q_2 = (t_1 \wedge t_2 \vee t3) \wedge (t4 \vee \neg(\neg t5 \wedge t6))$

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

→ 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

- $Sim_D(t_1 \vee t_2) = \sqrt{\frac{w_{t_1}^2 + w_{t_2}^2}{2}}$
- $Sim_D(t_1 \wedge t_2) = 1 - \sqrt{\frac{(1-w_{t_1})^2 + (1-w_{t_2})^2}{2}}$
- $Sim_D(\neg t_1) = 1 - w_{t_1}$

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:

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

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

$Q_1 = t_1 \wedge t_2$ $\qquad$ $Q_1 = t_1 \vee t_2$

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