Unit II : ASSOCIATION RULES

Basic Concepts - Market Basket Analysis - Frequent Itemsets, Closed Itemsets and Association Rules - Frequent Itemset Mining Methods – Apriori Algorithm – Generating Association Rules - Frequent pattern growth - **Mining Various Kinds of Association Rules**

Mining Frequent Patterns, Association and Correlations

- Basic concepts and a road map
- Efficient and scalable frequent itemset mining methods
- Mining various kinds of association rules
- Summary

Association Rule Mining: A Road Map

- Boolean vs. quantitative associations (Based on the types of values handled)
  - buys(x, "SQLServer") ^ buys(x, "DMBook") $\rightarrow$ buys(x, "DBMiner") [0.2%, 60%]
  - age(x, "30..39") ^ income(x, "42..48K") $\rightarrow$ buys(x, "PC") [1%, 75%]
- Single dimension vs. multiple dimensional associations (each distinct predicate of a rule is a dimension)
- Single level vs. multiple-level analysis (consider multiple levels of abstraction)
  - What brands of beers are associated with what brands of diapers?
- Extensions
  - Correlation, causality analysis
    - Association does not necessarily imply correlation or causality
  - Maxpatterns (a frequent pattern s.t. any proper subpattern is not frequent) and closed itemsets (if there exist no proper superset c' of c s.t. any transaction containing c also contains c')

What Is Association Mining?

- Association rule mining:
  - A transaction T in a database supports an itemset S if S is contained in T
  - An itemset that has support above a certain threshold, called minimum support, is termed *large* (*frequent*) itemset
  - Frequent pattern: pattern (set of items, sequence, etc.) that occurs frequently in a database
  - Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

Basic Concept: Association Rules

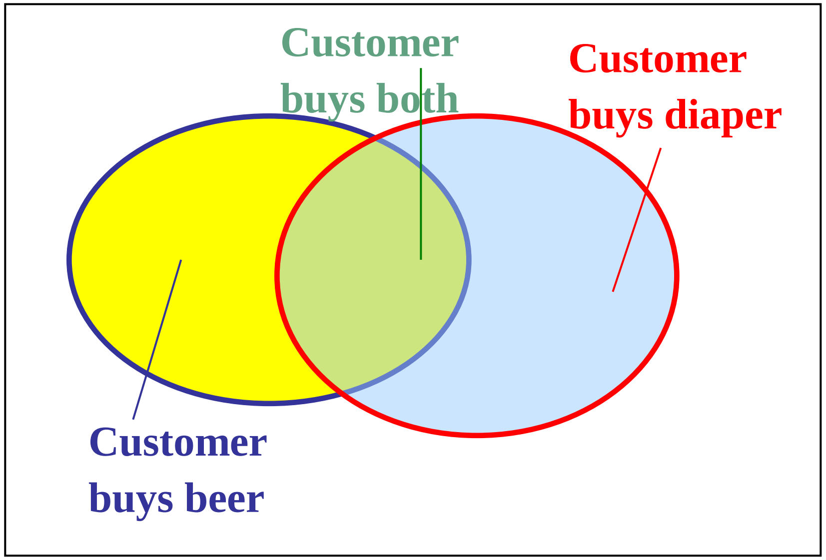
- Let $I = \{i_1, i_2, \ldots, i_n\}$ be the set of all distinct items
- The association rules can be represented as "$A \Rightarrow B$" where A and B are subsets, namely *itemsets*, of I
  - If A appears in one transaction, it is most likely that *B* also occurs in the same transaction

Basic Concept: Association Rules

- For example
  - "Bread ⇒ Milk"
  - "Beer ⇒ Diaper"
- The measurement of interestingness for association rules
  - support, *s*, probability that a transaction contains $A \cup B$
    - s = support("A⇒B") = P(A∪B)
  - confidence, *c*, conditional probability that a transaction having *A* also contains *B*.
    - c = confidence("A⇒B") = P(B|A)

Basic Concept: Association Rules

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F



- Let *min_support* = 50%, *min_conf* = 50%:
  - $A \rightarrow C$ (50%, 66.7%)
  - $C \rightarrow A$ (50%, 100%)

Basic Concepts: Frequent Patterns and Association Rules

- Association rule mining is a two-step process:
  - Find all frequent itemsets
  - Generate strong association rules from the frequent itemsets
    - For every frequent itemset *L*, find all non-empty subsets of *L*. For every such subset *A*, output a rule of the form "$A \Rightarrow (L-A)$" if the ratio of support(*L*) to support(*A*) is at least *minimum confidence*
- The overall performance of mining association rules is determined by the first step

Mining Association Rules—an Example

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

Min. support 50%
Min. confidence 50%

Frequent pattern	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

For rule $A \Rightarrow C$:

support = support($\{A\} \cup \{C\}$) = 50%

confidence = support($\{A\} \cup \{C\}$)/support($\{A\}$) = 66.6%

Mining Various Kinds of Association Rules

- Mining multilevel association
- Miming multidimensional association
- Mining quantitative association
- Mining interesting correlation patterns

Mining Multiple-Level Association Rules

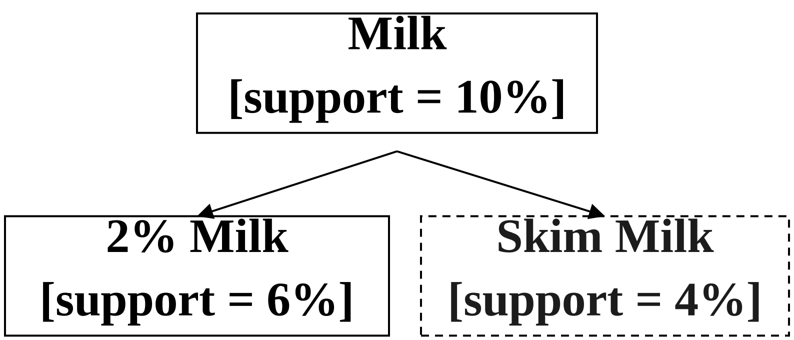
- Items often form hierarchies
- Flexible support settings
  - Items at the lower level are expected to have lower support
- Exploration of *shared* multi-level mining (Agrawal & Srikant@VLB'95, Han & Fu@VLDB'95)

uniform support

reduced support

Level 1
min_sup = 5%

Level 2
min_sup = 5%



Level 1
min_sup = 5%

Level 2
min_sup = 3%

Multi-level Association: Redundancy Filtering

- Some rules may be redundant due to "ancestor" relationships between items.
- Example
  - milk $\Rightarrow$ wheat bread [support = 8%, confidence = 70%]
  - 2% milk $\Rightarrow$ wheat bread [support = 2%, confidence = 72%]
- We say the first rule is an ancestor of the second rule.
- A rule is redundant if its support is close to the "expected" value, based on the rule's ancestor.

Mining Multi-Dimensional Association

- Single-dimensional rules:

  buys(X, "milk") ⇒ buys(X, "bread")

- Multi-dimensional rules: ≥ 2 dimensions or predicates
  - Inter-dimension assoc. rules (*no repeated predicates*)

    age(X,"19-25") ∧ occupation(X,"student") ⇒ buys(X, "coke")
  - hybrid-dimension assoc. rules (*repeated predicates*)

    age(X,"19-25") ∧ buys(X, "popcorn") ⇒ buys(X, "coke")
- Categorical Attributes: finite number of possible values, no ordering among values—data cube approach
- Quantitative Attributes: numeric, implicit ordering among values—discretization, clustering, and gradient approaches

Mining Quantitative Associations

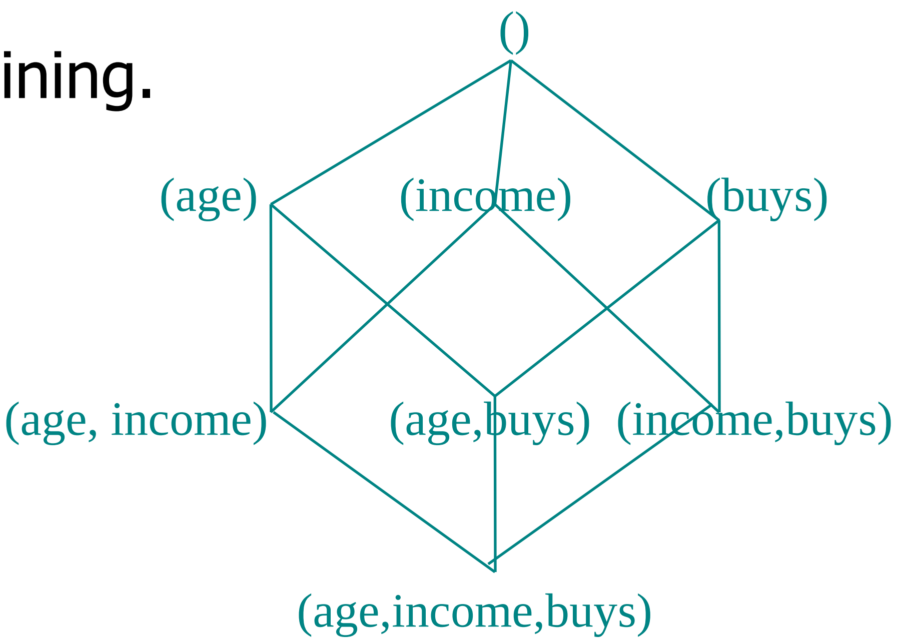
- Techniques can be categorized by how numerical attributes, such as age or salary are treated

1. Static discretization based on predefined concept hierarchies (data cube methods)
2. Dynamic discretization based on data distribution (quantitative rules, e.g., Agrawal & Srikant@SIGMOD96)
3. Clustering: Distance-based association (e.g., Yang & Miller@SIGMOD97)
   - one dimensional clustering then association
4. Deviation: (such as Aumann and Lindell@KDD99)

Sex = female => Wage: mean=$7/hr (overall mean = $9)

Static Discretization of Quantitative Attributes

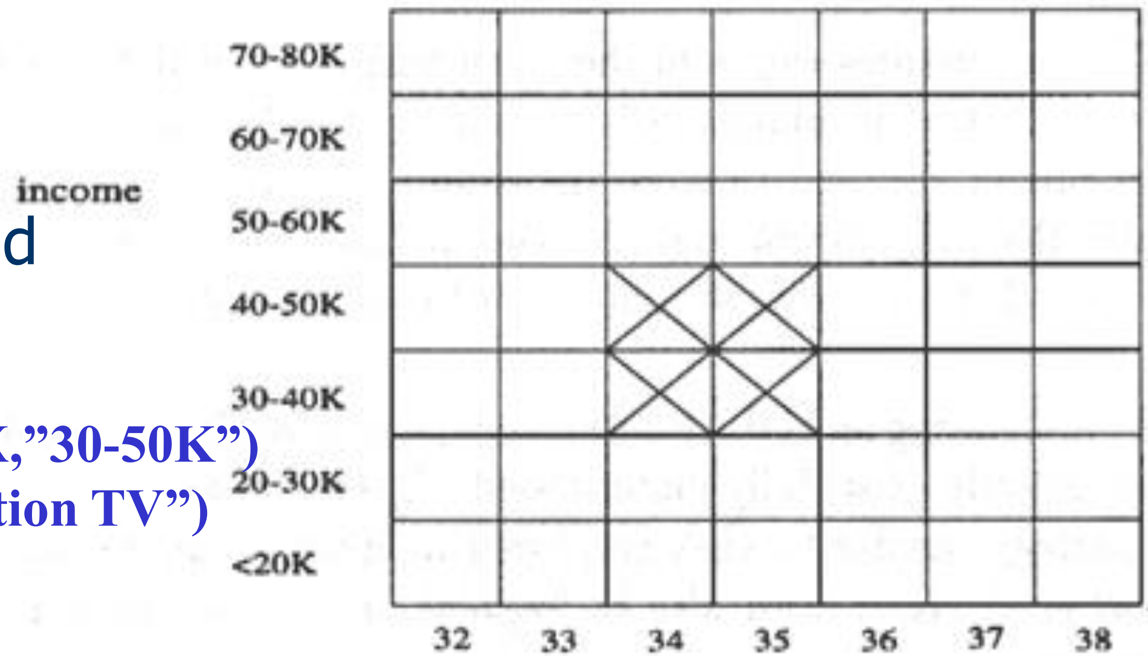
- Discretized prior to mining using concept hierarchy.
- Numeric values are replaced by ranges.
- In relational database, finding all frequent k-predicate sets will require *k* or *k*+1 table scans.
- Data cube is well suited for mining.
- The cells of an n-dimensional cuboid correspond to the predicate sets.
- Mining from data cubes can be much faster.



Quantitative Association Rules

- Proposed by Lent, Swami and Widom ICDE'97
- Numeric attributes are *dynamically* discretized
  - Such that the confidence or compactness of the rules mined is maximized
- 2-D quantitative association rules: $A_{quan1} \wedge A_{quan2} \Rightarrow A_{cat}$
- Cluster *adjacent* association rules to form general rules using a 2-D grid
- Example

age(X,"34-35") ∧ income(X,"30-50K") ⇒ buys(X,"high resolution TV")



Mining class association rules

- Normal association rule mining does not have any target.
- It finds all possible rules that exist in data, i.e., any item can appear as a consequent or a condition of a rule.
- However, in some applications, the user is interested in some targets.
  - E.g, the user has a set of text documents from some known topics. He/she wants to find out what words are associated or correlated with each topic.

An example

- **A text document data set**

  doc 1: Student, Teach, School : Education
  doc 2: Student, School : Education
  doc 3: Teach, School, City, Game : Education
  doc 4: Baseball, Basketball : Sport
  doc 5: Basketball, Player, Spectator : Sport
  doc 6: Baseball, Coach, Game, Team : Sport
  doc 7: Basketball, Team, City, Game : Sport

- Let *minsup* = 20% and *minconf* = 60%. The following are two examples of class association rules:

  Student, School $\rightarrow$ Education [sup= 2/7, conf = 2/2]
  game $\rightarrow$ Sport [sup= 2/7, conf = 2/3]

Mining Other Interesting Patterns

- Flexible support constraints (Wang et al. @ VLDB'02)
  - Some items (e.g., diamond) may occur rarely but are valuable
  - Customized sup_{min} specification and application
- Top-K closed frequent patterns (Han, et al. @ ICDM'02)
  - Hard to specify sup_{min}, but top-k with $length_{min}$ is more desirable
  - Dynamically raise sup_{min} in FP-tree construction and mining, and select most promising path to mine

Frequent-Pattern Mining: Summary

- Frequent pattern mining—an important task in data mining
- Scalable frequent pattern mining methods
  - Apriori (Candidate generation & test)
  - Projection-based (FPgrowth, CLOSET+, ...)
  - Vertical format approach (CHARM, ...)
- Mining a variety of rules and interesting patterns
- Constraint-based mining
- Mining sequential and structured patterns
- Extensions and applications

Frequent-Pattern Mining: Research Problems

- Mining fault-tolerant frequent, sequential and structured patterns
  - Patterns allows limited faults (insertion, deletion, mutation)
- Mining truly interesting patterns
  - Surprising, novel, concise, …
- Application exploration
  - E.g., DNA sequence analysis and bio-pattern classification
  - "Invisible" data mining