Data Warehousing & Mining
Unit-II
Association

Dr. K. Raghava Rao
Professor of CSE, Dept. of MCA
KL University
krraocse@gmail.com
http://datamining.blog.com

Mining Frequent Patterns, Association and Correlations

Definition:

>-Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set

>- Examples:

-- a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent itemset.

--buying first a PC , then a digital camera, and then a memory card, it occurs frequently in history database is frequently sequential pattern (Subsequence).

--it is a subgraph,subtrees or sublattices, which may be combined with itemsets or subsequences are called substructures.

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

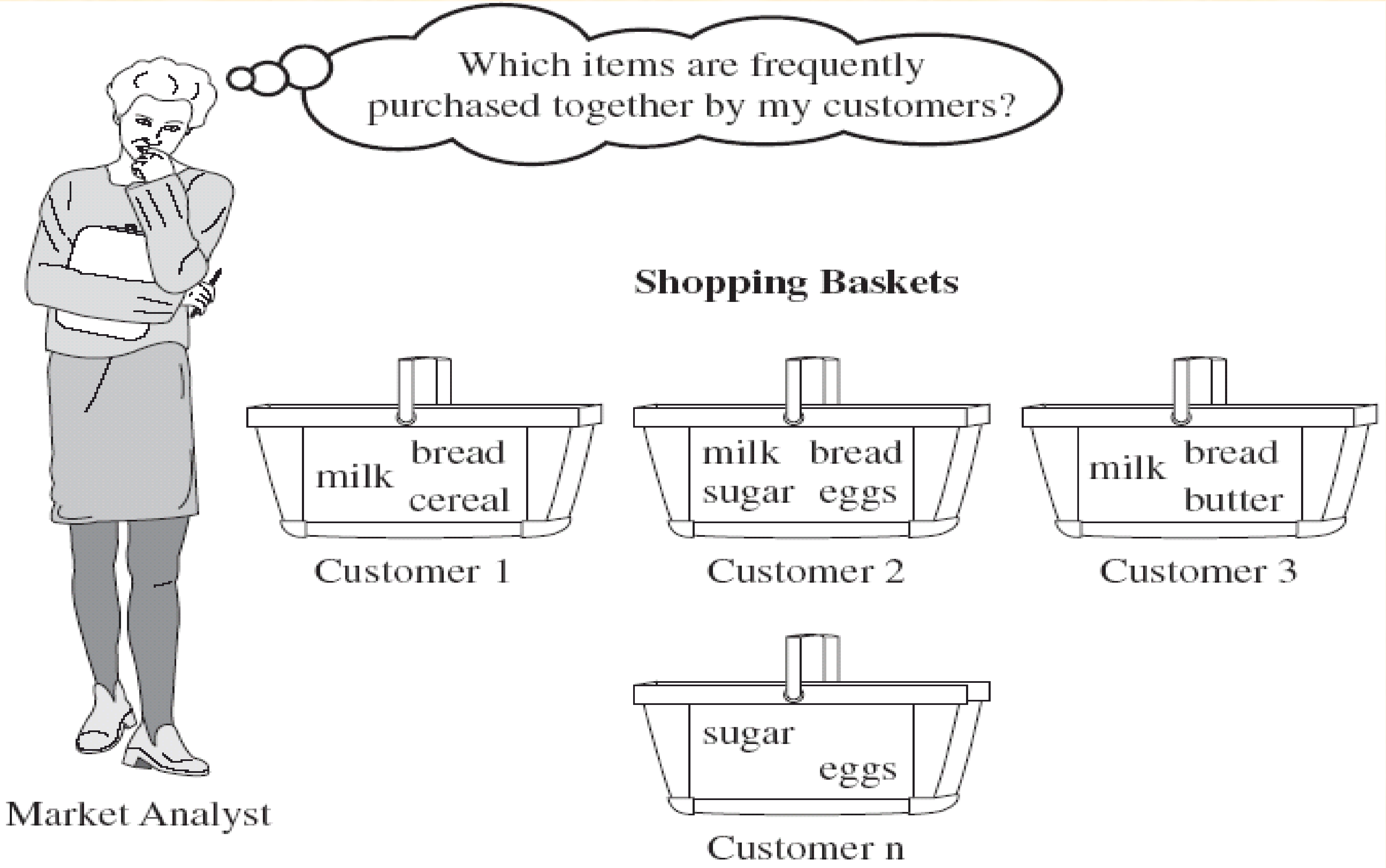
>-Frequent patterns mining searches for recurring relationships in a given dataset.

>-Example for frequent pattern mining for association rules: Market Basket data analysis.

>-Market basket analysis is a typical example of frequent itemset mining
Where Customers buying habits are divined by finding associations between different items that customers place in their “shopping baskets”
This information can be used to develop marketing strategies

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Market basket analysis: A motivating Example



Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Market basket analysis: A motivating Example

>-It is performed on retail data of customer transaction at store.

>-if customer who purchase computers also tend to buy antivirus software, at same time, then placing hardware display close to software display may help increase the sales of both of them.

>-another way placing above both opposite ends of store ,the customers pick up other items along the way.

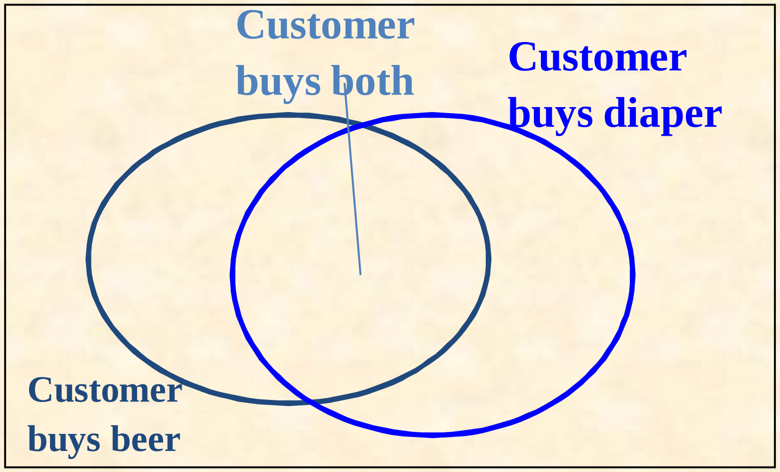
>-Market basket analysis can help retailers plan which items to put on sale at reduced prices.

>-set of items available at store , then each item represent with Boolean variable for absence or present.

>-each basket represents by a Boolean vector can be analyzed for buying patterns that reflect items that are frequently associated or purchased together.

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Market basket analysis: A motivating Example



Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

>-Find all the rules X & $Y \Rightarrow Z$ with minimum confidence and support

- support, *s*, probability that a transaction contains {X 🗐 Y 🗐 Z}
- confidence, *c*, conditional probability that a transaction having {X 🗐 Y} also contains *Z*

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication Rules:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of *n* binary attributes called *items*. Let $D = \{t_1, t_2, \ldots,$ be a set of transactions called the *database*. Each transaction in *D* has a unique transaction ID and contains a subset of the items in *I*

A *rule* is defined as an implication of the form

Example data base with 4 items and 5 transactions

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication

>-The set of items is *I* = {milk,bread,butter,beer} and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table to the right. An example rule for the supermarket could be $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ meaning that if milk and bread is bought, customers also buy butter.

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence. The *support* supp(*X*) of an itemset *X* is defined as the proportion of transactions in the data set which contain the itemset. In the example database, the itemset {milk,bread} has a support of 2 / 5 = 0.4 since it occurs in 40% of all transactions (2 out of 5 transactions

The *confidence* of a rule is defined $\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$
For example, the rule $\{\mathrm{milk, bread}\} \Rightarrow \{\mathrm{butter}\}$
has a confidence of 0.2 / 0.4 = 0.5 in the database, which means that for 50% of the transactions containing milk and bread the rule is correct. Confidence can be interpreted as an estimate of the probability *P*(*Y* | *X*), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication

>-Association rules are required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

>-To achieve this, association rule generation is a two-step process.

1)Find all frequent itemsets:
minimum support is applied to find all *frequent itemsets* in a database.

2)In this step, generate strong association rules from the frequent itemsets:
these rules must satisfy minimum support and minimum confidence.

>-While the second step is straight forward, the first step needs more attention.

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication

A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \ldots, a_{100}\}$ contains $({}_{100}{}^{1}) + ({}_{100}{}^{2}) + \ldots + ({}_{1}{}^{1}{}_{0}{}^{0}{}_{0}{}^{0}) = 2^{100} – 1 = 1.27*10^{30}$ 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 ɔ X, *with the same support* as X.

An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ɔ X

Closed pattern is a lossless compression of freq. patterns

Reducing the # of patterns and rules

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication

Closed Patterns and Max-Patterns

Exercise. DB = {$<a_1, \ldots, a_{100}>$, $< a_1, \ldots, a_{50}>$}

Min_sup = 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?

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication (Revised again)

The model: rules

- A transaction *t* contains *X*, a set of items (itemset) in *I*, if $X \subseteq t$.
- An association rule is an implication of the form:

  $X \rightarrow Y$, where $X, Y \subset I$, *and* $X \cap Y = \emptyset$

- An itemset is a set of items.
  - E.g., X = {milk, bread, cereal} is an itemset.
- A *k*-itemset is an itemset with *k* items.
  - E.g., {milk, bread, cereal} is a 3-itemset

Frequent Itemsets,Closed Itemsets and Assoication (Revised again)

Rule strength measures

- Support: The rule holds with support *sup* in *T* (the transaction data set) if sup% of transactions contain $X \cup Y$.
  - $sup = \Pr(X \cup Y)$.
- Confidence: The rule holds in *T* with confidence *conf* if *conf*% of tranactions that contain *X* also contain *Y*.
  - $conf = \Pr(Y \mid X)$
- An association rule is a pattern that states when *X* occurs, *Y* occurs with certain probability

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent Itemsets,Closed Itemsets and Assoication (Revised again)

Support and Confidence

- Support count: The support count of an itemset *X*, denoted by *X.count*, in a data set *T* is the number of transactions in *T* that contain *X*. Assume *T* has *n* transactions.
- Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Mining Frequent Patterns, Association and Correlations: Basic Concepts and Road Map

Frequent pattern Mining: A RoadMap

>-Frequent pattern mining can be classified in various ways, based on the following criteria:

--based on the completeness of patterns to be mined.

--based on the levels of abstraction involved in the rule set.

--based on the number of data dimensions involved in the rule set.

--based on no. of data dimensions involved in the rule.

--based on the types of values handled in the rule

--based on the kinds of rules to be mined.

--based on the kinds of patterns to be mined.

Mining Frequent Patterns, Association and Correlations: Efficient & Scalable Frequent Itemset Mining Methods

Mining simplest form of frequent patterns are:
-Single-dimensional
-Boolean frequent item sets

Scalable mining methods: Three major approaches

- Apriori is a basic algorithm for finding frequent itemsets.
- Freq. pattern growth is a improved version of apriori algorithm
- Vertical data format approach used to mine frequent itemsets without generating “candidates” frequent itemsets.

Switch to part2.1 unit-2 dw&dm slides.

Mining Frequent Patterns, Association and Correlations: Efficient & Scalable Frequent Itemset Mining Methods

Apriori: A Candidate Generation-and-Test Approach

Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!

-- Initially, scan DB once to get frequent 1-itemset

– Generate length (k+1) candidate itemsets from length k frequent itemsets

– Test the candidates against DB

– Terminate when no frequent or candidate set can be generated