Chapters 6 & 7, Frequent Pattern Mining

Young-Rae Cho
Associate Professor
Department of Computer Science
Baylor University

- Market Basket Problem
- Apriori Algorithm
- CHARM Algorithm
- Advanced Frequent Pattern Mining
- Advanced Association Rule Mining
- Constraint-Based Association Mining
Market Basket Problem

➢ Example
   ▪ "Customers who bought beer also bought diapers."

➢ Motivation
   ▪ To promote sales in retail by cross-selling

➢ Required Data
   ▪ Customers’ purchase patterns
     (Items often purchased together)

➢ Applications
   ▪ Store arrangement
   ▪ Catalog design
   ▪ Discount plans

Solving Market Basket Problem

➢ Basic Terms
   ▪ Transaction:
     a set of items (which are bought by one person at one time)
   ▪ Frequent itemset:
     a set of items (as a subset of a transaction) which occur frequently
     across transactions
   ▪ Association rule:
     one-direction relationship between two sets of items (e.g., A → B)

➢ Process
   ▪ Step 1, Generation of frequent itemsets
     e.g. \{ beer, nuts, diapers \}
   ▪ Step 2, Generation of association rules
     e.g. \{ beer \} → \{ nuts, diapers \} ➔ expected output knowledge
Frequent Itemsets

- **Transaction Table**
  
<table>
<thead>
<tr>
<th>T-ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bread, eggs, milk, diapers</td>
</tr>
<tr>
<td>2</td>
<td>coke, beer, nuts, diapers</td>
</tr>
<tr>
<td>3</td>
<td>eggs, juice, beer, nuts</td>
</tr>
<tr>
<td>4</td>
<td>milk, beer, nuts, diapers</td>
</tr>
<tr>
<td>5</td>
<td>milk, beer, diapers</td>
</tr>
</tbody>
</table>

- **Support**
  - Frequency of a set of items across transactions
  - \{ milk, diapers \}, \{ beer, nuts \}, \{ beer, diapers \} → 60% support
  - \{ milk, beer, diapers \}, \{ beer, nuts, diapers \} → 40% support

- **Frequent Itemsets**
  - Itemsets having support greater than (or equal to) a user-specified minimum support

Association Rules

- **Frequent Itemsets**
  ( min sup = 60%, size ≥ 2 )
  - \{ milk, diapers \}
  - \{ beer, nuts \}
  - \{ beer, diapers \}

- **Confidence**
  - For \( A \rightarrow B \), percentage of transactions containing \( A \) that also contain \( B \)
  - \{ milk \} → \{ diapers \}, \{ nuts \} → \{ beer \} : 100% confidence
  - \{ diapers \} → \{ milk \}, \{ beer \} → \{ nuts \}, \{ beer \} → \{ diapers \}, and \{ diapers \} → \{ beer \} : 75% confidence

- **Association Rules**
  - Rules having confidence greater than (or equal to) a user-specified minimum confidence
Generalized Formulas

- **Association Rules**
  - \( I = \{ I_1, I_2, \ldots, I_m \} \), \( T = \{ T_1, T_2, \ldots, T_n \} \), \( T_k \subseteq I \) for \( k \)
  - \( A \rightarrow B \) where \( A \subseteq I (A \neq \emptyset) \), \( B \subseteq I (B \neq \emptyset) \), \( A \subseteq T_i \) for \( \exists i \), \( B \subseteq T_j \) for \( \exists j \), and \( A \cap B = \emptyset \)

- **Computation of Support**
  - \( \text{support} (A \rightarrow B) = \frac{P(A \cup B)}{n} \)
  - where \( P(X) = \frac{|\{T_i \mid X \subseteq T_i\}|}{n} \)

- **Computation of Confidence**
  - \( \text{confidence} (A \rightarrow B) = \frac{P(B \mid A)}{P(A)} = \frac{P(A \cup B)}{P(A)} \)

Problem of Support & Confidence

- **Support Table**

<table>
<thead>
<tr>
<th></th>
<th>Tea</th>
<th>Not Tea</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>20</td>
<td>50</td>
<td>70</td>
</tr>
<tr>
<td>Not Coffee</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>SUM</td>
<td>30</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

- **Association Rule, \{Tea\} \rightarrow \{Coffee\}**
  - Support (\{Tea\} \rightarrow \{Coffee\}) ?
  - Confidence (\{Tea\} \rightarrow \{Coffee\}) ?

- **Problems in this Dataset ?**
Alternative Measures

- **Lift**
  
  \[
  \text{lift (A \rightarrow B)} = \frac{\text{confidence (A \rightarrow B)}}{P(B)}
  \]

  - The association rule \( A \rightarrow B \) is interesting if \( \text{lift}(A \rightarrow B) > 1 \)
  - However, it is the same to correlation between A and B

- **Correlation**

  \[
  \text{lift (A \rightarrow B)} = \frac{P(A \cup B)}{P(A) \times P(B)} = \text{correlation (A, B)}
  \]

  - Positive correlation if \( \text{correlation}(A,B) > 1 \)
  - Negative correlation if \( \text{correlation}(A,B) < 1 \)
  - \( A \leftrightarrow B \)

---

**\( \chi^2 \) Test (Chi-Square Test)**

- Evaluates whether an observed distribution in a sample differs from a theoretical distribution (i.e., hypothesis).
- Where \( E_i \) is an expected frequency and \( O_i \) is an observed frequency,
  \[
  \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}
  \]
  
  - The larger \( \chi^2 \), the more likely the variables are related (positively or negatively).
Chapters 6 & 7, Frequent Pattern Mining

- Market Basket Problem
  - Apriori Algorithm
- CHARM Algorithm
- Advanced Frequent Pattern Mining
- Advanced Association Rule Mining
- Constraint-Based Association Mining

Frequent Itemset Mining

- Process
  1. Find frequent itemsets → computational problem
  2. Find association rules

- Brute Force Algorithm for Frequent Itemset Generation
  - Enumerate all possible subsets of the total itemset, \( I \)
  - Count frequency of each subset
  - Select frequent itemsets

- Problem
  - Enumerating all candidates is not computationally acceptable
    → Efficient & scalable algorithm is required.
Apriori Algorithm

- **Motivations**
  - Efficient frequent itemset analysis
  - Scalable approach

- **Process**
  - Iterative increment of the itemset size
    1. Candidate itemset generation → computational problem
    2. Frequent itemset selection

- **Downward Closure Property**
  - Any superset of an itemset $X$ cannot have higher support than $X$.
    → If an itemset $X$ is frequent (support of $X$ is higher than min. sup.),
    then any subset of $X$ should be frequent.

Candidate Itemset Generation

- **Process**
  - Two steps: (1) selective joining and (2) a priori pruning

- **Selective Joining**
  - Each candidate itemset with size $k$ is generated by joining two frequent itemsets with size $(k-1)$
  - The frequent itemsets with size $(k-1)$ which share a frequent sub-itemset with size $(k-2)$ are joined

- **A priori Pruning**
  - A frequent itemset with size $k$ which has any infrequent sub-itemsets with size $(k-1)$ is pruned
Detail of Apriori Algorithm

- **Basic Terms**
  - $C_k$: Candidate itemsets of size $k$
  - $L_k$: Frequent itemsets of size $k$
  - $sup_{min}$: Minimum support

- **Pseudo Code**
  
  
  $k \leftarrow 1$
  $L_k \leftarrow$ frequent itemsets with size 1
  
  while $L_k \neq \emptyset$
  
  
  $k \leftarrow k + 1$
  $C_k \leftarrow$ candidate itemsets by selective joining & a priori pruning from $L_{k-1}$
  $L_k \leftarrow$ frequent itemsets using $sup_{min}$
  
  end while
  
  return $U_k L_k$

Example of Apriori Algorithm

- $sup_{min} = 2$

<table>
<thead>
<tr>
<th>T-ID</th>
<th>Items</th>
<th>$L_1$</th>
<th>$C_1$</th>
<th>$L_2$</th>
<th>$C_2$</th>
<th>$L_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>A, C, D</td>
<td>2</td>
<td>(A) 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>B, C, E</td>
<td>3</td>
<td>(B) 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>A, B, C, E</td>
<td>1</td>
<td>(C) 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>B, E</td>
<td>1</td>
<td>(D) 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>A, B, C, E</td>
<td>2</td>
<td>(E) 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Sup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A,B)</td>
<td>1</td>
</tr>
<tr>
<td>(A,C)</td>
<td>2</td>
</tr>
<tr>
<td>(A,E)</td>
<td>1</td>
</tr>
<tr>
<td>(B,C)</td>
<td>2</td>
</tr>
<tr>
<td>(B,E)</td>
<td>3</td>
</tr>
<tr>
<td>(C,E)</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Sup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B,C,E)</td>
<td>2</td>
</tr>
</tbody>
</table>
Summary of Apriori Algorithm

➤ Features
  ▪ An iterative approach of a level-wise search
  ▪ Reducing search space by downward closure property

➤ References

Challenges of Apriori Algorithm

➤ Challenges
  ▪ Multiple scan of transaction database
  ▪ Huge number of candidates
  ▪ Tedious workload of support counting

➤ Solutions
  ▪ Reducing transaction database scans
  ▪ Shrinking number of candidates
  ▪ Facilitating support counting
Chapters 6 & 7, Frequent Pattern Mining

- Market Basket Problem
- Apriori Algorithm
  - **CHARM Algorithm**
- Advanced Frequent Pattern Mining
- Advanced Association Rule Mining
- Constraint-Based Association Mining

Association Rule Mining

- **Process**
  1. Find frequent itemsets → computational problem
  2. Find association rules → redundant rule generation

- **Example 1**
  - \{ beer \} \rightarrow \{ nuts \} \ (40\% \text{ support}, 75\% \text{ confidence})
  - \{ beer \} \rightarrow \{ nuts, diapers \} \ (40\% \text{ support}, 75\% \text{ confidence})
  - The first rule is not meaningful.

- **Example 2**
  - \{ beer \} \rightarrow \{ nuts \} \ (60\% \text{ support}, 75\% \text{ confidence})
  - \{ beer, diapers \} \rightarrow \{ nuts \} \ (40\% \text{ support}, 75\% \text{ confidence})
  - Both rules are meaningful.
Frequent Closed Itemsets

- **General Definition of Closure**
  - A frequent itemset \( X \) is **closed** if there exists no superset of \( X, Y \supseteq X \), with the same support as \( X \).  
  - Different from frequent **maximal** itemsets

- **Frequent Closed Itemsets** with **Min. Support of 40%**
  - \( \{ \text{ milk, diapers } \} \) 60%
  - \( \{ \text{ milk, beer } \} \) 60%
  - \( \{ \text{ beer, nuts } \} \) 60%
  - \( \{ \text{ beer, diapers } \} \) 60%
  - \( \{ \text{ nuts, diapers } \} \) 60%
  - \( \{ \text{ milk, beer, diapers } \} \) 40%
  - \( \{ \text{ beer, nuts, diapers } \} \) 40%

<table>
<thead>
<tr>
<th>T-ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bread, eggs, milk, diapers</td>
</tr>
<tr>
<td>2</td>
<td>coke, beer, nuts, diapers</td>
</tr>
<tr>
<td>3</td>
<td>eggs, juice, beer, nuts</td>
</tr>
<tr>
<td>4</td>
<td>milk, beer, nuts, diapers</td>
</tr>
<tr>
<td>5</td>
<td>milk, beer, diapers</td>
</tr>
</tbody>
</table>

- **Mapping of Items and Transactions**

  - **Mapping Functions**
    - \( I = \{ I_1, I_2, \ldots, I_m \}, \ T = \{ T_1, T_2, \ldots, T_n \}, \ X \subseteq I, \ Y \subseteq T \)
    - \( i: T \rightarrow I, \ i(Y): \) itemset that is contained in all transactions in \( Y \)
    - \( t: I \rightarrow T, \ t(X): \) set of transactions (tidset) that contain all items in \( X \)

  - **Properties**
    - \( X_1 \subseteq X_2 \rightarrow t(X_1) \supseteq t(X_2) \)
      - (e.g.) \( \{\text{ACW}\} \subseteq \{\text{ACTW}\} \rightarrow \{1345\} \supseteq \{135\} \)
    - \( Y_1 \subseteq Y_2 \rightarrow i(Y_1) \supseteq i(Y_2) \)
      - (e.g.) \( \{245\} \subseteq \{2456\} \rightarrow \{\text{CDW}\} \supseteq \{\text{CD}\} \)
    - \( X \subseteq i(t(X)), \ Y \subseteq t(i(Y)) \)
      - (e.g.) \( t(i(\{\text{AC}\})) = \{1345\}, \ i(\{1345\}) = \{\text{ACW}\} \)
      - (e.g.) \( i(\{134\}) = \{\text{ACW}\}, \ t(i(\{\text{ACW}\})) = \{1345\} \)
Definition of Closure

- **Closure Operator**
  - $c_i(x) = i(t(x))$, $c_i(y) = i(t(y))$

- **Formal Definition of Closure**
  - An itemset $X$ is closed if $X = c_i(x)$
  - A tid-set $Y$ is closed if $Y = c_i(y)$

Examples of Closed Itemsets

- **Examples**
  - $X = \{ACW\}$
    - $t(x) = \{1345\}$, $i(t(x)) = \{ACW\}$
    - $X$ is closed.
  - $X = \{AC\}$
    - $t(x) = \{1345\}$, $i(t(x)) = \{ACW\}$
    - $X$ is not closed.
  - $X = \{ACT\}$
    - $t(x) = \{135\}$, $i(t(x)) = \{ACTW\}$
    - $X$ is not closed.
  - $X = \{CT\}$
    - $t(x) = \{1356\}$, $i(t(x)) = \{CT\}$
    - $X$ is closed.
CHARM Algorithm

- **Motivations**
  - Efficient frequent closed itemset analysis
  - Non-redundant rule generation

- **Property**
  - Simultaneous exploration of itemset space and tid-set space
  - Not enumerating all possible subsets of a closed itemset
  - Early pruning strategy for infrequent and non-closed itemsets

- **Process**
  - for each itemset pair
    - computing the frequency of their union set
    - pruning all infrequent and non-closed branches

---

Frequency Computation

- **Operation**
  - Tid-set of the union of two itemsets, $X_1$ and $X_2$
  - Intersection of two tid-sets, $t(X_1)$ and $t(X_2)$

$$t(X_1 \cup X_2) = t(X_1) \cap t(X_2)$$

- **Example**
  - $X_1 = \{AC\}, \ X_2 = \{D\}$
  - $t(X_1 \cup X_2) = t(\{ACD\}) = \{45\}$
  - $t(X_1) \cap t(X_2) = \{1345\} \cap \{2456\} = \{45\}$

<table>
<thead>
<tr>
<th>T-ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A, C, T, W</td>
</tr>
<tr>
<td>2</td>
<td>C, D, W</td>
</tr>
<tr>
<td>3</td>
<td>A, C, T, W</td>
</tr>
<tr>
<td>4</td>
<td>A, C, D, W</td>
</tr>
<tr>
<td>5</td>
<td>A, C, D, T, W</td>
</tr>
<tr>
<td>6</td>
<td>C, D, T</td>
</tr>
</tbody>
</table>
Pruning Strategy

Pruning
- Suppose two itemsets $X_1 \leq X_2$
  1. $t(X_1) = t(X_2)$ \rightarrow $t(X_1) \cap t(X_2) = t(X_1) = t(X_2)$
     \rightarrow Replace $X_1$ with $(X_1 \cup X_2)$, and prune $X_2$
  2. $t(X_1) \subset t(X_2)$ \rightarrow $t(X_1) \cap t(X_2) = t(X_1) \neq t(X_2)$
     \rightarrow Replace $X_1$ with $(X_1 \cup X_2)$, and keep $X_2$
  3. $t(X_1) \supset t(X_2)$ \rightarrow $t(X_1) \cap t(X_2) = t(X_2) \neq t(X_1)$
     \rightarrow Replace $X_2$ with $(X_1 \cup X_2)$, and keep $X_1$
  4. $t(X_1) \neq t(X_2)$ \rightarrow $t(X_1) \cap t(X_2) \neq t(X_1) \neq t(X_2)$
     \rightarrow Keep $X_1$ and $X_2$

Example of CHARM Algorithm

Subset Lattice

<table>
<thead>
<tr>
<th>T-ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A, C, T, W</td>
</tr>
<tr>
<td>2</td>
<td>C, D, W</td>
</tr>
<tr>
<td>3</td>
<td>A, C, T, W</td>
</tr>
<tr>
<td>4</td>
<td>A, C, D, W</td>
</tr>
<tr>
<td>5</td>
<td>A, C, D, T, W</td>
</tr>
<tr>
<td>6</td>
<td>C, D, T</td>
</tr>
</tbody>
</table>

50% minimum support
Summary of CHARM Algorithm

- **Advantages**
  - No need multiple scan of transaction database
    → Revision and enhancement of Apriori algorithm
  - No loss of information

- **References**

---

CSI 4352, Introduction to Data Mining

**Chapters 6 & 7, Frequent Pattern Mining**

- Market Basket Problem
- Apriori Algorithm
- CHARM Algorithm
  - Advanced Frequent Pattern Mining
- Advanced Association Rule Mining
- Constraint-Based Association Mining