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

# TRUY XUẤT THÔNG TIN

## CHƯƠNG I - DẪN NHẬP

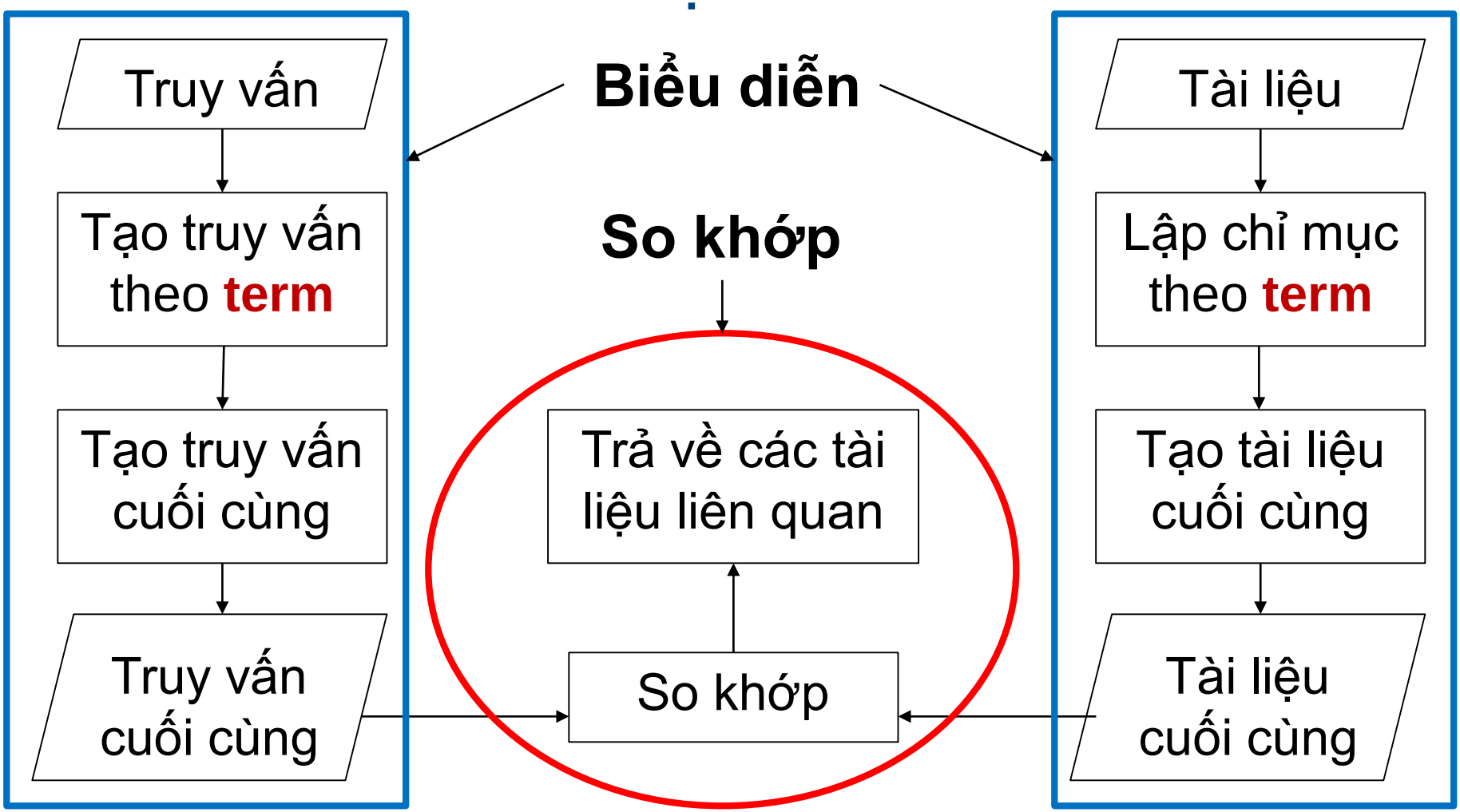
Nguyễn Trọng Chỉnh
chinhnt@uit.edu.vn

# 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

## ❖CƠ SỞ PHÂN LOẠI MÔ HÌNH IR



# CÁC MÔ HÌNH IR CĂN BẢN

## ❖CƠ SỞ PHÂN LOẠI MÔ HÌNH IR

Các mô hình so khớp (matching model) gồm:

- **Mô hình theo lý thuyết tập hợp:**
  - Boolean
  - Extended Boolean
  - Fuzzy
- **Mô hình đại số** (mô hình Vector)
- **Mô hình xác suất**
- **Mô hình khác** (Các mô hình mạng neuron)

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- Vấn đề so khớp được giải quyết dựa trên đại số Bool
- Mục tiêu là xác định các tập hợp thành viên và xử lý các phép toán trên tập hợp
- Các tập hợp thành viên thường được xác định dựa trên việc có hay không có sự xuất hiện của một từ khóa (còn gọi là từ vựng) trong một tài liệu.

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Các phép toán**

| Tập hợp | | Đại số Bool | |
|---|---|---|---|
| Phép hợp | $\cup$ | Phép tuyển (AND) | $\vee$ |
| Phép giao | $\cap$ | Phép hội (OR) | $\wedge$ |
| Phép lấy phần bù | \ | Phép phủ định (NOT) | $\overline{A}$ |

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Truy vấn trong mô hình Boolean**

Truy vấn gồm:

- Các biểu thức Boolean được biểu diễn dưới dạng biểu thức trung tố (phép phủ định có dạng tiền tố).

  VD: Romeo AND Ebook

- Ngôn ngữ tự nhiên.

  VD: Romeo Ebook

- Ngôn ngữ tự nhiên có quy ước.

  VD: Title:Romeo Type:Ebook

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Truy vấn trong mô hình Boolean**

Truy vấn được xử lý như sau:

- Mỗi tài liệu là một phần tử trong tập hợp
- Mỗi từ khóa là một tập hợp.
- Truy vấn được khử phép toán phủ định biểu thức theo các luật De Morgan.
- Truy vấn được phân tích theo dạng cây với nút lá là các từ khóa, nút trong là các phép toán hội và tuyển. Phép phủ định được gắn liền từ khóa.

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Truy vấn trong mô hình Boolean**

Truy vấn được xử lý như sau:

- Trường hợp truy vấn chỉ chứa phủ định các từ khóa thì không xử lý.
- Thực hiện các phép toán trên cây truy vấn theo thứ tự từ node lá về gốc.

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Ví dụ**

Truy vấn: Romeo OR NOT(Juliet OR NOT Hamlet)

Giả sử có các tài liệu sau, X cho biết tài liệu có chứa

| DOC | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **Romeo** | x | x | x | x | | | | |
| **Hamlet** | x | x | | | x | X | | |
| **Juliet** | x | | x | | x | | x | |

# CÁC MÔ HÌNH IR CĂN BẢN

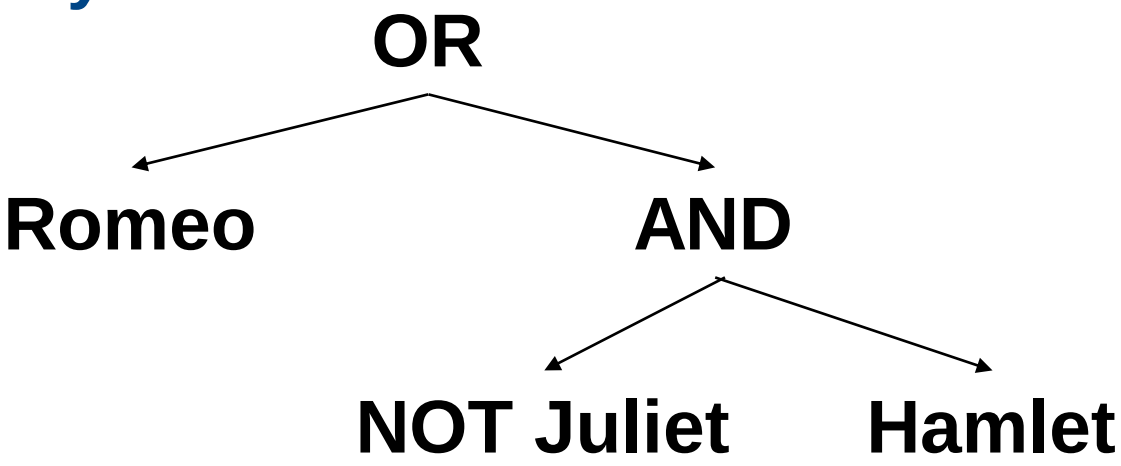
## ❖MÔ HÌNH BOOLEAN

- **Ví dụ**

Áp dụng luật De Morgan cho truy vấn được truy vấn mới

Romeo OR NOT Juliet AND Hamlet

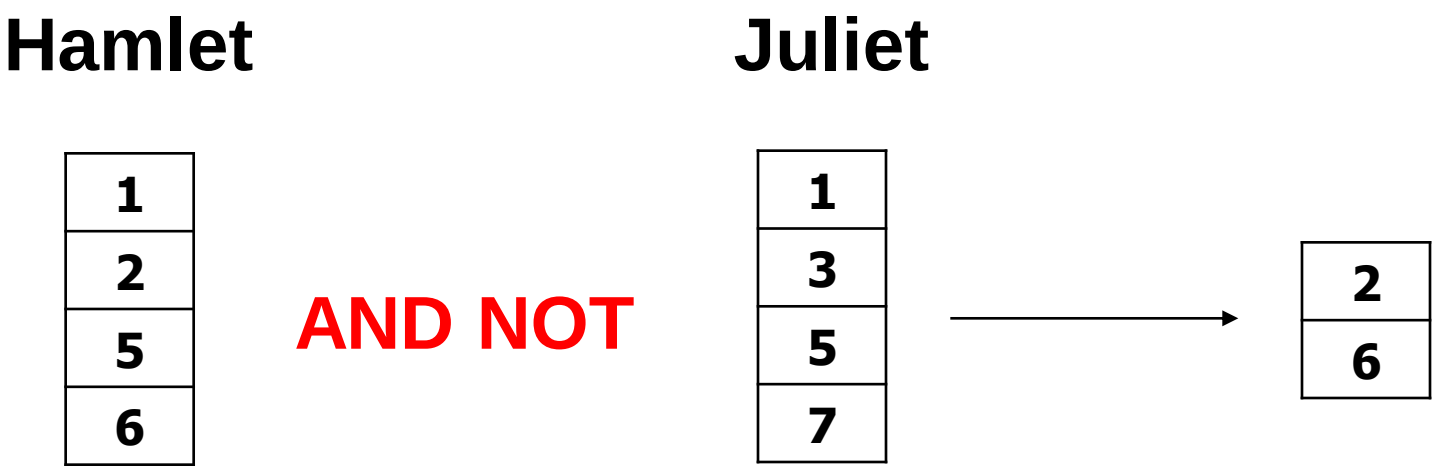
**Cây truy vấn:**



# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Ví dụ**

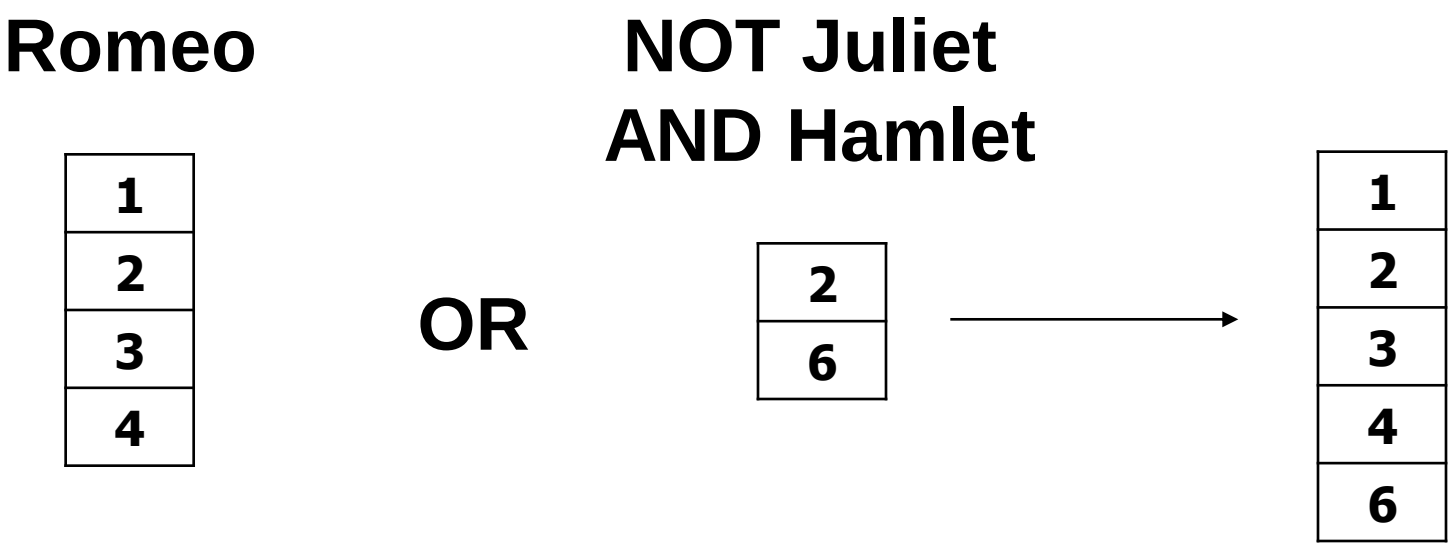



Lưu ý phép phủ định luôn đi kèm với phép hội

# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

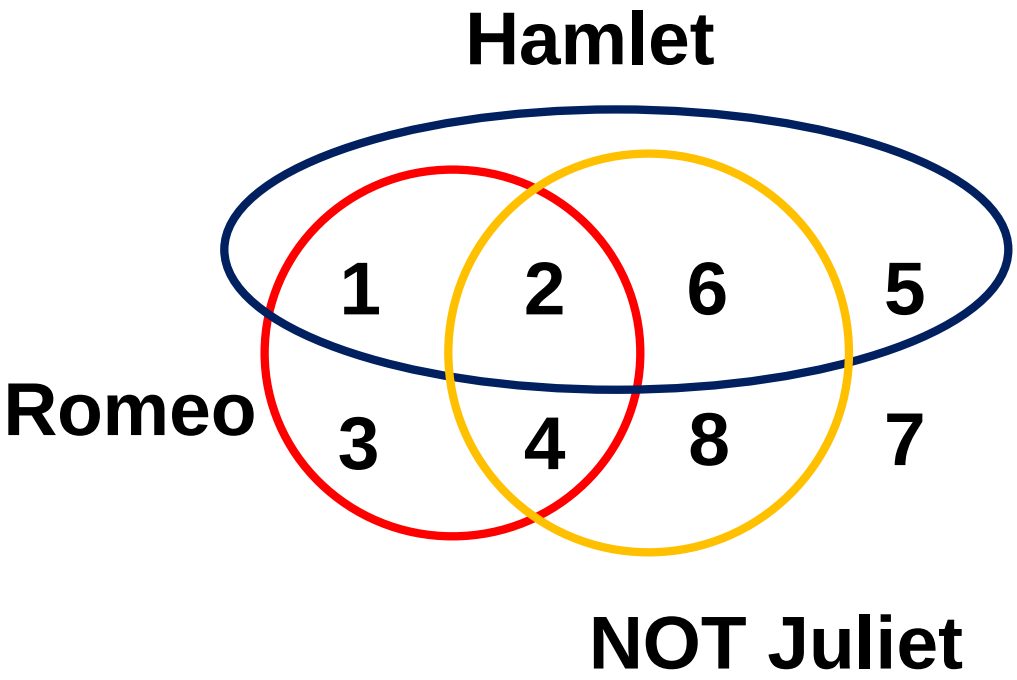
- **Ví dụ**



# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

Kết quả truy vấn là các tài liệu: 1, 2, 3, 4, 6,



# CÁC MÔ HÌNH IR CĂN BẢN

## ❖MÔ HÌNH BOOLEAN

- **Ưu điểm của mô hình Boolean**
- Truy vấn đơn giản và dễ hiểu
- Tương đối dễ cài đặt
- **Nhược điểm của mô hình Boolean**
- Khó khăn trong việc thể hiện chính xác yêu cầu thông tin
- Kết quả trả về hoặc quá nhiều hoặc quá ít.
- Kết quả trả về không có thứ tự cho biết mức độ liên quan của tài liệu so với truy vấn.

# LẬP CHỈ MỤC

## ❖KHÁI NIỆM

- Để xác định tài liệu d có chứa từ khóa t cần phải duyệt toàn bộ nội dung.
- Trường hợp tài liệu d dài và kích thước tập lưu trữ lớn → thời gian tìm kiếm rất lâu

→ Lập chỉ mục là việc tổ chức lại tài liệu sao cho việc xác định tài liệu chứa từ khóa nào đó trở nên hiệu quả

# LẬP CHỈ MỤC

## ❖CHỈ MỤC ĐẢO NGƯỢC

Chỉ mục đảo ngược (Inverted Index) là cách lưu trữ các từ khóa sao cho việc xác định những tài liệu chứa từ khóa đó trở nên hiệu quả.

Chỉ mục nghịch đảo trong các mô hình IR khác nhau sẽ chứa những dạng giá trị khác nhau. Với mô hình Boolean, giá trị lưu trữ trong chỉ mục nghịch đảo là chân trị của sự xuất hiện từ khóa trong tài liệu, được biểu diễn bằng 0 và 1.

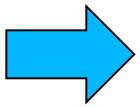
# LẬP CHỈ MỤC

## ❖CHỈ MỤC ĐẢO NGƯỢC

Ví dụ: giả sử có 5 tài liệu $d_i$, i=1..5, trong đó có tất cả 4 từ khóa $t_j$, j=1..4, khi đó, mỗi tài liệu được biểu diễn như sau:

**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 |



**Ma trận từ khóa (Term-doc)**

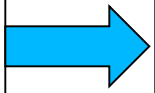
| DOC | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| $t_1$ | 1 | 1 | 1 | 0 | 0 |
| $t_2$ | 1 | 1 | 0 | 0 | 1 |
| $t_3$ | 1 | 0 | 0 | 1 | 0 |
| $t_4$ | 1 | 0 | 0 | 0 | 1 |

# LẬP CHỈ MỤC

## ❖QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

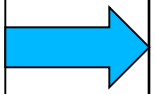
Với mỗi tài liệu, tách các từ khóa để tạo thành danh sách gồm từ khóa và chỉ số tài liệu. Chỉ số tài liệu là thứ tự mà tài liệu đó được xử lý.

**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**



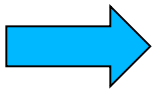
| Từ khóa | Chỉ số tài liệu |
|---|---|
| chỉ | 2 |
| mục | 2 |
| từ | 2 |
| khóa | 2 |

# 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 |



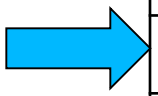
| Từ khóa | Chỉ số tài liệu |
|---|---|
| chỉ | 1 |
| chỉ | 2 |
| khóa | 2 |
| lập | 1 |
| liệu | 1 |
| mục | 1 |
| mục | 1 |
| mục | 2 |
| tài | 1 |
| từ | 2 |

# 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

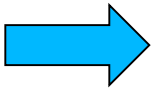
## ❖KHÁI NIỆM

- 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 đòng trong danh sách posting được trỏ 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 | 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**

| Từ khóa | số tài liệu | Tần 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 | Tần số |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 2 | 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

## ❖VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Từ vựng là đơn vị cơ bản cấu tạo thành tài liệu. Vấn 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

## ❖VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Trường hợp từ vựng được xác định là những cụm ký tự phân tách bởi khoảng trắng. Tập từ vựng và danh sách posting như sau:

# TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

- VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

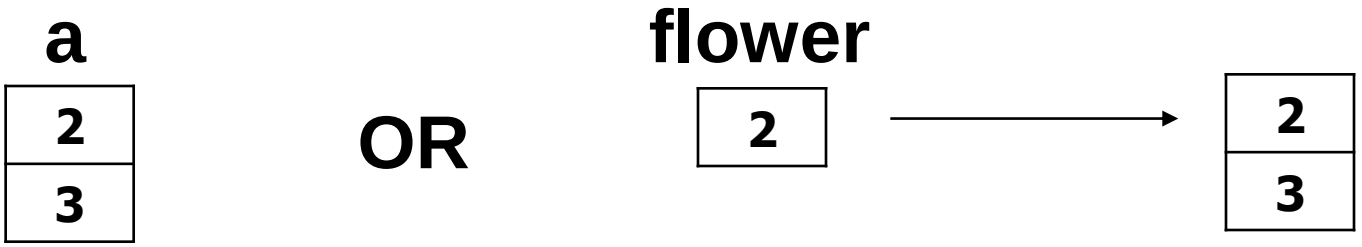
| Từ khóa | số tài liệu | Tần số |
|---|---|---|
| a | 2 | 3 |
| flower | 1 | 1 |
| flowers | 1 | 1 |
| in | 1 | 1 |
| is | 1 | 1 |
| lady | 1 | 1 |
| rose | 2 | 2 |
| sun | 1 | 1 |

| Chỉ số tài liệu | Tần số |
|---|---|
| 2 | 2 |
| 3 | 1 |
| 2 | 1 |
| 1 | 1 |
| 3 | 1 |
| 2 | 1 |
| 3 | 1 |
| 2 | 1 |
| 3 | 1 |
| 1 | 1 |

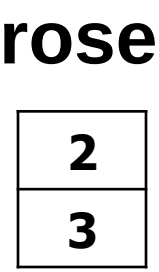
# TẬP TỪ VỰNG VÀ DANH SÁCH POSTING

- VẤN ĐỀ XÁC ĐỊNH TẬP TỪ VỰNG

Truy vấn q1: a flower




Truy vấn q2: rose



# 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

Việc truy vấn chỉ mục được thực hiện khi so khớp tài liệu với truy vấn theo mô hình tương ứng. Quá trình xử lý các phép toán được thực hiện sau khi xác định được tập tài liệu thuộc từ khóa đang xét.

Việc xác định tập tài liệu thuộc từ khóa đang xét được thực hiện theo 2 bước:

- Xác định vị trí từ khóa trên danh sách từ vựng (từ điển)
- Xác định danh sách tài liệu dựa trên liên kết giữa vị trí của từ khóa với danh sách posting với số phần tử được xác định tại mục từ.

# 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?