Mining Association Rules

——Association and Correlations——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

Association and Correlations

- **Association and Correlations**
- **Efficient and Scalable Frequent Itemset Mining Methods**
- **Mining Various Kinds of Association Rules**
- **From Association Mining to Correlation Analysis**
- **Constraint-based Association Mining**

2

Market-Basket Problem(1)

- A large set of *items*, e.g., things sold in a supermarket.
- A large set of *baskets*, each of which is a small set of the items, e.g., the things one customer buys on one day.
- Simplest question: find sets of items that appear "frequently" in the baskets.
- *Support* for itemset I = the number of baskets containing all items in I.
- Given a support *threshold* s, sets of items that appear in $\geq s$ baskets are called *frequent itemsets*.

3

Market-Basket Problem(2)

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

  B_1 = {m, c, b} B_2 = {m, p, j}

  B_3 = {m, b} B_4 = {c, j}

  B_5 = {m, p, b} B_6 = {m, c, b, j}

  B_7 = {c, b, j} B_8 = {b, c}
- Frequent itemsets: {m}, {c}, {b}, {j}, {m, b}, {c, b}, {j, c}.

4

Potential Applications (1)

- **Real market baskets: chain stores keep terabytes of information about what customers buy together.**
  - **Tells how typical customers navigate stores, lets them position tempting items.**
  - **Suggests tie-in "tricks," e.g., run sale on beer and raise the price of diapers.**
  - **Basket data analysis, cross-marketing, catalog design, sale campaign analysis**

Potential Applications (2)

- **"Baskets" = reviews; "items" = words in those crawled information from Internet.**
  - **Let us find review hotspots that appear together frequently, i.e., *review sentiment analysis*.**
- **"Baskets" = credit card bills, "items" = transactions in these business bank database.**
  - **Items that appear together too often could represent customers' consumption patterns, i.e. consumer behavior analysis.**
- **"Baskets" = Web pages; "items" = browsed pages.**
  - **Items that appear together too often could represent net citizens' browsing patterns, i.e. *Internet Behavior Analysis*.**

Important Hints

- "Market Baskets" is an abstraction that models any many-many relationship between two concepts: "items" and "baskets."
  - Items need not be "contained" in baskets.
- The only difference is that we count co-occurrences of items related to a basket, not vice-versa.
- Scale of Problem
  - WalMart sells 100,000 items and can store billions of baskets.
  - The Web has over 100,000,000 words and billions of pages.

7

What Is Frequent Pattern Analysis?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- 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(成簇)

8

Frequent Patterns and Association Rules (1)

- Itemset X = $\{x_1, \ldots, x_k\}$
- Find all the rules *X* → *Y* with minimum support and confidence
  - **support**, *s*, probability that a transaction contains X ∪ Y

$$s = \frac{(X \cup Y).Count}{n}$$

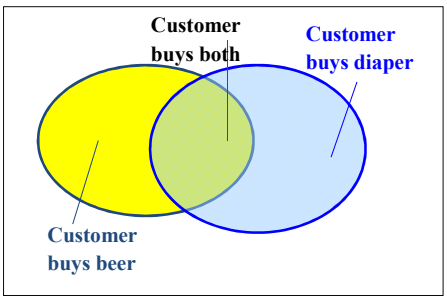
  - **confidence**, *c*, conditional probability that a transaction having X also contains *Y*

$$c = \frac{(X \cup Y).Count}{X.Count}$$

Frequent Patterns and Association Rules (2)

- *Let supmin = 50%, confmin = 50%*
- *Freq. Pat.: {A:3, B:3, D:4, E:3, AD:3}*
- Association rules:
  - *A* → *D* (60%, 100%)
  - *D* → *A* (60%, 75%)

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



Closed Patterns and Max-Patterns

- A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \ldots, a_{100}\}$ contains $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{1}{1}\binom{0}{0}\binom{0}{0} = 2^{100} - 1 = 1.27*10^{30}$ 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 ↄ X, *with the same support* as X (proposed by Pasquier, et al. @ ICDT' 99)
- An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ↄ X (proposed by Bayardo @ SIGMOD' 98)
- Closed pattern is a lossless compression of freq. patterns
  - Reducing the number of patterns and rules

11

Thanks !

12