

关联规则的挖掘

从交易数据、关系数据库以及其它的数据集中发现项和对象的频繁的模式(frequent patterns)、关联(associations)的过程

关联分析的应用

交叉销售

关联规则

- 规则 (Rule) :

$$\{x_1, x_2, x_3, \dots, x_n\} \rightarrow Y$$

- 可信度 (Confidence) 和最小可信度
 - 购买 x_1, x_2, \dots, x_n 的情况下购买Y的可能性, 条件概率
 - $Confidence(A \rightarrow B) = P(B | A)$
- 支持度 (Support) 和最小支持度
 - 同时购买 x_1, x_2, \dots, x_n 和Y的可能性
 - $Support(A \rightarrow B) = P(A \cup B)$
- 频繁项目集

满足最小支持度的项目集

$$confidence(X \rightarrow Y) = P(Y|X) = \frac{support(XY)}{support(X)}$$

Scalable Methods For Mining Frequent Pattern

Apriori

- 规则: 若存在某些项集是不平凡的, 那么我们就没必要生成他们的超集对他们进行检验与测试
- **Get Frequent itemset**
 - How to generate candidates(minsup)
 - p and q are two itemsts in L_k
 - if $p_{item1} = q_{item1}, p_{item2} = q_{item2}, \dots, p_{itemk-1} = q_{itemk-1}, p_{itemk} \neq q_{itemk}$
 - combine them to generate C_{k+1}
 - Delete those (k+1) length itemsets which include infrequent k length itemsets
- **Get Association Rule**

- For each frequent itemset l , generate every non-empty subset S , if satisfied
- $confidence((l - S) \rightarrow S) = P(Y|X) = \frac{support(l)}{support(l - S)} \geq minconf$
 - then, we can get the association rule $(l - S) \rightarrow S$

FPgrowth

Compare

Apriori : use a generate-and-test approach generates candidate itemsets and tests if they are frequent

FP-Growth : all frequent itemsets discovery without candidate generation

- Benefits of **FP-tree Structure**
 - Completeness
 - Preserve complete information for frequent pattern mining
 - Never break a long pattern of any transaction
 - Compactness
 - Reduce irrelevant info
 - (Items in frequency descending order) the more frequently occurring, the more likely to be shared
 - Never be larger than the original database

Why is FP-Growth the Winner

- Divide - and - conquer
 - decompose both the mining task and database according to the frequent pattern obtained so far
- no candidate generation, no candidate test
- compressed database
- no repeated scan of entire database

Notice!

Support and **confidence** are not good to represent **correlation**

Lift(增益、提升度)

$$lift = \frac{P(A \cup B)}{P(A)P(B)} = \frac{conf(A \rightarrow B)}{sup(B)}$$

$$lift \begin{cases} > 1 & \text{positively correlated} \\ = 1 & \text{independent} \\ < 1 & \text{negatively correlated} \end{cases}$$

Closed Pattern and Max-Pattern

1. An itemset X is a **max-pattern** if X is frequent and there exists no frequent super-pattern $Y \supset X$ (proposed by Bayardo @ SIGMOD'98)

Closed pattern is a **lossless compression** of frequent patterns