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 Chỉnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

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

VẤN 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 tại 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 thưa)

  + 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 dung 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

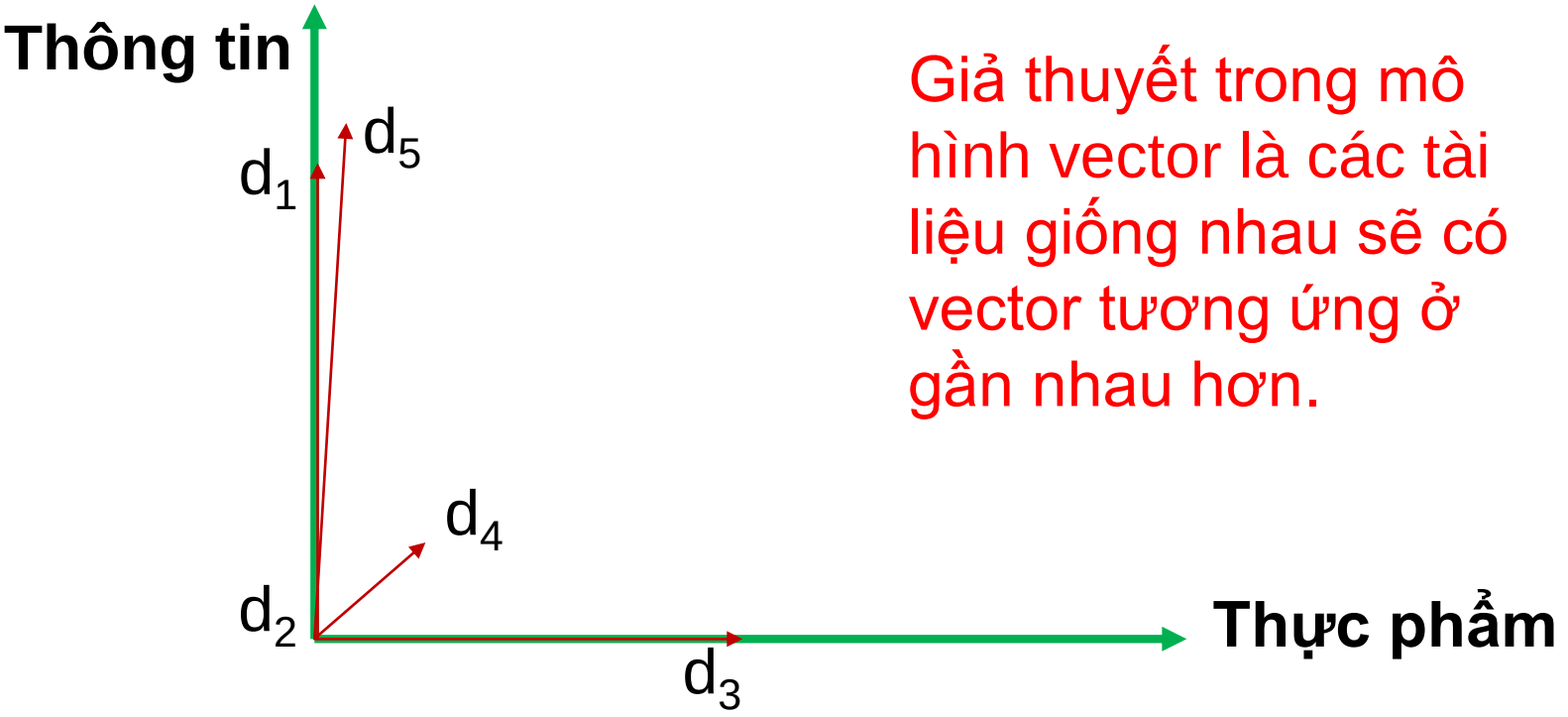
- Với cách tính trọng số là tần số xuất hiện của từ khóa trong tài liệu thì vector d1 cho biết:

+ 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":



Giả thuyết trong mô hình vector là các tài liệu giống nhau sẽ có vector tương ứng ở gần nhau hơn.

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.

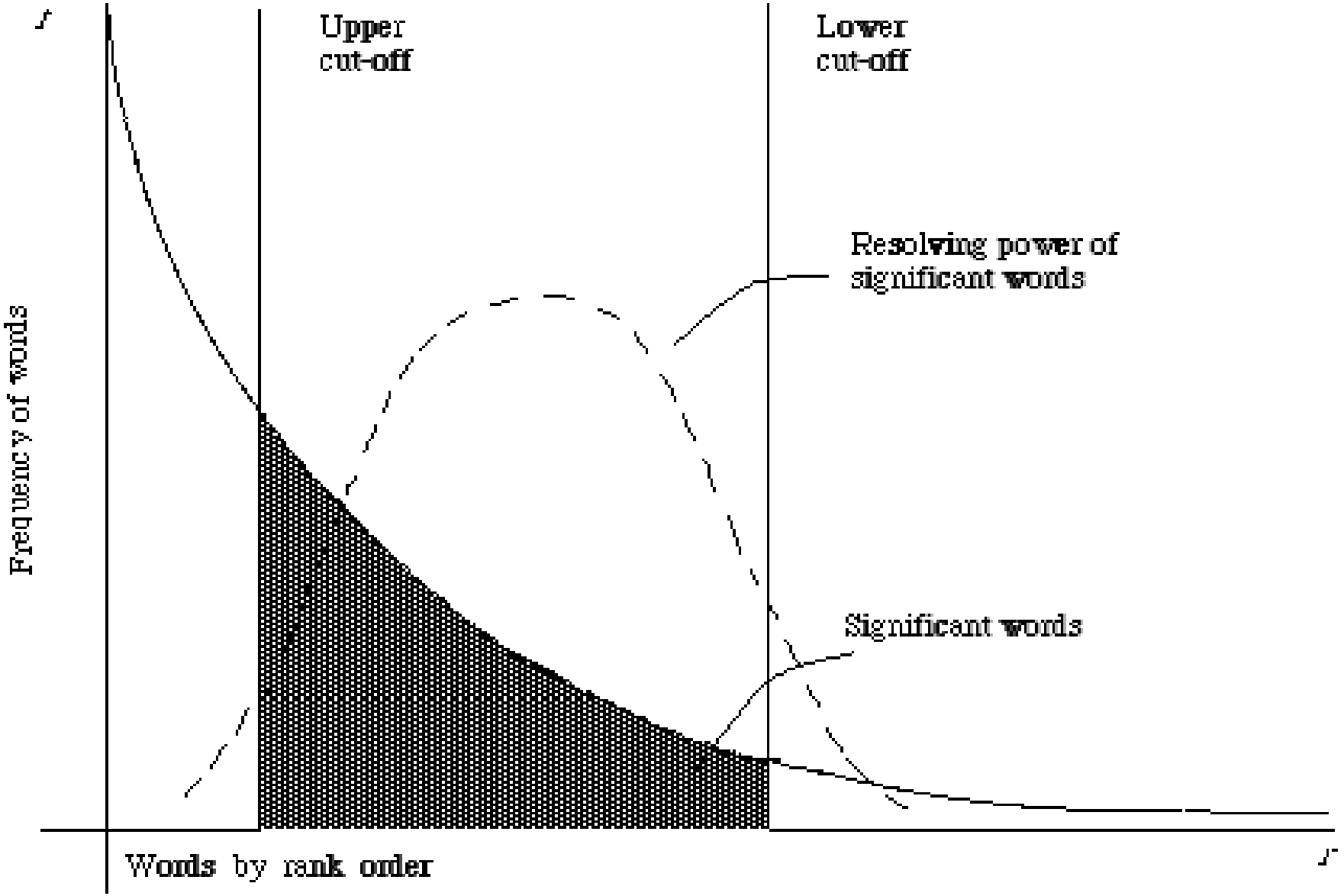
Ma trận tài liệu (Doc-Term)

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



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:**

$$freq_{kd} = \left\{ \begin{array}{c} \{0,1\} \\ \dfrac{freq_{kd}}{\max(freq_{kd})} \\ \dfrac{1}{2} + \dfrac{1}{2}\dfrac{freq_{kd}}{\max(freq_{kd})} \\ \ln(freq_{kd}) + 1 \end{array} \right\}$$

Nhị phân

Chuẩn max

Tăng cường

Logarith

TRỌNG SỐ CỦA TERM

❖CHUẨN HÓA TRỌNG SỐ

*** Trọng số theo hệ thống SMART:**

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

Nghịch đảo

Bình phương

Xác suất

Tần suất

TRỌNG SỐ CỦA TERM

❖CHUẨN HÓA TRỌNG SỐ

*** Trọng số theo hệ thống SMART:**

$$norm = \left\{ \begin{array}{l} \sum_{vector} w_j \\ \sqrt{\sum_{vector} w_j^2} \\ \sum_{vector} w_j^4 \\ \max_{vector} \left(w_j \right) \end{array} \right\}$$

Chuẩn tổng

Chuẩn cosine

Chuẩn Bậc bốn

Chuẩn max

LẬP CHỈ MỤC

❖VẤN ĐỀ

- Lưu trữ các vector có số chiều rất lớn
- Có nhiều trường hợp là dạng vector thưa.
- 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

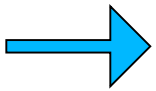
Từ khóa	Chỉ số tài liệu
chỉ	2
mục	2
từ	2
khóa	2

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



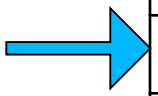
Từ khóa	Chỉ số tài liệu
chỉ	1
chỉ	2
khóa	2
lập	1
liệu	1
mục	1
mục	1
mục	2
tài	1
từ	2

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

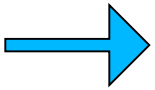


Từ khóa	Chỉ số tài liệu	Tần số
chỉ	1	1
chỉ	2	1
khóa	2	1
lập	1	1
liệu	1	1
mục	1	2
mục	2	1
tài	1	1
từ	2	1

LẬP CHỈ MỤC

❖QUÁ TRÌNH LẬP CHỈ MỤC

Từ khóa	Chỉ số tài liệu	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

LẬP CHỈ MỤC

❖QUÁ TRÌNH LẬP CHỈ MỤC

Tập từ vựng

Từ khóa	số tài liệu	Tần 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	Tần 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
$\mathbf{d_1}$	8	20	2	0
$\mathbf{d_2}$	10	0	31	0
$\mathbf{d_3}$	1	0	42	14
$\mathbf{d_4}$	0	3	0	3
$\mathbf{d_5}$	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
wood tree

4)

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 = wood

2) Q_2 =