



**ĐẠI HỌC QUỐC GIA TP HCM  
TRƯỜNG ĐẠI HỌC  
CÔNG NGHỆ THÔNG TIN**

# **TRUY XUẤT THÔNG TIN**

## **CHƯƠNG II – MÔ HÌNH KHÔNG GIAN VECTOR**

Nguyễn Trọng Chính  
chinhnt@uit.edu.vn

# **NỘI DUNG TRÌNH BÀY**

- ❖ **VẤN ĐỀ TRONG MÔ HÌNH BOOLEAN**
- ❖ **MÔ HÌNH VECTOR**
- ❖ **TRỌNG SỐ CỦA TERM**
- ❖ **TÍNH VECTOR SCORE**
- ❖ **TRUY VẤN TRÊN MÔ HÌNH KHÔNG GIAN VECTOR**

# TRUY XUẤT THÔNG TIN

## ❖ VẤN ĐỀ TRONG MÔ HÌNH BOOLEAN

- 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 <sub>1</sub>	t <sub>2</sub>	t <sub>3</sub>	t <sub>4</sub>
d <sub>1</sub>	1	1	1	1
d <sub>2</sub>	1	1	0	0
d <sub>3</sub>	1	0	0	0
d <sub>4</sub>	0	0	1	0
d <sub>5</sub>	0	1	0	1

# TRUY XUẤT THÔNG TIN

## ❖ VẤN ĐỀ TRONG MÔ HÌNH BOOLEAN

- 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

# TRUY XUẤT THÔNG TIN

## ❖ 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 thưa)
  - + Tính toán dựa trên hướng và độ lớn.

# TRUY XUẤT THÔNG TIN

## ❖ MÔ HÌNH VECTOR

- 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.

# TRUY XUẤT THÔNG TIN

## ❖ MÔ HÌNH VECTOR

- 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

# TRUY XUẤT THÔNG TIN

## ❖ MÔ HÌNH VECTOR

### Ma trận tài liệu (Doc-Term)

DOC	truy xuất	thông tin	Công nghệ	Thực phẩm
$d_1$	8	20	2	0

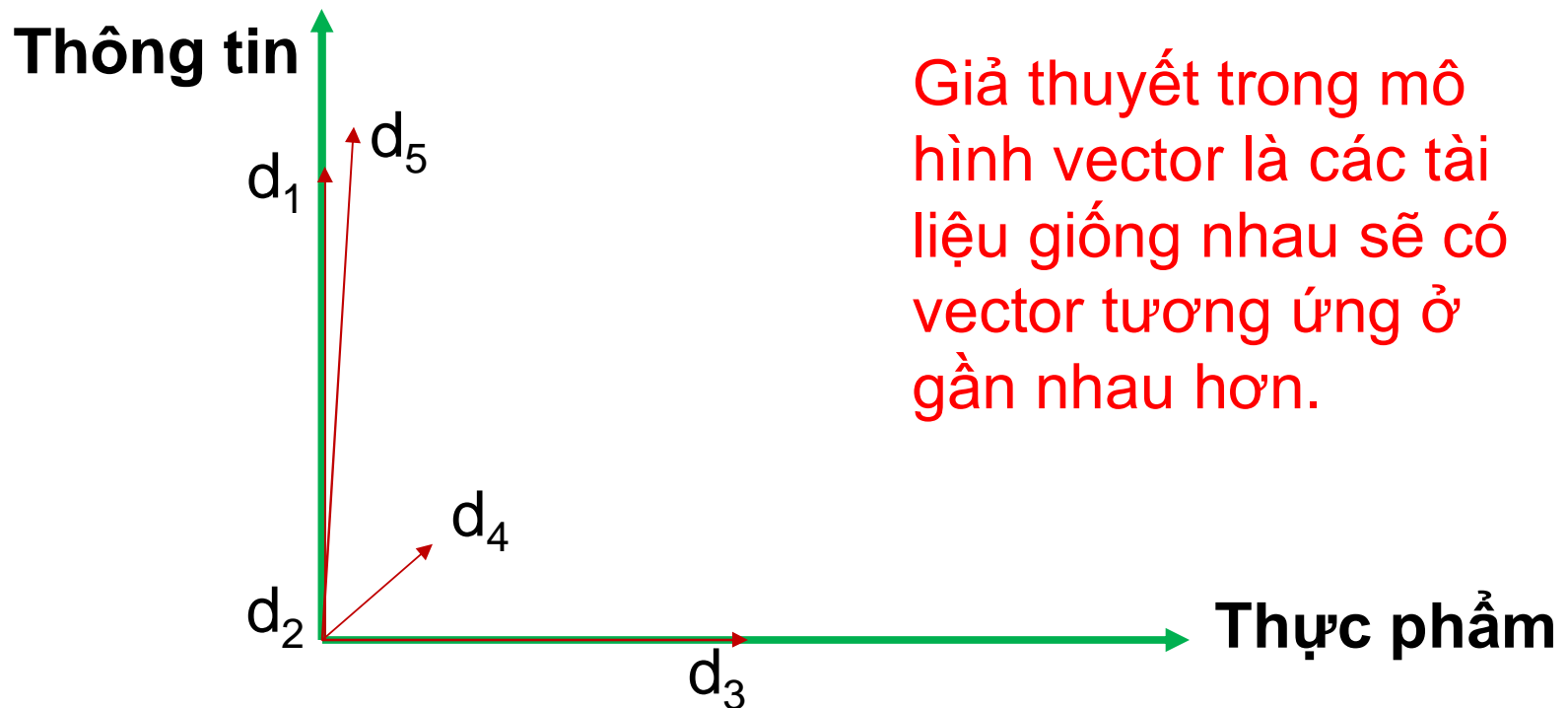
- Với cách tính trọng số là tần số xuất hiện của từ khóa trong tài liệu thì vector  $d_1$  cho biết:
  - + Từ “truy xuất” xuất hiện 8 lần
  - + Từ “thông tin” xuất hiện 20 lần
  - + Từ “công nghệ” xuất hiện 2 lần
  - + Từ “thực phẩm” không xuất hiện.



# TRUY XUẤT THÔNG TIN

## ❖ MÔ HÌNH VECTOR

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



# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

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.

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

\* 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

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

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

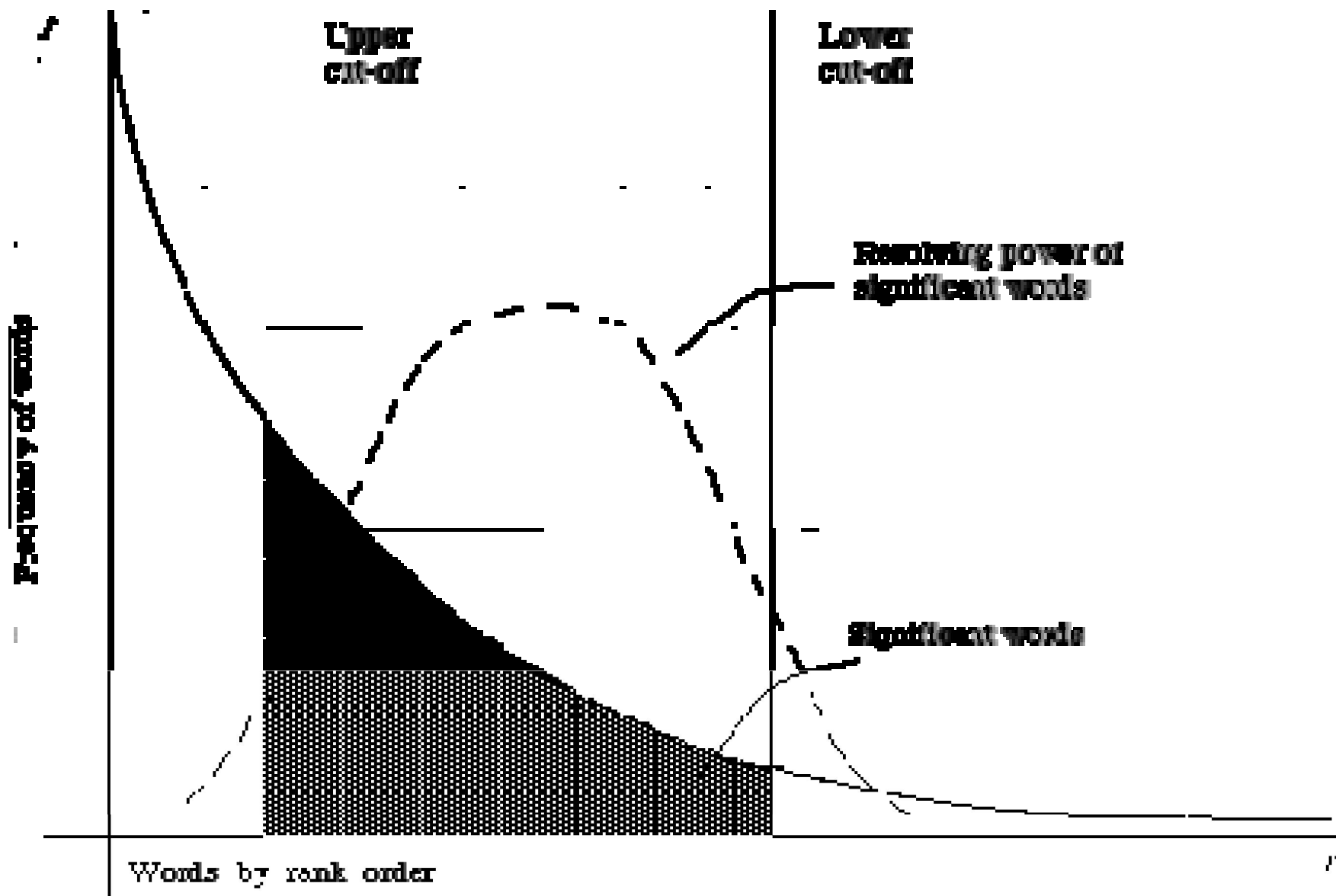
**Ma trận tài liệu (Doc-Term)**

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

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

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**

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

\* 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.

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

### \* 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.

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



# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

### \* 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

$$\text{idf}_{\text{truy xuất}} = \log(1000/38) = 1.42$$

$$\text{idf}_{\text{thông tin}} = \log(1000/200) = 0.7$$

$$\text{idf}_{\text{công nghệ}} = \log(1000/102) = 0.99$$

$$\text{idf}_{\text{thực phẩm}} = \log(1000/11) = 1.96$$

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

### \* 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

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

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

### \* 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:

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

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

\* Trọng số theo hệ thống SMART:

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

*Nhị phân*

*Chuẩn max*

*Tăng cường*

*Logarith*

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

\* Trọng số theo hệ thống SMART:

$$collect_k = \left\{ \begin{array}{l} \log \frac{NDoc}{Doc_k} \\ \left( \log \frac{NDoc}{Doc_k} \right)^2 \\ \log \frac{NDoc - Doc_k}{Doc_k} \\ \frac{1}{Doc_k} \end{array} \right\} \begin{array}{l} \text{Nghịch đảo} \\ \text{Bình phương} \\ \text{Xác suất} \\ \text{Tần suất} \end{array}$$

# TRUY XUẤT THÔNG TIN

## ❖ TRỌNG SỐ CỦA TERM

\* Trọng số theo hệ thống SMART:

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

***Chuẩn tổng***

***Chuẩn cosine***

***Chuẩn Bậc bốn***

***Chuẩn max***

# TRUY XUẤT THÔNG TIN

## ❖ TÍNH VECTOR SCORE

Phản ánh **khoảng cách** giữa hai vector, cho biết độ tương đồng giữa hai tài liệu hoặc tài liệu và truy vấn.

- Vector score giữa hai tài liệu  $D_i$  và  $D_j$  được tính theo chuẩn tích vô hướng (cosine) như sau:

$$\text{sim}(D_i, D_j) = \sum_{k=1}^t w_{ik} * w_{jk}$$

+ k là chiều của term trong không gian.

Các trọng số đã được chuẩn hóa trước.

# TRUY XUẤT THÔNG TIN

## ❖ TÍNH VECTOR SCORE

- Vector score giữa tài liệu  $D_i$  và truy vấn  $Q$  được tính theo chuẩn tích vô hướng (cosine) như sau:  
+ Trường hợp trọng số term của  $Q$  đã được chuẩn hóa:

$$\text{sim}(Q, D_i) = \sum_{j=1}^t w_{qj} * w_{d_{ij}}$$



# TRUY XUẤT THÔNG TIN

## ❖ TÍNH VECTOR SCORE

+ Trường hợp trọng số term của  $Q$  và  $D_i$  chưa được chuẩn hóa

$$sim(Q, D_i) = \frac{\sum_{j=1}^t w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^t (w_{qj})^2 * \sum_{j=1}^t (w_{d_{ij}})^2}}$$

# TRUY XUẤT THÔNG TIN

## ❖ TRUY VẤN TRÊN MÔ HÌNH VECTOR

### \* 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 XUẤT THÔNG TIN

## ❖ TRUY VẤN TRÊN MÔ HÌNH VECTOR

### \* 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)**

<b>DOC</b>	<b>truy xuất</b>	<b>thông tin</b>	<b>Công nghệ</b>	<b>Thực phẩm</b>
<b>d<sub>1</sub></b>	8	20	2	0
<b>d<sub>2</sub></b>	10	0	31	0
<b>d<sub>3</sub></b>	1	0	42	14
<b>d<sub>4</sub></b>	0	3	0	3
<b>d<sub>5</sub></b>	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)  $Q_2 = \text{ship wood}$