

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



BTLT TUẦN 6: KHAI THÁC DỮ LIỆU

GVHD: NGUYỄN THANH BÌNH

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

Ngày 20 tháng 4 năm 2024

1. Consider the transaction database in the table below:

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

Determine the absolute support of itemsets $\{a, e, f\}$, and $\{d, f\}$.

Convert the absolute support to the relative support.

Bài giải

Absolute support là số lượng các giao dịch (transaction) có chứa itemset trong dữ liệu giao dịch
Itemset $\{a, e, f\}$ xuất hiện ở giao dịch 3 và 4 và itemset của $\{d, f\}$ xuất hiện ở giao dịch 3 và 5 nên:

\Rightarrow absolute support của $\{a, e, f\} = 2$

\Rightarrow absolute support của $\{d, f\} = 2$

Relative support là số lượng các giao dịch (transaction) có chứa itemset trên tổng giao dịch

\Rightarrow relative support của $\{a, e, f\} = 2/5 = 0.4$

\Rightarrow relative support của $\{d, f\} = 2/5 = 0.4$

2. For the database in Exercise 1, compute all frequent patterns at absolute minimum support values of 2, 3, and 4.

Áp dụng thuật toán *Apriori* với minimum support = 2 (suy ra minimum support = 3, 4) để tính mẫu liên hệ

- k = 1

$i \in C_1$	a	b	c	d	e	f
sup(i)	3	3	2	3	3	4

Với minimum support = 2, $F_1 = \{a, b, c, d, e, f\}$

Với minimum support = 3, $F_1 = \{a, b, d, e, f\}$

Với minimum support = 4, $F_1 = \{f\}$

- k = 2

$i \in C_2$	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

Với minimum support = 2, $F_2 = \{\{a, d\}, \{a, e\}, \{a, f\}, \{b, c\}, \{b, d\}, \{b, f\}, \{d, f\}, \{e, f\}\}$

Với minimum support = 3, $F_2 = \{e, f\}$

Với minimum support = 4, $F_2 = \{\}$

- k = 3

$i \in C_3$	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
$\{b, d, e\}$	$\{\{b, e\}, \{d, e\}\} \notin F_2$	
$\{b, f, e\}$	$\{b, e\} \notin F_2$	
$\{d, f, e\}$	$\{d, e\} \notin F_2$	

Với minimum support = 2, $F_3 = \{a, e, f\}$

Với minimum support = 3, $F_3 = \{\}$

Với minimum support = 4, $F_3 = \{\}$

Vì $|F_3| = 1$ nên $|C_4| = 0 \Rightarrow$ dừng

Kết luận

- Với minimum support = 2,
Frequent patterns = $\{a, b, c, d, e, f, \{a, d\}, \{a, e\}, \{a, f\}, \{b, c\}, \{b, d\}, \{b, f\}, \{d, f\}, \{e, f\}, \{a, e, f\}\}$
- Với minimum support = 3, frequent patterns = $\{a, b, d, e, f, \{e, f\}\}$
- Với minimum support = 4, frequent patterns = $\{f\}$

3. For the database in Exercise 1, determine all the maximal frequent patterns at absolute minimum support values of 2, 3, and 4.

Definition 4.2.4 (Maximal Frequent Itemsets) A frequent itemset is maximal at a given minimum support level minsup, if it is frequent, and no superset of it is frequent.

Vì vậy từ bài 2 ta có thể suy ra:

- Với minimum support = 2, maximal frequent patterns = $\{\{a, d\}, \{b, c\}, \{b, d\}, \{b, f\}, \{d, f\}, \{a, e, f\}\}$
(vì $\{\{a, e\}, \{a, f\}, \{e, f\}\} \subset \{a, e, f\}$ và $\{a, b, c, d, e\} \subset \{\{a, d\}, \{b, c\}, \{b, d\}, \{b, f\}, \{d, f\}\}$)
- Với minimum support = 3, maximal frequent patterns = $\{a, b, d, e, \{e, f\}\}$ (vì $\{e, f\} \subset \{e, f\}$)
- Với minimum support = 4, maximal frequent patterns = $\{f\}$

4. Represent the database of Exercise 1 in vertical format.

Dữ liệu bài 1 dưới dạng vertical:

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

5. Consider the transaction database in the table below:

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

Determine all frequent patterns and maximal patterns at support levels of 3, 4, and 5

Bài giải

Áp dụng thuật toán *Apriori* với minimum support = 3

- k = 1

$i \in C_1$	a	b	c	d	e	f
sup(i)	2	3	4	6	8	6

Với minimum support = 3, $F_1 = \{b, c, d, e, f\}$

Với minimum support = 4, $F_1 = \{c, d, e, f\}$

Với minimum support = 5, $F_1 = \{d, e, f\}$

- k = 2

$i \in C_2$	Lý do cắt (nếu có)	sup(i)
{b, c}		1
{b, d}		2
{b, e}		3
{b, f}		3
{c, d}		3
{c, e}		4
{c, f}		2
{d, e}		6
{d, f}		4
{e, f}		6

Với minimum support = 3, $F_2 = \{\{b, e\}, \{b, f\}, \{c, d\}, \{c, e\}, \{d, e\}, \{d, f\}, \{e, f\}\}$

Với minimum support = 4, $F_2 = \{\{c, e\}, \{d, e\}, \{d, f\}, \{e, f\}\}$

Với minimum support = 5, $F_2 = \{\{d, e\}, \{e, f\}\}$

- k = 3

$i \in C_3$	Lý do cắt (nếu có)	sup(i)
{b, e, f}		3
{b, e, c}	$\{b, c\} \notin F_2$	
{b, e, d}	$\{b, d\} \notin F_2$	
{b, f, c}	$\{\{f, c\}, \{b, d\}\} \notin F_2$	
{b, f, d}	$\{b, d\} \notin F_2$	
{c, d, e}		3
{c, d, f}	$\{c, f\} \notin F_2$	
{c, e, f}	$\{c, f\} \notin F_2$	
{d, e, f}		4

Với minimum support = 3, $F_3 = \{\{b, e, f\}, \{c, d, e\}, \{d, e, f\}\}$

Với minimum support = 4, $F_3 = \{d, e, f\}$

Với minimum support = 5, $F_3 = \{\}$

- $k = 4$

$i \in C_3$	Lý do cắt (nếu có)	$\text{sup}(i)$
$\{b, e, f, c\}$	$\{\{b, e, c\}, \{b, f, c\}, \{e, f, c\}\} \notin F_3$	
$\{b, e, f, d\}$	$\{\{b, e, d\}, \{b, f, d\}\} \notin F_3$	
$\{c, d, e, f\}$	$\{\{c, d, f\}, \{c, e, f\}\} \notin F_3$	

Với minimum support = 3, $F_4 = \{\}$

Với minimum support = 4, $F_4 = \{\}$

Với minimum support = 5, $F_4 = \{\}$

Vì $|F_4| = 0 \Rightarrow$ dừng

Kết luận

- Với minimum support = 3
 - Frequent patterns = $\{b, c, d, e, f, \{b, e\}, \{b, f\}, \{c, d\}, \{c, e\}, \{d, e\}, \{d, f\}, \{e, f\}, \{b, e, f\}, \{c, d, e\}, \{d, e, f\}\}$
 - Maximal frequent patterns = $\{\{b, e, f\}, \{c, d, e\}, \{d, e, f\}\}$
- Với minimum support = 4
 - Frequent patterns = $\{c, d, e, f, \{c, e\}, \{d, e\}, \{d, f\}, \{e, f\}, \{d, e, f\}\}$
 - Maximal frequent patterns = $\{\{c, e\}, \{d, e, f\}\}$
- Với minimum support = 5
 - Frequent patterns = $\{d, e, f, \{d, e\}, \{e, f\}\}$
 - Maximal frequent patterns = $\{\{d, e\}, \{d, f\}\}$