EE226 Big Data Mining Lecture 3

Basic Data Mining Algorithms

Liyao Xiang
http://xiangliyao.cn/
Shanghai Jiao Tong University

http://jhc.sjtu.edu.cn/public/courses/EE226/

Notice

- There will be a quiz in the next week’s class. Please take a piece of paper and pens.

Reference and Acknowledgement

- Most of the slides are credited to Prof. Jiawei Han's book "Data Mining: Concepts and Techniques."

Outline

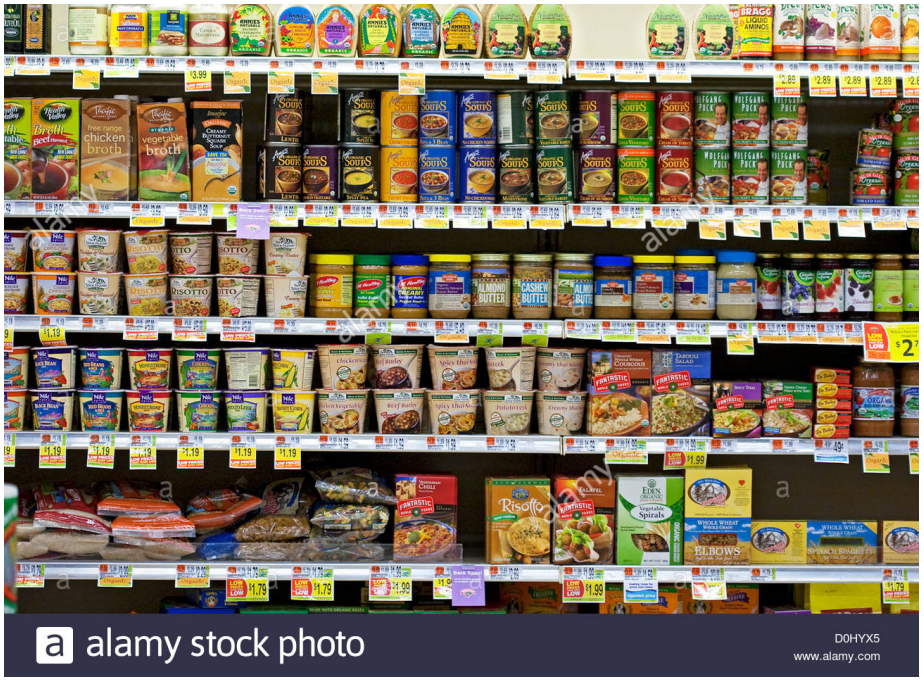
- Basic Concepts in Frequent Pattern Mining
- Frequent Itemset Mining Methods
- Pattern Evaluation Methods

Outline

- Basic Concepts in Frequent Pattern Mining
- Frequent Itemset Mining Methods
- Pattern Evaluation Methods

Basic Concepts

- Frequent pattern: a pattern (a set of items, subsequences, substructures …) that appear frequently in a database



- Finding frequent patterns is key to mining associations, correlations, clustering, classification and other relationships among data.
- Applications: basket data analysis, cross-marketing, catalog design …

Basic Concepts

- itemset: a set of one or more items
- k-itemset: $X = \{x_1, \ldots, x_k\}$
- (absolute) support, or support count of X: frequency or occurrence of an itemset X
- (relative) support: the fraction of transactions that contains X over all transaction
- An itemset X is frequent if X's support is no less than a defined threshold min_sup

TID	Items Purchased
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Basic Concepts

- support: probability that a transaction contains X∪Y

  support(X⇒Y) = P(X∪Y)

- confidence: conditional prob. that a transaction having X also contains Y

  confidence(X⇒Y) = P(Y|X)

$$P(Y|X) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$$

TID	Items Purchased
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Basic Concepts

- min_sup: minimum support threshold
- min_conf: minimum support confidence threshold
- e.g., find all rules X ⇒ Y with min_sup and min_conf

let min_sup = 50%, min_conf = 50%

frequent pattern: Beer: 3, Nuts: 3, Diaper: 4, Eggs: 3, {Beer, Diaper}: 3

- Association rules:

  Beer⇒Diaper (60%, 100%)

  Diaper⇒Beer (60%, 75%)

TID	Items Purchased
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Basic Concepts

- Association rule mining includes:
  1. Find all frequent itemsets: frequency of itemsets ≥ min_sup
  2. Generate strong association rules from the frequent itemsets
- 1 is the major step, but challenging in that there may be a huge number of itemsets satisfying min_sup
- An itemset is frequent ⇒ each of its subsets is frequent
- Solution: mine closed frequent itemset and maximal frequent itemset
- closed frequent itemset X: X is frequent and there is no super-itemset Y ⊃ X with the same support count as X
  - closed frequent itemset is a lossless compression of frequent itemset
- maximal frequent itemset X: X is frequent and there is no super-itemset Y ⊃ X which is frequent

Basic Concepts

- e.g., {<a_1, …, a_{100}>, < a_1, …, a_{50}>}, min_sup = 1
- What is the set of closed frequent itemset?
  - <a_1, …, a_{100}>: 1, < a_1, …, a_{50}>: 2
- What is the set of maximal frequent itemset?
  - <a_1, …, a_{100}>: 1
- We can assert <a_2, a_{45}> is frequent since a_2, $a_{45} \in$ < a_1, …, a_{50}> but cannot assert their actual support count
- How many itemsets are potentially to be generated in the worst case?
  - When min_sup is low, there exist potentially an exponential number of frequent itemsets
  - Worst case: M^N where M = # distinct items, N = max length of transactions

Summary

- frequent pattern
- k-itemset
- (absolute) support, support count, relative support
- min_sup, confidence
- closed frequent itemset, maximal frequent itemset

Outline

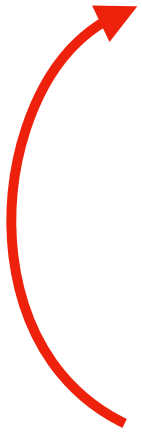
- Basic Concepts in Frequent Pattern Mining
- Frequent Itemset Mining Methods
- Pattern Evaluation Methods

Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- Improving the Efficiency of Apriori
- FP-Growth: A Frequent Pattern-Growth Approach
- ECLAT: Frequent Pattern Mining with Vertical Data Format

Apriori

- Downward Closure Property: any subset of a frequent itemset must be frequent
  - e.g., if {beer, diaper, nuts} is frequent, so is {beer, diaper} since every transaction having {beer, diaper, nuts} also contains {beer, diaper}
- Apriori employs a level-wise search where k-itemsets are used to explore (k + 1)-itemsets. Steps:



1. Scan database once to get frequent 1-itemsets L_1
2. Join the k-frequent itemsets L_k to generate length (k+1) candidate itemsets C'_{k+1}
3. Prune C'_{k+1} against the database to get C_{k+1}
4. Scan (Test) database for the count of each candidate in C_{k+1}, obtain L_{k+1}
5. Terminate when no frequent or candidate set can be generated