ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC

BTLT TUẦN 7: KHAI THÁC DỮ LIỆU

GVHD: NGUYỄN THANH BÌNH

Sinh viên thực hiện: Nguyễn Công Hoài Nam
Mã số sinh viên: 21280099

Ngày 27 tháng 4 năm 2024

6. *Represent the transaction database of Exercise (5) in vertical format.*

Dữ liệu bài (5) dưới dạng vertical:

items	tid
a	1, 2
b	3, 4, 5
c	1, 3, 6, 7
d	1, 2, 3, 4, 6, 8
e	1, 2, 3, 4, 5, 6, 7, 8
f	2, 3, 4, 5, 7, 8

7. *Determine the confidence of the rules {a} ⇒ {f}, and {a, e} ⇒ {f} for the transaction database in Exercise (1).*

Với X, Y là hai set của items, ta có công thức cho độ đo "độ tin cậy" hay "confidence" là:

$$conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$$

Suy ra:

$$conf(\{a\} \Rightarrow \{f\}) = \frac{sup(\{a\} \cup \{f\})}{sup(\{a\})} = \frac{sup(\{a, f\})}{sup(\{a\})}$$
$$conf(\{a, e\} \Rightarrow \{f\}) = \frac{sup(\{a, e\} \cup \{f\})}{sup(\{a, e\})} = \frac{sup(\{a, e, f\})}{sup(\{a, e\})}$$

Nhắc lại cơ sở dữ liệu cho bài (1)

tid	items
1	a, b, c, d
2	b, c, e, f
3	a, d, e, f
4	a, e, f
5	b, d, f

Ta tính được độ tin cậy của {a} ⇒ {f}

$$\left.\begin{array}{l} sup(\{a, f\}) = 2 \text{ (tid 3, 4)} \\ sup(\{a\}) = 3 \text{ (tid 1, 3, 4)} \end{array}\right\} \Rightarrow conf(\{a\} \Rightarrow \{f\}) = \frac{2}{3} = 0.667$$

Và độ tin cậy của {a,e} ⇒ {f}

$$\left.\begin{array}{l} sup(\{a, e, f\}) = 2 \text{ (tid 3, 4)} \\ sup(\{a, e\}) = 2 \text{ (tid 3, 4)} \end{array}\right\} \Rightarrow conf(\{a, e\} \Rightarrow \{f\}) = \frac{2}{2} = 1$$

8. *Determine the confidence of the rules {a} ⇒ {f}, and {a, e} ⇒ {f} for the transaction database in Exercise (5).*

Cơ sở dữ liệu cho bài (5)

tid	items
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

Tương tự bài (7), ta có

$$conf(\{a\} \Rightarrow \{f\}) = \frac{sup(\{a\} \cup \{f\})}{sup(\{a\})} = \frac{sup(\{a, f\})}{sup(\{a\})}$$
$$conf(\{a, e\} \Rightarrow \{f\}) = \frac{sup(\{a, e\} \cup \{f\})}{sup(\{a, e\})} = \frac{sup(\{a, e, f\})}{sup(\{a, e\})}$$

Ta tính được độ tin cậy của {a} ⇒ {f}

$$\left.\begin{array}{l} sup(\{a,f\}) = 1 \text{ (tid 2)} \\ sup(\{a\}) = 2 \text{ (tid 1, 2)} \end{array}\right\} \Rightarrow conf(\{a\} \Rightarrow \{f\}) = \frac{1}{2} = 0.5$$

Và độ tin cậy của {a,e} ⇒ {f}

$$\left.\begin{array}{l} sup(\{a,e,f\}) = 1 \text{ (tid 2)} \\ sup(\{a,e\}) = 2 \text{ (tid 1, 2)} \end{array}\right\} \Rightarrow conf(\{a,e\} \Rightarrow \{f\}) = \frac{1}{2} = 0.5$$

9. ***Show the candidate itemsets and the frequent itemsets in each level-wise pass of the Apriori algorithm in Exercise (1). Assume an absolute minimum support level of 2.***

Ở bài tập (1) ta đã áp dụng thuật toán *Apriori*, nhắc lại

- k = 1

i ∈ C_1	a	b	c	d	e	f
sup(i)	**3**	**3**	**2**	**3**	**3**	**4**

⇒

i ∈ F_1	a	b	c	d	e	f
sup(i)	**3**	**3**	**2**	**3**	**3**	**4**

C_1 (1-items candidate itemsets)

F_1 (1-items frequent itemsets)

- k = 2

i ∈ C_2	**Lý do cắt (nếu có)**	**sup(i)**
{a, b}		1
{a, c}		1
{a, d}		**2**
{a, e}		**2**
{a, f}		**2**
{b, c}		**2**
{b, d}		**2**
{b, e}		1
{b, f}		**2**
{c, d}		1
{c, e}		1
{c, f}		1
{d, e}		1
{d, f}		**2**
{e, f}		**3**

C_2 (2-items candidate itemsets)

⇒

i ∈ F_2	**sup(i)**
{a, d}	**2**
{a, e}	**2**
{a, f}	**2**
{b, c}	**2**
{b, d}	**2**
{b, f}	**2**
{d, f}	**2**
{e, f}	**3**

F_2 (2-items frequent itemsets)

- k = 3

i ∈ C_3	**Lý do cắt (nếu có)**	**sup(i)**
{a, d, e}	$\{d,e\} \notin F_2$	
{a, d, f}		1
{a, d, b}	$\{a,b\} \notin F_2$	
{a, d, c}	$\{\{d,c\},\{a,c\}\} \notin F_2$	
{a, e, f}		**2**
{a, e, b}	$\{\{a,b\},\{b,e\}\} \notin F_2$	
{a, e, c}	$\{\{a,c\},\{e,c\}\} \notin F_2$	
{a, f, b}	$\{a,b\} \notin F_2$	
{a, f, c}	$\{\{a,c\},\{f,c\}\} \notin F_2$	
{b, c, d}	$\{c,d\} \notin F_2$	
{b, c, f}	$\{c,f\} \notin F_2$	
{b, c, e}	$\{\{b,e\},\{c,e\}\} \notin F_2$	
{b, d, f}		1
{b, d, e}	$\{\{b,e\},\{d,e\}\} \notin F_2$	
{b, f, e}	$\{b,e\} \notin F_2$	
{d, f, e}	$\{d,e\} \notin F_2$	

C_3 (3-items candidate itemsets)

⇒

i ∈ F_3	**sup(i)**
{a, e, f}	**2**

F_3 (3-items frequent itemsets)

Vì $|F_3| = 1$ nên $|C_4| = 0 \Rightarrow$ dừng.
Vậy candidate itemsets và frequent itemsets ở mỗi level-wise pass là:

- k = 1 (pass 1)
  - $C_1 = a, b, c, d, e, f$
  - $F_1 = a, b, c, d, e, f$
- k = 2 (pass 2)
  - $C_1 = \{a, b\}, \{a, c\}, \{a, d\}, \{a, e\}, \{a, f\}, \{b, c\}, \{b, d\}, \{b, e\}, \{b, f\}, \{c, d\}, \{c, e\}, \{c, f\}, \{d, e\}, \{d, f\}, \{e, f\}$
  - $F_1 = \{a, d\}, \{a, e\}, \{a, f\}, \{b, c\}, \{b, d\}, \{b, f\}, \{d, f\}, \{e, f\}$
- k = 3 (pass 3)
  - $C_1 = \{a, d, f\}, \{a, e, f\}, \{b, d, f\}$
  - $F_1 = \{a, e, f\}$