Frequent Itemset Mining

Nadjib LAZAAR

LIRMM- UM
COCONUT Team

(PART II)

IMAGINA 18/19

Webpage: http://www.lirmm.fr/~lazaar/teaching.html
Email: lazaar@lirmm.fr

Condensed representation of Frequent Itemsets: Closed and Maximal Itemsets

Maximal Itemsets

- The set of Maximal (frequent) Itemsets:

$$M_\theta = \{P \subset \mathcal{I} |\; freq(P) \geq \theta \wedge \forall P' \subset P : freq(P') < \theta\}$$

An itemset is maximal if it is frequent,
but none of its proper supersets is frequent.

- That is:

$$\forall \theta, \forall P \in F_\theta : (P \in M_\theta) \vee (\exists P' \supset P : freq(P') \geq \theta)$$

Maximal Itemsets

- Every frequent itemset has a maximal superset:

$$\forall \theta, \forall P \in F_\theta : (\exists P' \in M_\theta : P \subseteq P')$$

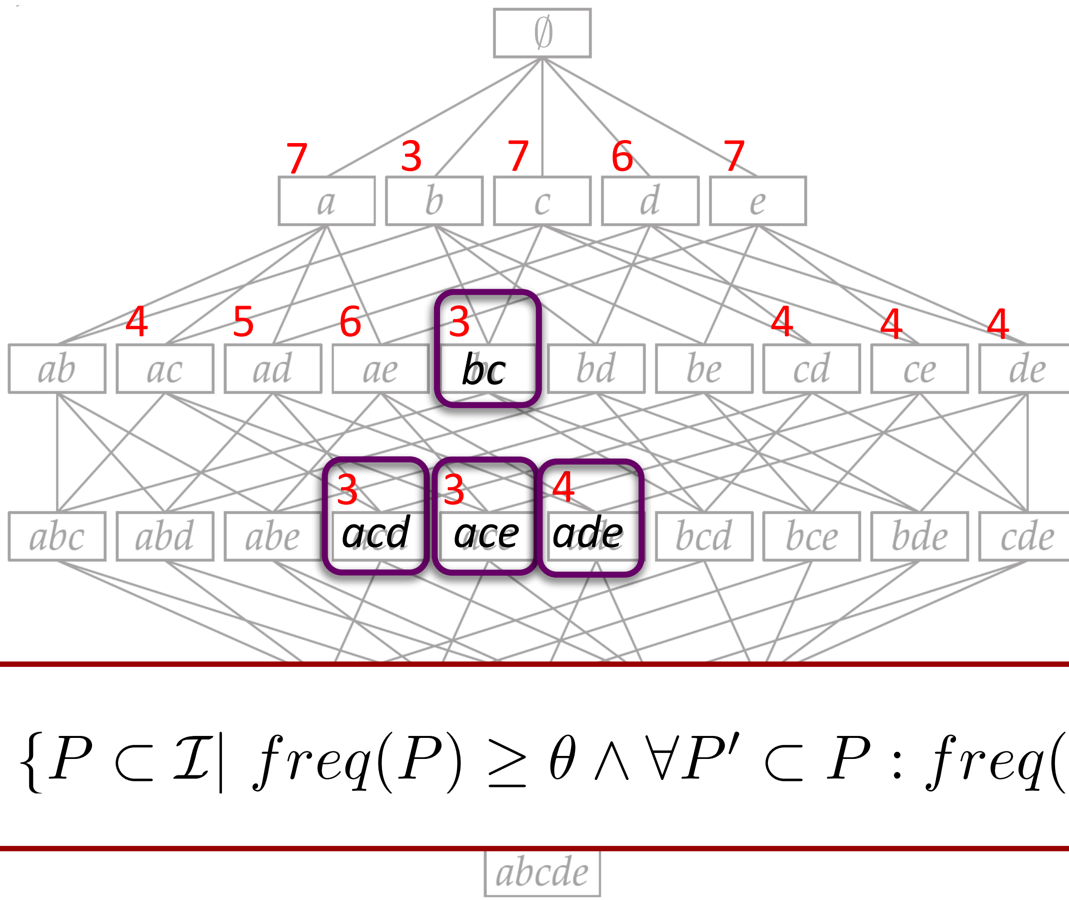
- The maximal itemsets are a condensed representation of the frequent itemsets where:

$$\forall \theta : \; F_\theta = \bigcup_{P \in M_\theta} 2^P$$

Example (5)

Here are the Frequent itemset with minsup θ=3

Q: What are the maximal itemsets minsup θ=3?



	a	b	c	d	e
1:	1	0	0	1	1
2:	0	1	1	1	0
3:	1	0	1	0	1
4:	1	0	1	1	1
5:	1	0	0	0	1
6:	1	0	1	1	0
7:	0	1	1	0	0
8:	1	0	1	1	1
9:	0	1	1	0	1
10:	1	0	0	1	1

matrix representation

$$M_\theta = \{P \subset \mathcal{I} |\ freq(P) \geq \theta \land \forall P' \subset P : freq(P') < \theta\}$$

Maximal Itemsets (limitation)

The set of maximal itemsets captures the set of all frequent itemsets

BUT it does not preserve the knowledge of all support values

THE NEED Can we have a condensed representation of the set of frequent itemsets, which preserves knowledge of all support values?

Closed Itemsets

- The set of Closed (frequent) Itemsets:

$$C_\theta = \{P \subset \mathcal{I} |\ freq(P) \geq \theta \wedge \forall P' \supset P : freq(P') < freq(P)\}$$

An itemset is closed if it is frequent,
but none of its proper supersets has the same support

- That is:

$$\forall \theta, \forall P \in F_\theta : (P \in C_\theta) \vee (\exists P' \supset P : freq(P') = freq(P))$$

Closed Itemsets

- Every frequent itemset has a closed superset:

$$\forall \theta, \forall P \in F_\theta : (\exists P' \in C_\theta : P \subseteq P')$$

- The closed itemsets are a condensed representation of the frequent itemsets where:

$$\forall \theta : \;\; F_\theta = \bigcup_{P \in C_\theta} 2^P$$

Closed Itemsets

- Every frequent itemset has a closed superset with the same support
- The set of all closed itemsets preserves knowledge of all support values:

$$\forall \theta, \forall P \in F_\theta : S(P) = \max_{P' \in C_\theta, P' \supseteq P} S(P')$$

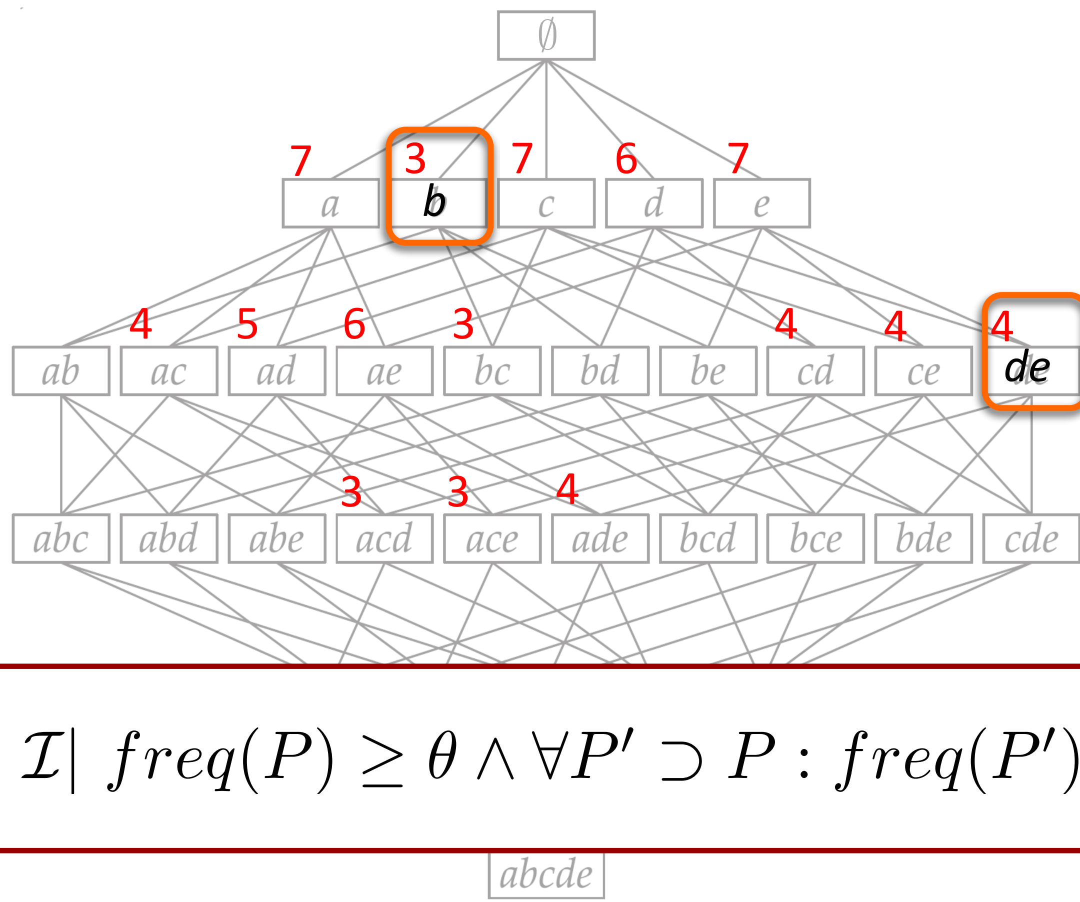
- Which is not the case with the maximal itemsets:

$$\forall \theta, \forall P \in F_\theta : S(P) \supseteq \max_{P' \in M_\theta, P' \supseteq P} S(P')$$

Example (6)

Here are the Frequent itemset with minsup θ=3

Q: are b and de Closed itemsets?

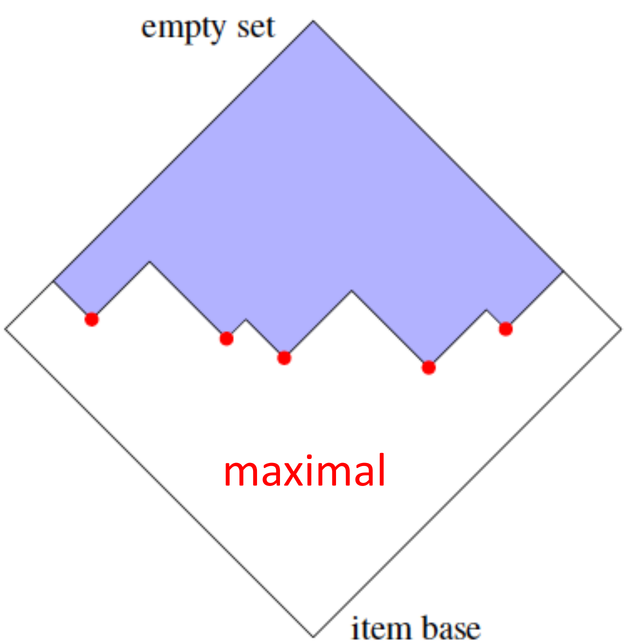


	a	b	c	d	e
1:	1	0	0	1	1
2:	0	1	1	1	0
3:	1	0	1	0	1
4:	1	0	1	1	1
5:	1	0	0	0	1
6:	1	0	1	1	0
7:	0	1	1	0	0
8:	1	0	1	1	1
9:	0	1	1	0	1
10:	1	0	0	1	1

matrix representation

$$C_\theta = \{P \subset \mathcal{I} |\ freq(P) \geq \theta \wedge \forall P' \supset P : freq(P') < freq(P)\}$$

Maximal/Closed itemsets

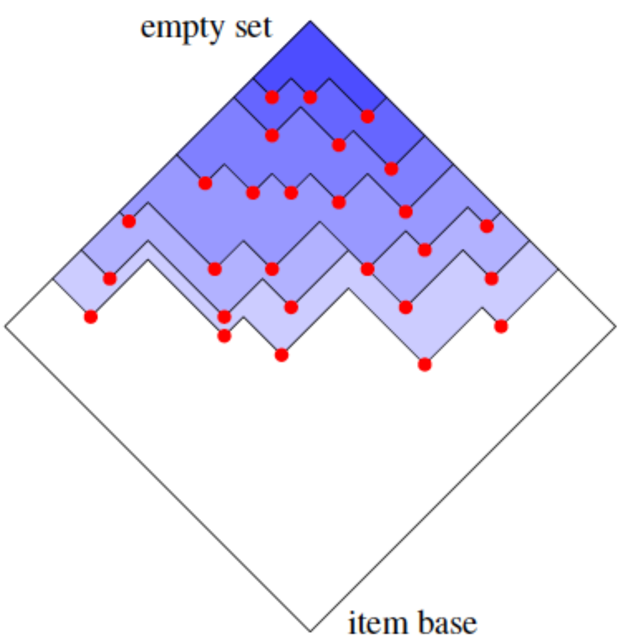


Q: Can the Closed itemsets be expressed using maximal ones?

$$C_\theta = \bigcup_{i \in \{\theta, \ldots, n\}} M_i$$

Frequent/Closed/Maximal

Dataset	#Frequent	#Closed	#Maximal
Zoo-1	151 807	3 292	230
Mushroom	155 734	3 287	453
Lymph	9 967 402	46 802	5 191
Hepatitis	$27 \cdot 10^7$+	1 827 264	189 205



http://fimi.ua.ac.be/data/

Tutorials

link: http://www.lirmm.fr/~lazaar/imagina/TD2.pdf