Mining Frequent Patterns, Association & Correlation

Part I: Basic Concepts

Hui Yang
Department of Computer Science
San Francisco State University
http://cose-stor.sfsu.edu/~huiyang
Outline

• Basic concepts
• 3 major mining algorithms to discover frequent association patterns
• Mining various types of association patterns
• From association mining to correlation analysis
• Constraint-based association mining
• Summary
What Is Frequent Pattern Analysis?

- The process of identifying frequent patterns in a dataset and applying such patterns to address a variety of issues.
  - **Frequent pattern**: a pattern that occurs *frequently* in a data set. A pattern can be a set of items, an ordered sequence of items, or a substructure.
  - *First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining*

**Examples of frequent patterns**

- *What products were often purchased together?*
  - *(Beer, diapers), (bread, butter, milk)*
- *What are the subsequent purchases after buying a PC?*
  - *Customers often buy MS Office Software within 10 days after they buy a PC.*
- *What sub-structures of amino acids are frequently appearing in proteins?*
Basic Concepts: Association Patterns and Association Rules

- **Given:**
  - \( \text{DB} \): a transactional database of \( n \) transactions: \( T_1, T_2, \ldots, T_n \),
  - \( I \): the set of unique items in \( \text{DB} \)
  - \( T_i \subseteq I \)
  - Ex: 5 TXs, \( I=\{A,B,C,D,E,F\} \)

- **Itemsets (or association sets)**
  - \( S = \{s_1, \ldots, s_k\} \), where \( S \subseteq I \)
  - \( S \) is often called a \( k \)-itemset
  - Ex: \{D\}, \{A,B\}, \{A,B,D\}, \{A,B,D,E,F\}
  - Implication: \{A,B\}--customers purchase A and B together

- **Association rules**
  - \( X \rightarrow Y \), where \( X, Y \subseteq I \) and \( X \cap Y = \emptyset \)
  - \( \text{Reads: } X \text{ implies } Y \), or “the chance that } Y \text{ occurs given that } X \text{ occurs”}
  - \( X \): the antecedent, \( Y \): the precedent
  - Ex: \( AB \rightarrow C, CDE \rightarrow B \)
Measuring Association Patterns and Association Rules

- Itemset \( S = \{s_1, \ldots, s_k\} \)
  - **Support**: the probability that a transaction contains \( S \);
    - Equivalently, the number or percentage of transactions that contain \( S \)
    - Ex: \( \text{supp}(A) = 3 \), \( \text{supp}(D)=4 \), \( \text{supp}(BCD) = 1 \)

- Association rule \( X \rightarrow Y \)
  - **Support**: \( \text{support}(X \cap Y) \)
  - **Confidence**, \( c \), conditional probability that a transaction having \( X \) also contains \( Y \)
    - \( \text{conf}(AB \rightarrow C)=\text{supp}(A,B,C)/\text{supp}(AB) \)
    - Example:
      - \( A \rightarrow D \) (3, 100%) \( \text{(conf}=3/3=100\%) \)
      - \( D \rightarrow A \) (3, 75%) \( \text{(conf}=3/4=75\%) \)

<table>
<thead>
<tr>
<th>Transaction-id</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>A, B, D</td>
</tr>
<tr>
<td>20</td>
<td>A, C, D</td>
</tr>
<tr>
<td>30</td>
<td>A, D, E</td>
</tr>
<tr>
<td>40</td>
<td>B, E, F</td>
</tr>
<tr>
<td>50</td>
<td>B, C, D, E, F</td>
</tr>
</tbody>
</table>

Customer buys both

Customer buys diaper

Customer buys beer

Customer bears beer

B, E, F 40
B, C, D, E, F 50
A, D, E 30
A, C, D 20
A, B, D 10
Mining Frequent Association Patterns

- **Given:**
  - DB: a transactional database of \( n \) transactions: \( T_1, T_2, \ldots, T_n \)
  - I: the set of unique items in DB
  - **User-specified parameters:** minimum support (\( \text{Supp}_{\text{min}} \)), minimum confidence (\( \text{Conf}_{\text{min}} \))

- **Output:**
  - The set of A.R.s that have support \( \geq \text{Supp}_{\text{min}} \) (and confidence \( \geq \text{Conf}_{\text{min}} \))

- **Example:**
  - Let \( \text{Supp}_{\text{min}} = 3 \), \( \text{Conf}_{\text{min}} = 50\% \)
  - Frequent itemsets
    - \{A:3, B:3, D:4, E:3, AD:3\}
  - Freq. association rules:
    - \( A \rightarrow D \) (60\%, 100\%)
    - \( D \rightarrow A \) (60\%, 75\%)

<table>
<thead>
<tr>
<th>Transaction-id</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>A, B, D</td>
</tr>
<tr>
<td>20</td>
<td>A, C, D</td>
</tr>
<tr>
<td>30</td>
<td>A, D, E</td>
</tr>
<tr>
<td>40</td>
<td>B, E, F</td>
</tr>
<tr>
<td>50</td>
<td>B, C, D, E, F</td>
</tr>
</tbody>
</table>
Why Frequent Pattern Analysis?

• Disclose inherent regularities (or correlations) in data
  – *What products were often purchased together?— Beer and diapers?!*
  – *What are the subsequent purchases after buying a PC?*

• Address issues in different applications
  – *Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.*

• Form the foundation of many essential data mining tasks
  – *Pattern analysis in spatiotemporal, multimedia, and stream data*
  – *Classification: associative classification*
  – *Cluster analysis: frequent pattern-based clustering*
  – *Information retrieval: frequent pattern-based approximate indexing*
  – *Semantic data compression (or summarization)*
Closed Patterns and Max-Patterns

• A long pattern contains a combinatorial number of sub-patterns, e.g., \{a_1, \ldots, a_{100}\} contains \(_{100}^1 + \(_{100}^2 + \ldots + \(_{100}^{100}\) = 2^{100} – 1 = 1.27 \times 10^{30}\) sub-patterns!

• Solution: Mine closed patterns and max-patterns instead

• An itemset \(X\) is closed if \(X\) is frequent and there exists no super-pattern \(Y \supset X\), with the same support as \(X\) (proposed by Pasquier, et al. @ ICDT’99)

• An itemset \(X\) is a max-pattern if \(X\) is frequent and there exists no frequent super-pattern \(Y \supset X\) (proposed by Bayardo @ SIGMOD’98)

• Closed pattern is a lossless compression of freq. patterns
  – Reducing the # of patterns and rules
Closed Patterns and Max-Patterns

• Exercise. DB = \{<a_1, \ldots, a_{100}>, < a_1, \ldots, a_{50} >\}
  – \text{Supp}_{\text{min}} = 1

• What is the set of closed itemset?
  – <a_1, \ldots, a_{100}>: 1
  – < a_1, \ldots, a_{50} >: 2

• What is the set of max-pattern?
  – <a_1, \ldots, a_{100}>: 1

• What is the set of all patterns?
  – !!
Freq. Association Patterns: A Main Property

- The **downward closure** (or anti-monotonicity) property of frequent patterns (also:)
  - Any subset of a frequent itemset must be frequent
    - If \{\text{beer, diaper, nuts}\} is frequent, so is \{\text{beer, diaper}\}  
      i.e., every transaction having \{\text{beer, diaper, nuts}\} also contains \{\text{beer, diaper}\}
  - Conversely: If an itemset is not frequent, none of its supersets will be frequent.
    - (\text{beer, diaper}) is not freq.⇒(\text{beer,diaper,nuts}) is not freq.

- Scalable mining methods: 3 major approaches
  - Apriori \((Agrawal & Srikant@VLDB’94)\)
  - Freq. pattern growth \((FPgrowth—Han, Pei & Yin @SIGMOD’00)\)
  - Vertical data format approach \((Charm—Zaki & Hsiao @SDM’02)\)