



Big Data Analytics

– Week 5: Maximal Frequent Itemsets –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

Maximal & Closed Frequent Itemsets

Big challenge: database contains potentially a huge number of frequent itemsets (especially if minSup is set too low).

- A frequent itemset of length 100 contains $2^{100}-1$ many frequent subsets

Closed frequent itemset:

An itemset X is *closed* in a data set D if there exists no proper super-itemset Y such that $support(X) = support(Y)$ in D .

- The set of closed frequent itemsets contains complete information regarding its corresponding frequent itemsets.

Maximal frequent itemset:

An itemset X is *maximal* in a data set D if there exists no proper super-itemset Y such that $support(Y) \geq minSup$ in D .

- The set of maximal itemsets does not contain the complete support information
- More compact representation

Maximal Frequent Itemsets

Given a binary database $\mathbf{D} \subseteq \mathcal{T} \times \mathcal{I}$, over the tids $\mathcal{T}$ and items $\mathcal{I}$, let $\mathcal{F}$ denote the set of all frequent itemsets, that is,

$$\mathcal{F} = \{X \mid X \subseteq \mathcal{I} \text{ and } \text{sup}(X) \geq \text{minsup}\}$$

A frequent itemset $X \in \mathcal{F}$ is called *maximal* if it has no frequent supersets. Let $\mathcal{M}$ be the set of all maximal frequent itemsets, given as

$$\mathcal{M} = \{X \mid X \in \mathcal{F} \text{ and } \nexists Y \supset X, \text{ such that } Y \in \mathcal{F}\}$$

The set $\mathcal{M}$ is a condensed representation of the set of all frequent itemset $\mathcal{F}$, because we can determine whether any itemset X is frequent or not using $\mathcal{M}$. If there exists a maximal itemset Z such that $X \subseteq Z$, then X must be frequent; otherwise X cannot be frequent.

Example for Maximal Itemsets

Transaction database

Tid	Itemset
1	ABDE
2	BCE
3	ABDE
4	ABCE
5	ABCDE
6	BCD

Frequent itemsets ($minsup = 3$)

<i>sup</i>	Itemsets
6	<i>B</i>
5	<i>E, BE</i>
4	<i>A, C, D, AB, AE, BC, BD, ABE</i>
3	<i>AD, CE, DE, ABD, ADE, BCE, BDE, ABDE</i>

GenMax Algorithm

```

// Initial Call:  $\mathcal{M} \leftarrow \emptyset$ ,
 $P \leftarrow \{\langle i, \mathbf{t}(i) \rangle \mid i \in \mathcal{I}, \text{sup}(i) \geq \text{minsup}\}$ 
GENMAX ( $P, \text{minsup}, \mathcal{M}$ ):
1  $Y \leftarrow \bigcup X_i$ 
2 if  $\exists Z \in \mathcal{M}$ , such that  $Y \subseteq Z$  then
3   return // prune entire branch
4 foreach  $\langle X_i, \mathbf{t}(X_i) \rangle \in P$  do
5    $P_i \leftarrow \emptyset$ 
6   foreach  $\langle X_j, \mathbf{t}(X_j) \rangle \in P$ , with  $j > i$  do
7      $X_{ij} \leftarrow X_i \cup X_j$ 
8      $\mathbf{t}(X_{ij}) = \mathbf{t}(X_i) \cap \mathbf{t}(X_j)$ 
9     if  $\text{sup}(X_{ij}) \geq \text{minsup}$  then  $P_i \leftarrow P_i \cup \{\langle X_{ij}, \mathbf{t}(X_{ij}) \rangle\}$ 
10  if  $P_i \neq \emptyset$  then GENMAX ( $P_i, \text{minsup}, \mathcal{M}$ )
11  else if  $\nexists Z \in \mathcal{M}, X_i \subseteq Z$  then
12     $\mathcal{M} = \mathcal{M} \cup X_i$  // add  $X_i$  to maximal set

```

GenMax Example

