Chapter 4 – Association Pattern Mining

Data Mining by Charu C. Aggarwal

Agenda

- Target Case study
- What is Association Mining?
- Association Mining Details
- Terminologies
- Apriori Principle
- Applications
- Q&A

Target Case (Circa 2012)

- The guest marketing analytics team at Target inferred based on purchasing data (involving elevated rates of buying unscented lotion, mineral supplements, and cotton balls) that one of its customers—a teenage girl in Minnesota, was pregnant.

What is Association Mining?

- Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional and/or relational databases, and other information repositories.

Association Mining

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Examples

{bread} ⇒ {milk}

{soda} ⇒ {chips}

{bread} ⇒ {jam}

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Terminologies frequently used

- **Itemset**
  - A collection of one or more items, e.g., {milk, bread, jam}
  - k-itemset, an itemset that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an itemset
  - σ({Milk, Bread}) = 3
    σ({Soda, Chips}) = 4
- **Support**
  - Fraction of transactions that contain an itemset
  - s({Milk, Bread}) = 3/8
    s({Soda, Chips}) = 4/8
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a minsup threshold

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Example

1. Consider the following association

   {Milk, Diapers} -> {Beer}

2. The support count for the above mentioned association

   {Milk, Diapers, Beer} is 2

3. Total number of transactions is 5
4. The Rule's support is 2/5 = 0.4
5. The rule's confidence is obtained by dividing the support count for {Milk, Diapers, Beer} by the support count for {Milk, Diapers}.
6. The confidence for this rule is 2/3 = 0.67

Table 6.1. An example of market basket transactions.

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Why Use Support and Confidence?

Support

- Support is an important measure because a rule that has very low support may occur by chance.
- A low support may also be less enticing from business point of view.
- Support is used to eliminate uninteresting rules.

Confidence

- Confidence measures the reliability of the inference made by a rule.
  - Example: X -> Y, the higher the confidence the more likelihood that Y will be present in transactions that contain X.
- Provides an estimate of the conditional probability of Y given X

Please Note…

Association analysis results should be interpreted with Caution!

Inference made by an association rule does not necessarily imply causality.

It suggests a strong co-occurrence relationship between items in the antecedent and consequent of the rule.

- Brute Force Algorithms
  - List all possible association rules.
  - Compute the support and confidence for each rule.
  - Prune rules that fail the ‘minsup’ and ‘minconf’ thresholds.
  - Not pragmatic.

Apriori principle

- Apriori is the first association rule mining algorithm that pioneered the use of support-based pruning to systematically control the exponential growth of candidate item-sets.
- If an item-set is frequent, then all of its subsets must also be frequent.

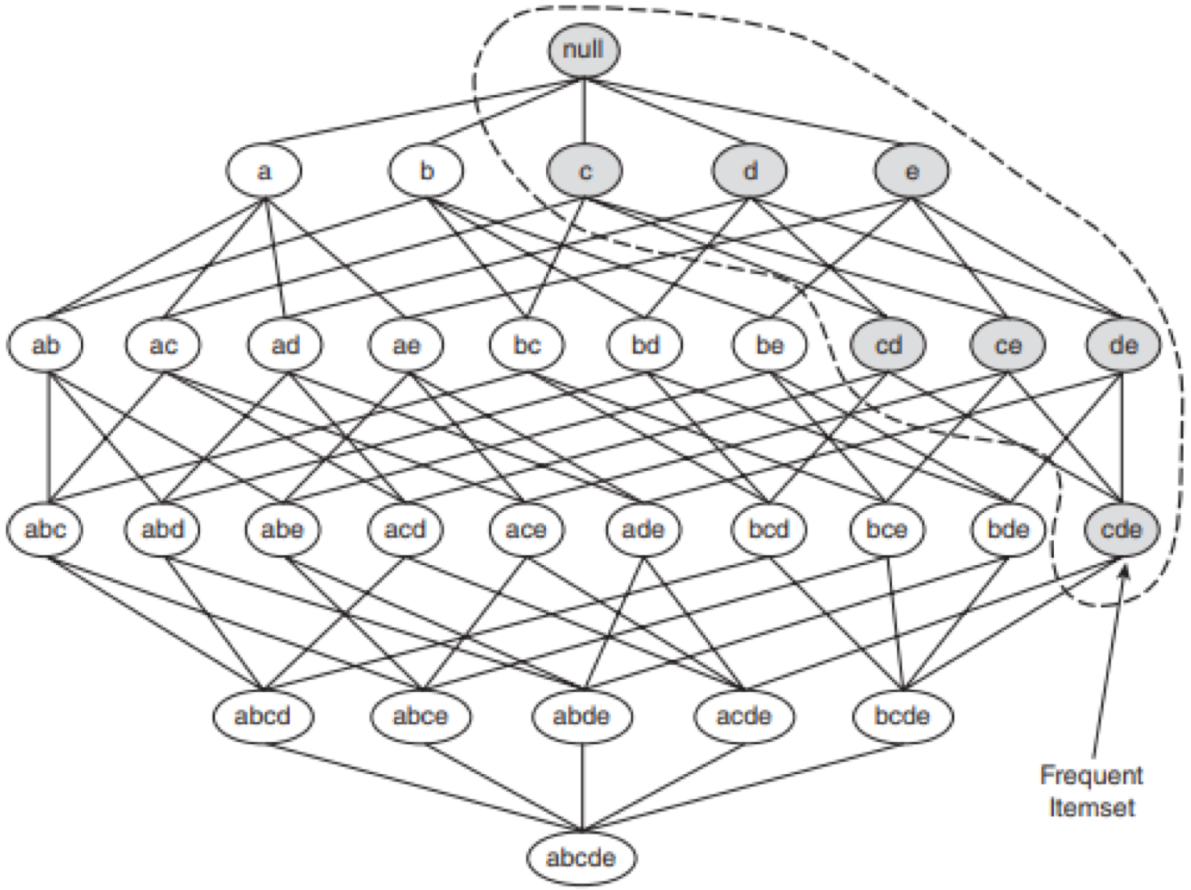
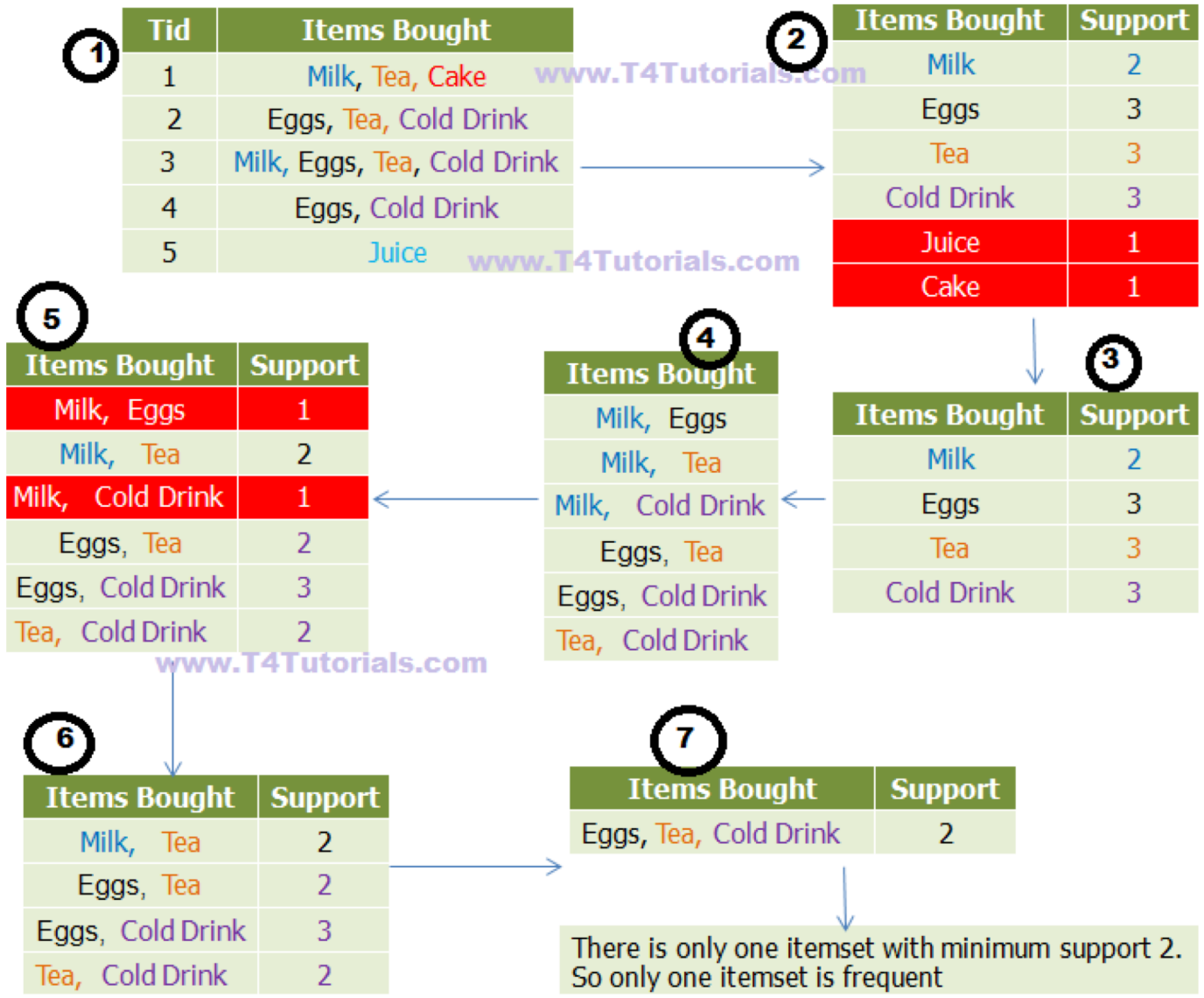


Figure 6.3. An illustration of the *Apriori* principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.

How does the Apriori principle work?



Applications

Market Basket Analysis

- A typical and widely-used example of association rule mining is Market basket analysis
  - Commercial world always want to know more about your purchasing patterns and everything about you!
  - Brand Promotions
  - Inventory Management
  - Customer Relationship Management

Other applications

- Medical Diagnosis
- Protein sequences
- Census Data
- Customer Relationship Management of Credit card business

Thank you!

Rita Chowdhury