WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Concise Representations of Frequent Itemsets

HUMAN CAPITAL
HUMAN – BEST INVESTMENT!
EUROPEAN UNION EUROPEAN SOCIAL FUND
Project is co-financed by European Union within European Social Fund

Do We Need to Know All Frequent Itemsets?

- The number of frequent itemsets is usually huge.
- Time of their discovery can be significant.
- There are cases in which one needs to know only a small subset of frequent itemsets! (*Representative* and minimal *non-redundant rules* can be derived directly from concise representations of frequent itemsets called *closed itemsets* and *generators*.)

2

Lossless Representations of Frequent Itemsets

- Itemsets representation is meant *lossless* if it allows derivation and support determination of all frequent itemsets without accessing the database.

3

Lossless Representations of Frequent Itemsets

Lossless representations of frequent itemsets are based on the following **sets subsuming other sets**:

- *closed itemsets*
- (*key*) *generators*
- (*generalized*) *disjunctive sets*

4

Simple example of reasoning about suports of itemsets

- Let *sup*({*ac*}) = 3 and *sup*({*abcde*}) = 3.
- This information is sufficient to determine the support of {*abce*} as folows:

  3 = *sup*({*ac*}) ≥ *sup*({*abce*}) ≥ *sup*({*abcde*}) = 3.

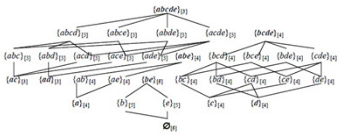
  Hence,

  sup({*ac*}) = *sup*({*abce*}) = *sup*({*abcde*}) = 3.
- In general, if $X \subseteq Y$ and *sup*(X) = *sup*(Y) = k, then for each itemset Z such that: $X \subseteq Z \subseteq Y$, its supports also equals k.

5

Reasoning about suports of itemsets

Id	*Transaction*
T_1	{*abcde*}
T_2	{*abcdef*}
T_3	{*abcdehi*}
T_4	{*abe*}
T_5	{*bcdehi*}

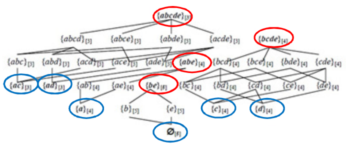


Example. *minSup*=2; *minConf*=77%.

6

Reasoning about suports of itemsets

Id	Transaction
T_1	{abcde}
T_2	{abcdef}
T_3	{abcdehi}
T_4	{abe}
T_5	{bcdehi}



Example. *minSup*=2; *minConf*=77%.

Supports and tid-lists of subsets/supersets

Lemma. Let $X \subseteq Y$. Then:

$$t(X) = t(Y) \Leftrightarrow sup(X) = sup(Y).$$

Proof

(⇒). Trivial.

(⇐). Let $X \subseteq Y$ and $sup(X) = sup(Y)$.

Then, $t(X) \supseteq t(Y)$ and $|t(X)| = |t(Y)|$.

Hence, $t(X) = t(Y)$.

Closures of itemsets

- *A closure of itemset X* is denoted as $\gamma(X)$ and defined as:

$$\gamma(X) = \bigcap\{T \in D \cup \{I\} \mid T \supseteq X\}.$$

- **Note:** An itemset has exactly one itemset as its closure!
- **Property.** The closure of itemset X is the greatest superset $Y \supseteq X$ such that

$$sup(Y) = sup(X).$$

Closed itemsets

- An itemset is defined as *closed* if it is equal to its closure.
- **Property:** Each closure is a closed itemset.
- **Important property of closed itemsets:** The set of all closed itemsets is sufficient to determine support of each itemset X in 2^I, namely:

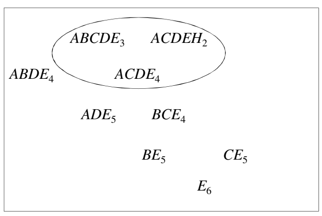
$$sup(X) = \max\{sup(Y) \mid Y \text{ is closed} \wedge Y \supseteq X\}.$$

Closed Itemsets Representation

- *Closed itemsets representation* (CR) consists of all frequent closed itemsets and the information about their supports.

Example: Reasoning with CR

- CR for *minSup*=1.
- $sup(ACD)$=?, $sup(AF)$=?



•$sup(ACD) = \max(sup(ACDE), sup(ABCDE), sup(ACDEH)) = 4$.

•AF has no superset in CR, so: $sup(AF) \leq minSup$.

Calculating *FC*s with CHARM

Id	Transaction
T1	{abcde}
T2	{abcdef}
T3	{abcdehi}
T4	{abe}
T5	{bcdehi}

$h(X\cup Y) = \Sigma_{i\in t(X\cup Y)}\, i$

- $t(X) = t(Y)$:
  - remove node Y from the tree;
  - replace each X with $X\cup Y$.
- $t(X) \subset t(Y)$:
  - replace each X with $X\cup Y$.
- $t(X) \supset t(Y)$:
  - remove node Y from the tree;
  - add $X\cup Y$ as a child of node X.
- otherwise:
  - add $X\cup Y$ as a child of node X.

13

Calculating *FC*s with dCHARM

Id	Transaction
T1	{abcde}
T2	{abcdef}
T3	{abcdehi}
T4	{abe}
T5	{bcdehi}

- $d(X\cup Y) = d(Y)\backslash d(X)$
- $sup(X\cup Y) = sup(X) - |d(X\cup Y)|$
- $h(X\cup Y) = h(X) - (\Sigma_{i\in d(X\cup Y)}\, i)$

- $d(X) = d(Y)$:
  - remove node Y from the tree;
  - replace each X with $X\cup Y$.
- $d(X) \supset d(Y)$:
  - replace each X with $X\cup Y$.
- $d(X) \subset d(Y)$:
  - remove node Y from the tree;
  - add $X\cup Y$ as a child of node X.
- otherwise:
  - add $X\cup Y$ as a child of node X.

14

(Key) generators

Id	Transaction
T1	{abcde}
T2	{abcdef}
T3	{abcdehi}
T4	{abe}
T5	{bcdehi}

15

Generator of an Itemset

- Y is defined *a generator of itemset X* if it is a minimal subset of X such that
  $$\gamma(Y) = \gamma(X).$$
- **Note:** An itemset may have more than one generator!
- **Property.** *A generator of itemset X* is a minimal subset $Y\subseteq X$ such that
  $$sup(Y) = sup(X).$$

16

(Key) Generator

- An itemset X is defined as *a (key) generator* if X's generator is X.
- **Theorem.**
  - All subsets of a generator are generators.
  - All supersets of a non-generator are not generators.

17

Supersets of non-generators

Theorem.

If $X\notin G$, then $\forall Y\supset X,\ Y\notin G$.

Proof. Let $X\notin G$, $Y\supset X$. Then:

$\exists X'\in G(X)$ such that $X'\subset X$ and

$\exists Z\neq\emptyset$ such that $Z = Y\backslash X$

$\Rightarrow t(X') = t(X)$ and

$t(Y)=t(X\cup Z)= t(X)\cap t(Z) = t(X')\cap t(Z) = t(X'\cup Z)$

$\Rightarrow sup(Y) = sup(X\cup Z) = sup(X'\cup Z)$ and

$Y = X\cup Z \supset X'\cup Z$

$\Rightarrow Y\notin G$.

18

Subsets of generators

Theorem A. If $X \notin G$, then $\forall Y \supset X$, $Y \notin G$.

Theorem B.

If $X \in G$, then $\forall Y \subset X$, $Y \in G$.

Proof (by contradiction).

Let $X \in G$, $Y \subset X$ and $Y \notin G$. Then:

By Theorem A all proper supersets of Y (and thus also X) are not generators.

Hence, $X \notin G$, which contradicts the assumption.

19

(Key) Generators

- **Important Property of Generators:**
  The set of all generators is sufficient to determine the support of each itemset X in 2^I, namely:

  $sup(X) = \min\{sup(Y) \mid Y$ is a generator $\wedge\ Y \subseteq X\}$.

20

Generators representation (GR)

- *The generators representation* (GR) consists of:
  1) all frequent generators and the information about their supports,
  2) the border of minimal infrequent generators.

21

GR border

Theorem. X is a minimal infrequent generator $\Leftrightarrow$ X is a minimal infrequent itemset.

Proof ($\Rightarrow$). X is a minimal infrequent generator

$\Rightarrow$ X is a minimal infrequent generator & all proper subsets of X are generators

$\Rightarrow$ X is a minimal infrequent generator & all proper subsets of X are frequent generators

$\Rightarrow$ X is infrequent & all proper subsets of X are frequent

$\Rightarrow$ X is a minimal infrequent itemset.

22

GR border

Theorem. X is a minimal infrequent generator $\Leftrightarrow$ X is a minimal infrequent itemset.

Proof ($\Leftarrow$). X is a minimal infrequent itemset

$\Rightarrow$ X is infrequent & proper subsets of X are frequent

$\Rightarrow$ X is infrequent & all proper subsets of X are frequent & have supports different from $sup(X)$

$\Rightarrow$ X is an infrequent generator & all its proper subsets are frequent

$\Rightarrow$ X is an infrequent generator & all proper subsets of X are frequent generators $\Rightarrow$

$\Rightarrow$ X is a minimal infrequent generator.

23

Calculating *GR* with GR-Apriori

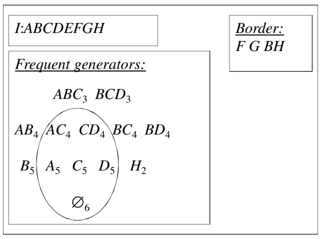
Id	*Transaction*
T1	*{abcde}*
T2	*{abcdef}*
T3	*{abcdehi}*
T4	*{abe}*
T5	*{bcdehi}*

24

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Reasoning with GR

- GR for *minSup* = 1.
- *sup*(*AF*) = ?, *sup*(*ACD*) = ?



- *AF* is infrequent, as it is a superset of *F*, which is an infrequent generator.
- *sup*(*ACD*) = min(*sup*(*AC*), *sup*(*CD*), *sup*(*A*), *sup*(*C*), *sup*(*D*), *sup*(∅)) = 4.

25

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Generalized Disjunctive and Disjunction-Free Sets

- An itemset X is defined *generalized disjunctive* if there is a certain rule based on X, that is if:

$$\exists A_1 \vee \ldots \vee A_n \in X \text{ such that}$$
$$X \setminus \{A_1,\ldots, A_n\} \rightarrow A_1 \vee \ldots \vee A_n \text{ is a certain rule.}$$

- Otherwise, X is called *generalized disjunction-free*.

26

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Certain Generalized Disjunctive Rule and Set

Id	*Transaction*
T_1	***ABCDEG***
T_2	***ABCDEF***
T_3	***ABCDEH***
T_4	*ABDE*
T_5	***ACDEH***
T_6	*BCE*

$\{A,C\} \rightarrow F \vee G \vee H$ is a certain rule.

Hence, $\{A,C,F,G,H\}$ is generalized disjunctive.

27

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Property of Certain Generalized Disjunctive Rules

- If $X \setminus \{A_1,\ldots, A_n\} \rightarrow A_1 \vee \ldots \vee A_n$ is certain, then $XY \setminus \{A_1,\ldots, A_n\} \rightarrow A_1 \vee \ldots \vee A_n$ is also certain.

Id	*Transaction*
T_1	***ABCDEG***
T_2	***ABCDEF***
T_3	***ABCDEH***
T_4	*ABDE*
T_5	***ACDEH***
T_6	*BCE*

Since $\{A,C\} \rightarrow F \vee G \vee H$ is a certain rule, then $\{A,\mathbf{B},C\} \rightarrow F \vee G \vee H$ is also certain.

28

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Certain Generalized Disjunctive Rules and Supports of Sets

- $X \setminus \{A_1,\ldots, A_n\} \rightarrow A_1 \vee \ldots \vee A_n$ is a certain rule **iff** support of X can be calculated from the supports of its proper subsets:

$$sup(X) = (-1)^{|Y|} \times \{-sup(X) + \Sigma_{i=1..|Y|-1} (-1)^{i-1} \times [\Sigma_{i\text{-itemsets } Z \subset Y}\ sup(X \cup Z)]\},$$

where $Y = \{A_1, \ldots, A_n\}$.

29

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Reasoning about Supports of Supersets

- Any generalized certain rule based on X, determines a method of calculating support of X based on its proper subsets.

Id	*Transaction*
T_1	***ABCDEG***
T_2	***ABCDEF***
T_3	***ABCDEH***
T_4	*ABDE*
T_5	***ACDEH***
T_6	*BCE*

$\{A,C\} \rightarrow F \vee G \vee H$ is a certain rule, so $\{A,\mathbf{B},C\} \rightarrow F \vee G \vee H$ is also certain.

Thus, since $\{A,C,F,G,H\}$ is a generalized disjunctive set, then $\{A,\mathbf{B},C,F,G,H\}$ is also a generalized disjunctive set.

30

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Generalized Disjunction Free Generators Representation

- *Generalized disjunction-free generators representation* (GDFGR) consists of:
  - all frequent generalized disjunction-free generators (and their supports),
  - minimal frequent generalized disjunctive generators (and their supports or certain generalized disjunctive rules),
  - minimal infrequent generators...

31

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References…

- Marzena Kryszkiewicz: Concise Representation of Frequent Patterns Based on Disjunction-Free Generators. ICDM 2001: 305-312
- Marzena Kryszkiewicz, Marcin Gajek: Concise Representation of Frequent Patterns Based on Generalized Disjunction-Free Generators. PAKDD 2002: 159-171
- Marzena Kryszkiewicz: Concise Representations of Frequent Patterns and Association Rules, Warsaw: Publishing House of Warsaw University of Technology (2002)

32

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References

- Pasquier N.: Data mining: Algorithmes d'extraction et de Réduction des Règles d'association dans les Bases de Données. Thèse de Doctorat, Université Blaise Pascal – Clermont–Ferrand II (2000)
- Mohammed Javeed Zaki, Ching-Jui Hsiao: Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. IEEE Trans. Knowl. Data Eng. 17(4): 462-478 (2005)

33