



# **KHAI THÁC TẬP KHÔNG HỮU ÍCH**

## **(MINING ERASABLE ITEMSETS)**

# Nội dung

---

- Giới thiệu
- Bài toán khai thác EI
- Thuật toán META
- Tổng kết

# Giới thiệu



*Sản xuất các loại sản phẩm...*

$P_1 (i_2, i_3, i_4, i_6)$

*Lợi nhuận: 20 triệu*

$P_2 (i_2, i_5, i_7)$

*Lợi nhuận: 50 triệu*

$P_3 (i_1, i_2, i_3, i_5)$

*Lợi nhuận: 30 triệu*

*Tổng lợi nhuận khi  
bán toàn bộ sản phẩm  
100 triệu*

*Nguyên liệu  
 $i_1, i_2, i_3, i_4, i_5, i_6, i_7$*

# Giới thiệu



Công ty không có đủ tiền mua nguyên liệu...

Công ty phải ngừng sản xuất một số loại sản phẩm và không mua những nguyên vật liệu tương ứng...



***Ngừng sản xuất những loại sản phẩm nào?***

*Những loại sản phẩm mà không làm giảm tổng lợi nhuận quá một **ngưỡng** nào đó...*

# Giới thiệu

## Ví dụ:

Với ngưỡng giảm lợi nhuận chấp nhận được là 25%, công ty có thể bỏ loại sản phẩm  $P_1$  và không mua các nguyên liệu  $i_4$  và  $i_6$ .

Sản phẩm	Lợi nhuận
$P_1 (i_2, i_3, i_4, i_6)$	20
$P_2 (i_2, i_5, i_7)$	50
$P_3 (i_1, i_2, i_3, i_5)$	30

*$\{i_4, i_6\}$  gọi là một tập thành phần không hữu ích*

Vấn đề trở thành tìm  
những *itemsets* như vậy...

# Nội dung

---

- Giới thiệu
- Bài toán khai thác EI
- Thuật toán META
- Tổng kết

# Bài toán khai thác EI

- Dữ liệu
  - Ngưỡng  $\xi$
  - Tập thành phần  $I = \{i_1, i_2, \dots, i_m\}$
  - Cơ sở dữ liệu  $DB = \{P_1, P_2, \dots, P_n\}$

Cơ sở dữ liệu thí dụ  $DB_e$  gồm 6 loại sản phẩm và 7 thành phần, tổng lợi nhuận là 10000.

<i>Product</i>	<i>PID</i>	<i>Items</i>	<i>Val</i>
$P_1$	1	$\{i_2, i_3, i_4, i_6\}$	500
$P_2$	2	$\{i_2, i_5, i_7\}$	200
$P_3$	3	$\{i_1, i_2, i_3, i_5\}$	500
$P_4$	4	$\{i_1, i_2, i_4\}$	8000
$P_5$	5	$\{i_6, i_7\}$	300
$P_6$	6	$\{i_3, i_4\}$	500

# Bài toán khai thác EI

- Định nghĩa 1 (*Gain*)

Cho itemset  $A (\subseteq I)$ , *Gain* của  $A$  được tính như sau:

$$Gain(A) = \sum_{\{P_k | A \cap P_k.Items \neq \emptyset\}} P_k.Val$$

Ví dụ:

$A = \{i_6, i_7\}$ , các loại sản phẩm có chứa  $i_6$

hoặc  $i_7$  hay cả hai là  $P_1, P_2, P_5$ , do đó:

$$\begin{aligned} Gain(A) &= P_1.Val + P_2.Val + P_5.Val \\ &= 500 + 200 + 300 = 1000 \end{aligned}$$

$P_i$	PID	Items	Val
$P_1$	1	$\{i_2, i_3, i_4, i_6\}$	500
$P_2$	2	$\{i_2, i_5, i_7\}$	200
$P_3$	3	$\{i_1, i_2, i_3, i_5\}$	500
$P_4$	4	$\{i_1, i_2, i_4\}$	8000
$P_5$	5	$\{i_6, i_7\}$	300
$P_6$	6	$\{i_3, i_4\}$	500



# Bài toán khai thác EI

- Định nghĩa 2 ( $EI$ )

Cho trước một ngưỡng  $\xi$  và tập  $DB$ , một tập  $A$  gọi là *tập không hữu ích* nếu:

$$Gain(A) \leq \left( \sum_{P_k \in DB} P_k.Val \right) \times \xi$$

Ví dụ:

Tổng lợi nhuận 10000,  $\xi = 15\%$

$$\begin{aligned} Gain(\{i_6, i_7\}) &= 50 + 20 + 30 \\ &= 100 \leq (10000 \times 15\%) \end{aligned}$$

Do đó  $\{i_6, i_7\}$  là một tập không hữu ích.


$P_i$	PID	Items	Val
$P_1$	1	$\{i_2, i_3, i_4, i_6\}$	500
$P_2$	2	$\{i_2, i_5, i_7\}$	200
$P_3$	3	$\{i_1, i_2, i_3, i_5\}$	500
$P_4$	4	$\{i_1, i_2, i_4\}$	8000
$P_5$	5	$\{i_6, i_7\}$	300
$P_6$	6	$\{i_3, i_4\}$	500

# Bài toán khai thác EI

- Phát biểu bài toán: Cho cơ sở dữ liệu sản phẩm  $DB$  và một ngưỡng  $\xi$ , hãy tìm tất cả các tập không hữu ích trong  $DB$ .

$PID$	$Items$	$Val$
1	$\{i_2, i_3, i_4, i_6\}$	500
2	$\{i_2, i_5, i_7\}$	200
3	$\{i_1, i_2, i_3, i_5\}$	500
4	$\{i_1, i_2, i_4\}$	8000
5	$\{i_6, i_7\}$	300
6	$\{i_3, i_4\}$	500

$\xi = 15\%$



$Itemset$	$Gain$
$\{i_3\}$	1500
$\{i_5\}$	700
$\{i_6\}$	800
$\{i_7\}$	500
$\{i_5, i_6\}$	1500
$\{i_5, i_7\}$	1000
$\{i_6, i_7\}$	1000
$\{i_5, i_6, i_7\}$	1500

$EIs$

# Nội dung

---

- Giới thiệu
- Bài toán khai thác EI
- Thuật toán META
- Tổng kết

# Thuật toán META

- **META** Mining Erasable iTemsets with the Antimonotone property algorithm

Thuật toán đầu tiên khai thác *EIs* được nhóm tác giả Zhi-Hong Deng giới thiệu vào năm 2009.

- Các tính chất

**Tính chất 1:** Cho hai tập  $X \subseteq I$  và  $Y \subseteq I$ . Nếu  $X$  là tập con của  $Y$  ( $X \subseteq Y$ ) thì  $Gain(X) \leq Gain(Y)$ .

**Tính chất 2** (*anti-monotone*): Cho hai tập  $X \subseteq I$  và  $Y \subseteq I$ . Nếu  $X$  không phải là một *EI* và  $X \subseteq Y$  thì  $Y$  cũng không phải là một *EI*.

**Tính chất 3:** Nếu  $X$  là một *EI* và  $Y$  là tập con của  $X$  ( $Y \subseteq X$ ) thì  $Y$  phải là một *EI*.

# Thuật toán META

- Phương pháp: Tìm các *EI*s theo từng cấp độ (*level-wise search*)

**Input:** Cơ sở dữ liệu  $DB = \{P_1, P_2, \dots, P_n\}$ ; ngưỡng  $\xi$ ;

**Output:** Tập toàn bộ các tập không hữu ích *EI*;

$Sum\_val = 0$ ;

For ( $k = 1$ ;  $k \leq n$ ;  $k++$ )

$Sum\_val = Sum\_val + P_k.Val$ ;

$E_1 = \{EI_1\}$ ;

For ( $k = 2$ ;  $E_{k-1} \neq \emptyset$ ;  $k++$ )

$GC_k = \mathbf{Gen\_Candidate}(E_{k-1})$ ;

For each product  $P \in DB$  {

For each candidate itemset  $C \in GC_k$

If ( $C \cap P \neq \emptyset$ ) then

$C.gain = C.value + P.Val$ ;

$E_k = \{C \in GC_k \mid C.gain \leq \xi \times Sum\_val\}$

Return  $EI = \cup_k E_k$ ;

# Thuật toán META

## **Procedure Gen\_Candidate ( $E_{k-1}$ )**

*// Các items trong  $E_{k-1}$  được sắp theo thứ tự xuất hiện trong  $I$*

$Candidates = \emptyset$ ;

For each  $A_1 (= \{x_1, x_2, \dots, x_{k-2}, x_{k-1}\}) \in E_{k-1}$

For each  $A_2 (= \{y_1, y_2, \dots, y_{k-2}, y_{k-1}\}) \in E_{k-1}$

If  $((x_1=y_1) \wedge (x_2=y_2) \wedge \dots \wedge (x_{k-2}=y_{k-2}) \wedge (x_{k-1} < y_{k-1}))$  then

$X = \{x_1, x_2, \dots, x_{k-2}, x_{k-1}, y_{k-1}\}$ ;

If **No\_Unerasable\_Subset** ( $X, E_{k-1}$ ) then

add  $X$  to  $Candidates$ ;

Return  $Candidates$ ;

## **Procedure No\_Unerasable\_Subset ( $X, E_{k-1}$ )**

For each  $(k-1)$ -subset  $X_s$  of  $X$

If  $X_s \notin E_{k-1}$  then

Return FALSE;     *//anti-monotone*

Return TRUE;

# Thuật toán META

- Minh họa

$$I = \{i_1, i_2, i_3, i_4, i_5, i_6, i_7\}$$

$$DB = \{P_1, P_2, P_3, P_4, P_5, P_6\}$$

$$\xi = 18\%$$

<i>PID</i>	<i>Items</i>	<i>Val</i>
1	$\{i_2, i_3, i_4, i_6\}$	500
2	$\{i_2, i_5, i_7\}$	200
3	$\{i_1, i_2, i_3, i_5\}$	500
4	$\{i_1, i_2, i_4\}$	8000
5	$\{i_6, i_7\}$	300
6	$\{i_3, i_4\}$	500

## Bước 1: Khởi tạo

- Tính  $Sum\_val = 10000$
- Tìm  $E_1$

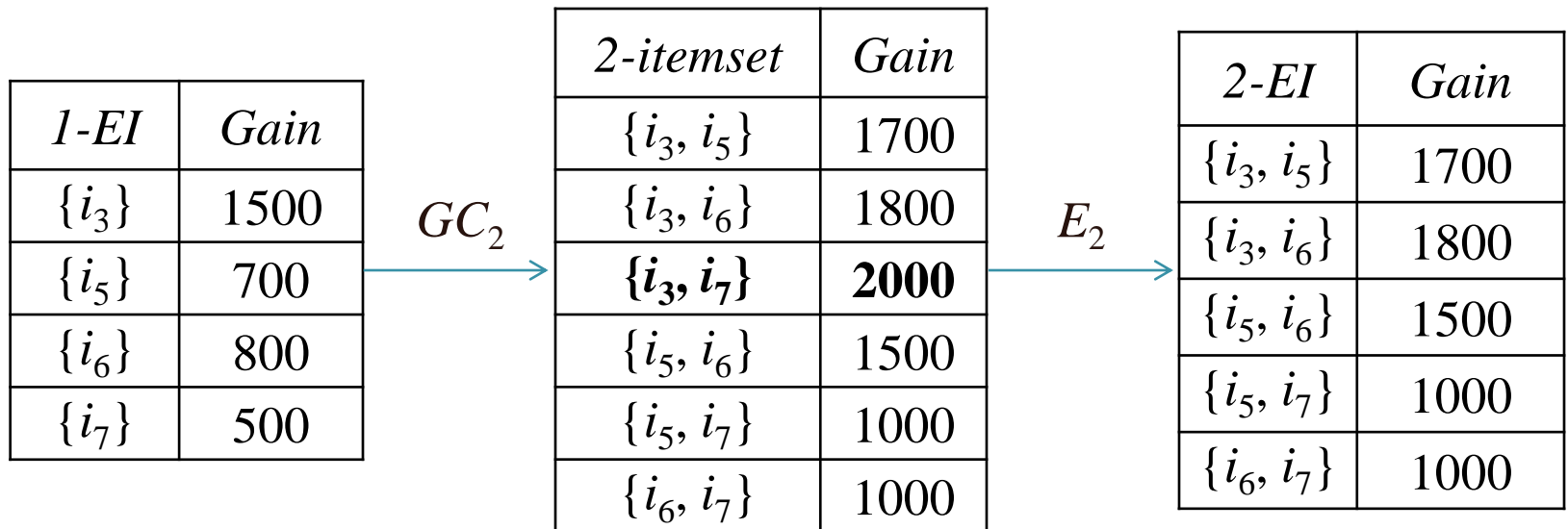
$E_1$	<i>1-itemset</i>	<i>Gain</i>
	$\{i_3\}$	1500
	$\{i_5\}$	700
	$\{i_6\}$	800
	$\{i_7\}$	500

# Thuật toán META

## Bước 2: Khai thác

$$\underline{k = 2}$$

- Xây dựng  $GC_2$  dựa trên  $E_1$
- Tìm tập  $E_2$  dựa trên  $GC_2$



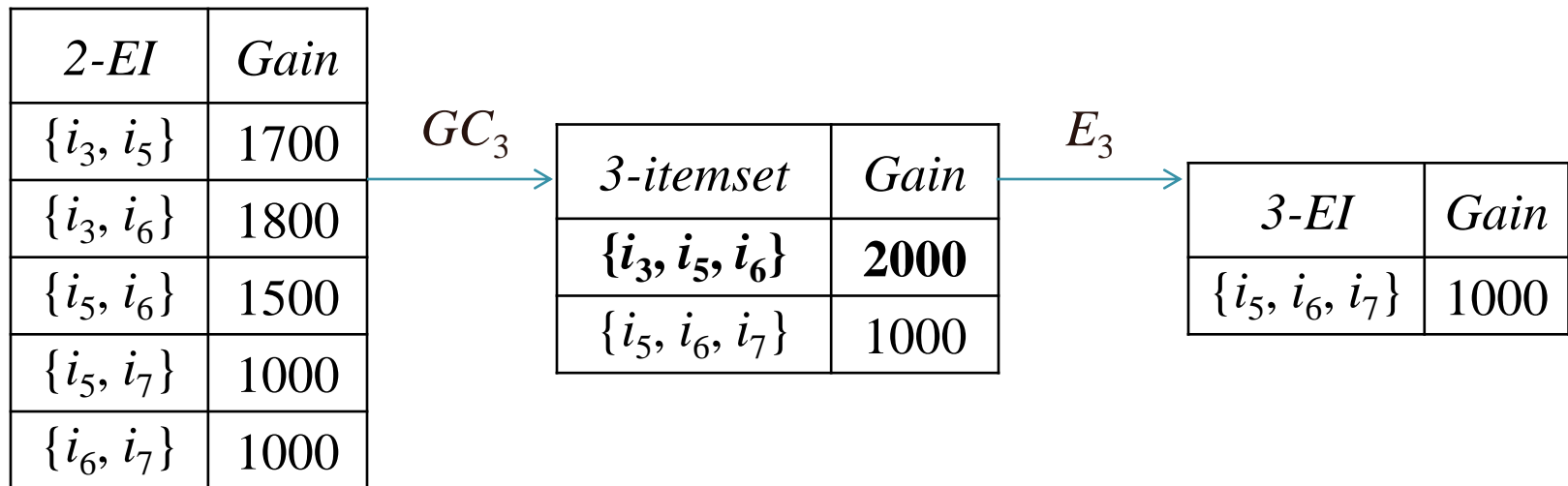


# Thuật toán META

## Bước 2: Khai thác

$$\underline{k = 3}$$

- Xây dựng  $GC_3$  dựa trên  $E_2$
- Tìm tập  $E_3$  dựa trên  $GC_3$



# Thuật toán META

---

Bước 2: Khai thác

$$\underline{k = 4}$$

$E_4 = \emptyset \rightarrow$  Thuật toán dừng

Bước 3: Trả về kết quả

$$EI = E_1 \cup E_2 \cup E_3$$

# Thuật toán META

- Nhận xét

- Đầu tiên, thuật toán duyệt *DB* tính tổng lợi nhuận. Trong  $k$  bước lặp tiếp theo, thuật toán tiếp tục duyệt *DB* để tính lợi nhuận của các itemset. Do đó chi phí thời gian rất lớn.

- Thuật toán không loại bỏ được dữ liệu dư thừa. Ví dụ xét itemset  $\{i_3\}$ , các loại sản phẩm chứa  $i_3$  là  $P_1, P_3$  và  $P_6$ , nhưng khi tính *Gain* của  $\{i_3\}$  phải duyệt toàn bộ *DB*.

<i>PID</i>	<i>Items</i>	<i>Val</i>
1	$\{i_2, i_3, i_4, i_6\}$	500
2	$\{i_2, i_5, i_7\}$	200
3	$\{i_1, i_2, i_3, i_5\}$	500
4	$\{i_1, i_2, i_4\}$	8000
5	$\{i_6, i_7\}$	300
6	$\{i_3, i_4\}$	500

# Nội dung

---

- Giới thiệu
- Bài toán khai thác EI
- Thuật toán META
- Tổng kết

# Tổng kết

- Khai thác các tập thành phần không hữu ích là một trong những tác vụ mới trong khai thác dữ liệu.
- Về mặt kỹ thuật, khai thác *EI* cũng tương tự khai thác mẫu phổ biến *FP*. Cả hai cùng khai thác các itemset quan tâm.
- Tuy nhiên, khai thác *EI* và khai thác *FP* có sự khác biệt.
  - Khai thác *FP* ra đời trong bối cảnh một siêu thị bán lẻ muốn tìm mối quan hệ giữa các mặt hàng được khách hàng mua.
  - Khai thác *EI* xuất hiện trong bối cảnh công ty sản xuất sản phẩm cần lập kế hoạch sản xuất phù hợp khi nền kinh tế rơi vào suy thoái.