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

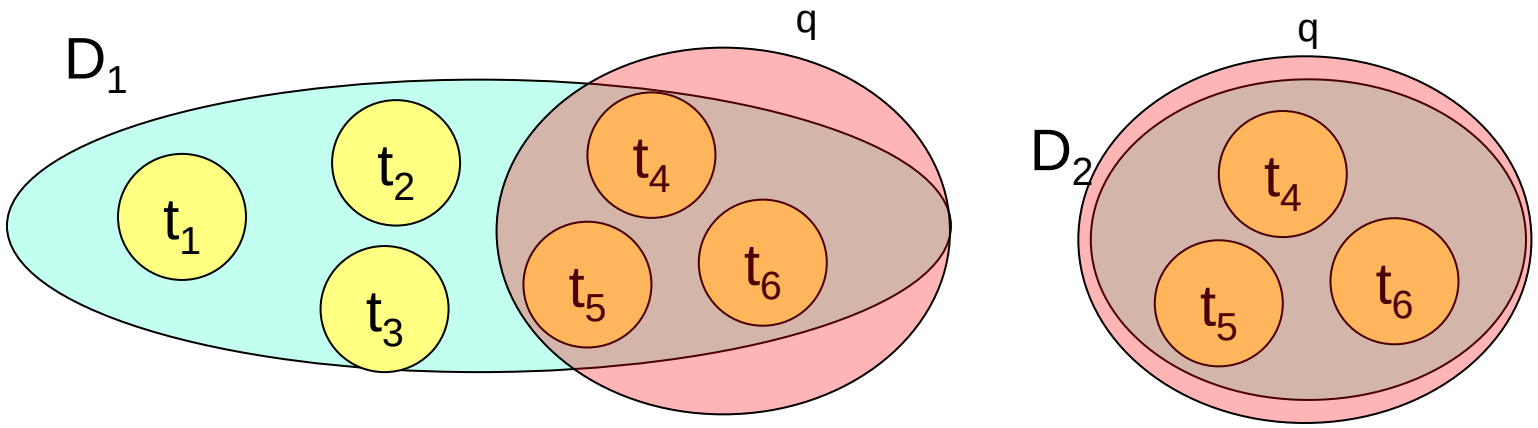
- TRUY XUẤT THÔNG TIN
- CÁC MÔ HÌNH TRUY XUẤT THÔNG TIN CĂN BẢN
- LẬP CHỈ MỤC
- TẬP TỪ VỰNG VÀ DANH SÁCH “POSTING”
- TRUY VẤN CHỈ MỤC

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 chứa trong, truy vấn q={t_1, t_2}

d	1	1	1	1
q	1	1	0	0
$q \cap d = q_t$ x d_t	1	1	0	0
Sim(d_1,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)

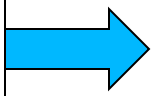
Doc / Term	d_1	d_2	d_3	d_4	d_5
t_1	1	1	1	0	0
t_2	1	1	0	0	1
t_3	1	0	0	1	0
t_4	1	0	0	0	1

LẬP CHỈ MỤC

❖QUÁ TRÌNH XÂY DỰNG CHỈ MỤC

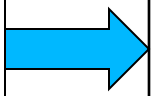
Với mỗi tài liệu, tách các từ khóa để tạo thành danh sách gồm từ khóa và chỉ số tài liệu. Chỉ số tài liệu là thứ tự mà tài liệu đó được xử lý.

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



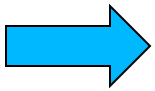
Từ khóa	Chỉ số tài liệu
chỉ	2
mục	2
từ	2
khóa	2

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



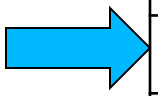
Từ khóa	Chỉ số tài liệu
chỉ	1
chỉ	2
khóa	2
lập	1
liệu	1
mục	1
mục	1
mục	2
tài	1
từ	2

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

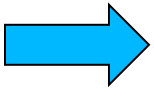
❖KHÁI NIỆM

- 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: sử dụng mô hình tập hợp:

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

d_1 = {sun, flowers} d_2 = {a, rose, is, flower}

d_3 = {a, lady, in, rose}

q_1 = {a, flower} q_2 = {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 = {sun, flowers} d_2 = {a, rose, is, flower}

d_3 = {a, lady, in, rose}

q_1 = {a, flower} q_2 = {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

❖VẤN 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 = {sun, flower} d_2 = {a, rose, be, flower}

d_3 = {a, lady, in, rose}

q_1 = {a, flower} q_2 = {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:

Từ khóa	số tài liệu
a	2
flower	1
flowers	1
in	1
is	1
lady	1
rose	2
sun	1

Chỉ số tài liệu
2
3
2
1
3
2
3
2
3
1

TRUY VẤN CHỈ MỤC

❖MÔ HÌNH TẬP HỢP

Ví dụ:

truy vấn q1 = {a, flower}, R = {}

- Truy vấn "a" trên chỉ mục:

R → {}
"a" → {2,3}

R → {(2,1), (3,1)}

- Truy vấn "flower" trên chỉ mục:

R → {(2,1), (3,1)}
“flower" → {2}

R → {(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)

Từ khóa	số tài liệu
a	2
flower	1
flowers	1
in	1
is	1
lady	1
rose	2
sun	1

Chỉ số tài liệu
2
3
2
1
3
2
3
2
3
1

TRUY VẤN CHỈ MỤC

❖MÔ HÌNH BOOLEAN

Ví dụ: truy vấn:

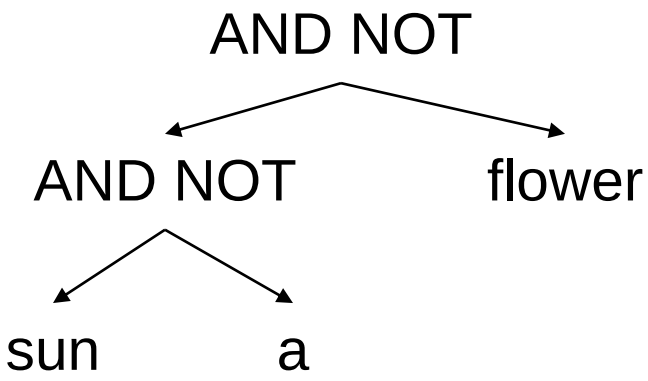
sun AND NOT (a OR flower)

- Chuyển về dạng OR:

sun AND (NOT a AND NOT flower)

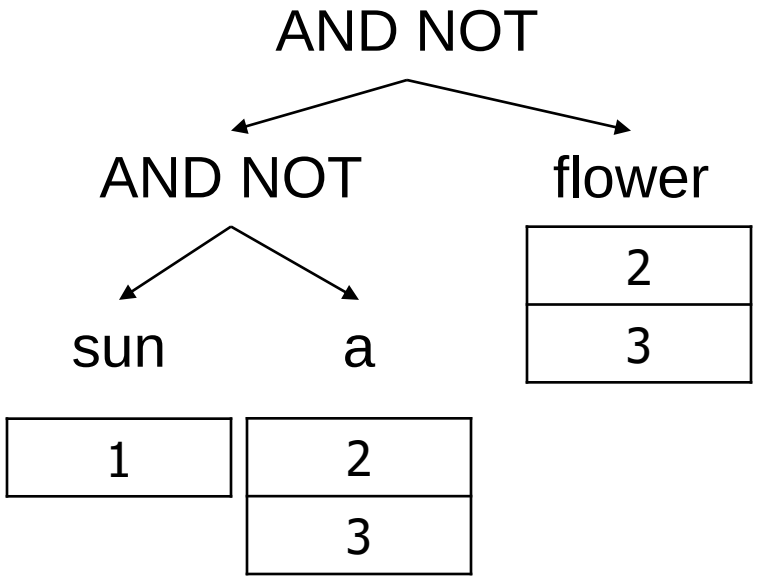
= sun AND NOT a AND NOT flower

- Chuyển thành dạng cây



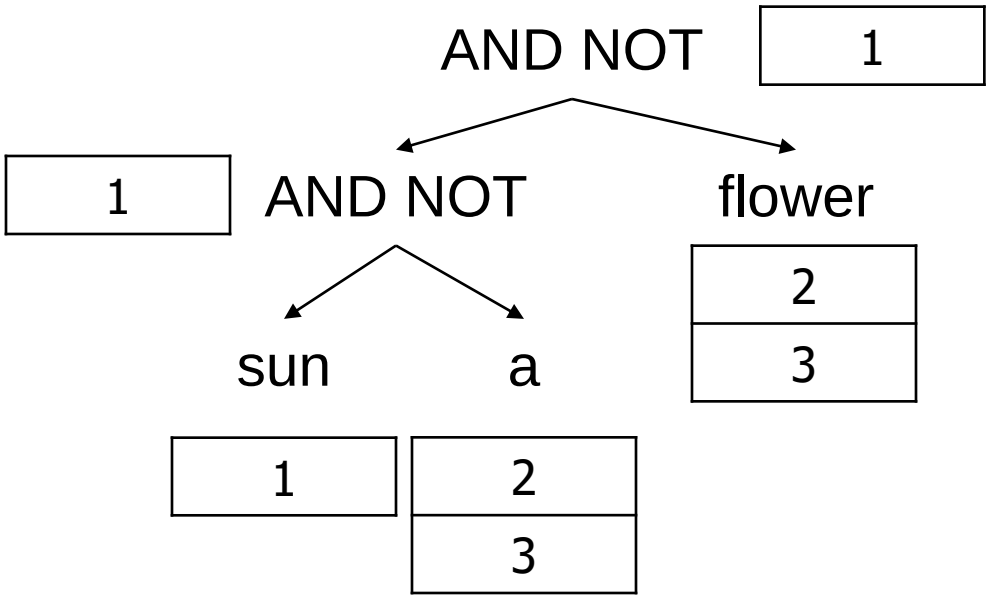
TRUY VẤN CHỈ MỤC

- 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?