COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 14
TOPICS TO BE COVERED:

- Data mining techniques:
- Association rules

WHAT IS FREQUENT PATTERN ANALYSIS?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- **Frequent patterns are itemsets, subsequences, or substructures.**
- For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a *frequent itemset.*
- *A subsequence, such as* buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (*frequent) sequential pattern.*
- *A substructure can refer* to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (*frequent) structured pattern.*
- *Finding such frequent patterns plays an essential role in* mining associations, correlations, and many other interesting relationships among data.
- Moreover, it helps in data classification, clustering, and other data mining tasks as well.

WHAT IS FREQUENT PATTERN ANALYSIS?

- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications
  - Market basket analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

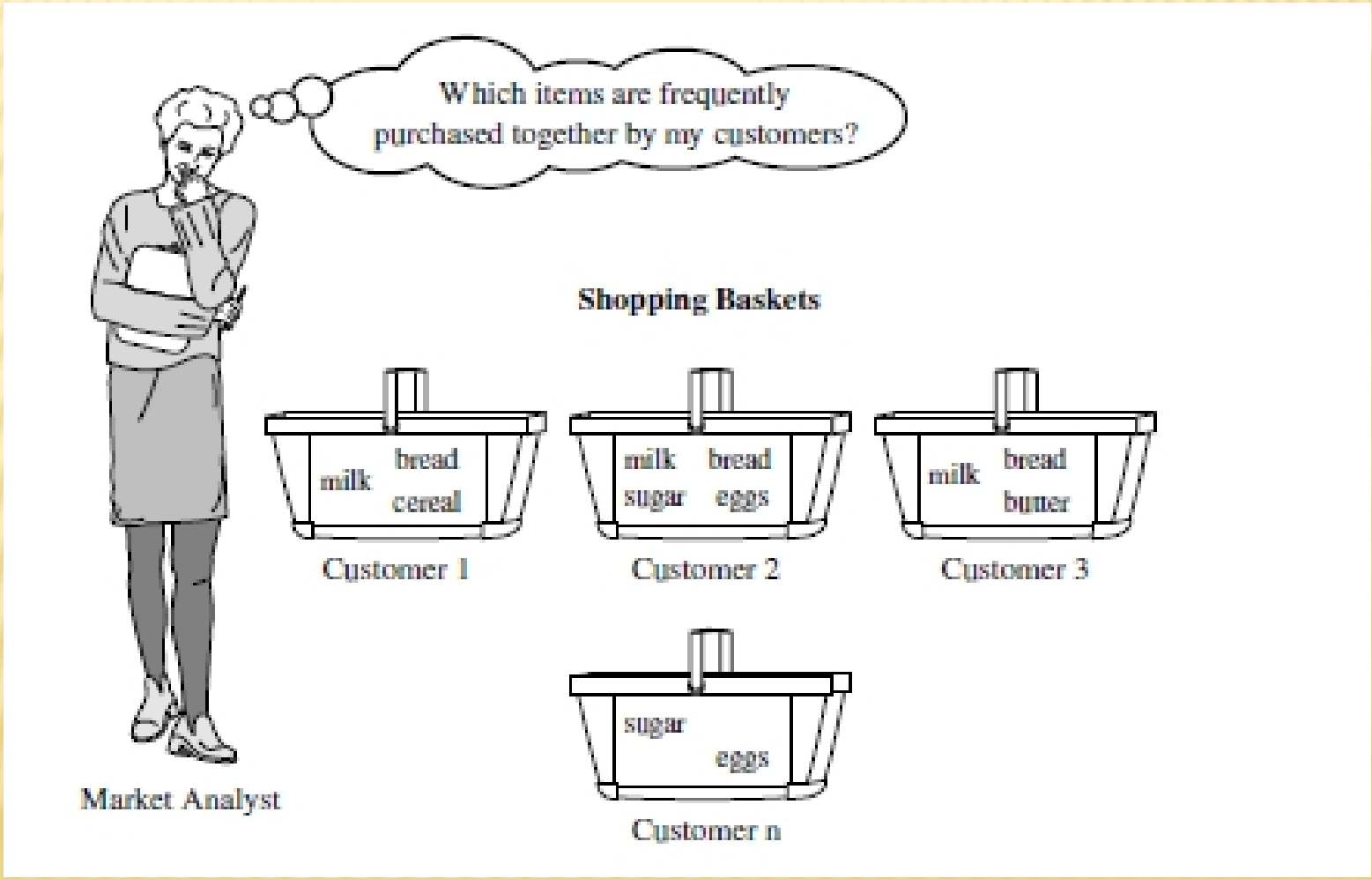
WHY IS FREQ. PATTERN MINING IMPORTANT?

- Discloses an intrinsic and important property of data sets
- Forms the foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: associative classification
  - Cluster analysis: frequent pattern-based clustering
  - Data warehousing: iceberg cube and cube-gradient
  - Semantic data compression: fascicles
  - Broad applications

MARKET BASKET ANALYSIS

- This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets” .
- The discovery of such association scan help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.
- which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence).

MARKET BASKET ANALYSIS



Market Basket Analysis

- you would like to learn more about the buying habits of your customers.

  “Which groups or sets of items are customers likely to purchase on a given trip to the store?”
- Market basket analysis can also help retailers plan which items to put on sale at reduced prices.
- If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers *as well as computers.*

Market Basket Analysis

- If the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently *associated or purchased together. These patterns can be represented in the* form of association rules.
- For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule below:

$$computer \Rightarrow antivirus\_software \ [support = 2\%, confidence = 60\%]$$

MARKET BASKET ANALYSIS

- **Rule support** and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association

  Rule means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.
- **A confidence** of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.

FREQUENT ITEMSETS, CLOSED ITEMSETS AND ASSOCIATION RULES

- A set of items is referred to as an **itemset**.An itemset that contains *k items is a k-itemset.*

  The set {computer, antivirus software} is a 2-itemset.

- *The occurrence* frequency of an itemset is the number of transactions that contain the itemset. This is also known, as the **frequency, support count, or count of the itemset**

FREQUENT ITEMSETS, CLOSED ITEMSETS AND ASSOCIATION RULES

- A major challenge in mining frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support (*min sup) threshold, especially when min sup is set low. This is because if an itemset is* frequent, each of its subsets is frequent as well. A long itemset will contain a combinatorial number of shorter, frequent sub-itemsets.

- For example, a frequent itemset of length 100, such as {*a1, a2,..., a100}, contains* ($_1^{100}$) =100 frequent 1-itemsets: *a1, a2, … , a100,*
- ($_2^{100}$) frequent 2-itemsets: (*a1, a2), (a1, a3), …, (a99, a100), and so on. The total number* of frequent itemsets that it contains is thus

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}.$$

FREQUENT ITEMSETS, CLOSED ITEMSETS AND ASSOCIATION RULES

- This is too huge a number of itemsets for any computer to compute or store. To overcome this difficulty,
- Solution: *Mine closed frequent itemset and max-frequent itemset*

- An itemset *X is closed in a data set S if there exists no proper super-itemset Y such* that *Y has the same support count as X in S. An itemset X is a* ***closed frequent itemset*** in set *S if X is both closed and frequent in S. An itemset X is a* ***maximal frequent itemset*** (or max-itemset) in set *S if X is frequent, and there exists no super-itemset Y such that X is a proper subset of Y and Y is frequent in S.*

FREQUENT PATTERNS MINING

Frequent pattern mining can be categorized in many different ways according to various criteria, such as the following:

1. Based on the completeness of patterns to be mined, categories of frequent pattern

mining include mining the *complete set of frequent itemsets, the closed frequent itemsets, the maximal frequent itemsets, and constrained frequent itemsets.*

2. Based on the levels of abstraction and dimensions of data involved in the rule, categories can

include the mining of *single-level association rules,multilevel association rules, singledimensional association rules, and multidimensional association rules.*

3. **Based on the types of values handled in the rule, the categories can include mining**

   Boolean association rules and quantitative association rules.

4. **Based on the kinds of rules to be mined, categories include mining**

   association rules and *correlation rules.*

5. **Based on the kinds of patterns to be mined, frequent pattern mining can be classified**

   into *frequent itemset mining, sequential pattern mining, structured pattern*

EFFICIENT & SCALABLE FREQUENT ITEMSET MINING METHOD

- Many efficient and scalable algorithms have been developed for frequent itemset mining, from which association and correlation rules can be derived. These algorithms can be classified into three categories:

  (1) *Apriori algorithms,*

  (2) frequent pattern growth-based algorithms, such as FP-growth, and

  (3) Algorithms that use the vertical data format.

THE APRIORI ALGORITHM: FINDING FREQUENT ITEMSETS USING CANDIDATE GENERATION

- Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge of frequent itemset properties.*
- TheApriori algorithm is a seminal algorithm for mining frequent itemsets for Boolean association rules. It explores the level-wise mining Apriori property that *all nonempty subsets of a frequent itemset must also be frequent.*