Chapter 3: Frequent Itemset Mining & Association Rules

崔金华
电子邮箱: jhcui@hust.edu.cn
个人主页: https://csjhcui.github.io/

致谢:课件主要参考《 Mining of Massive Datasets》(Second Edition) J. Leskovec, A. Rajaraman, J. Ullman的相关课件, 特此致谢!!!

3.1.1 Rules

The Hobby of Detectives!

3.1.1 Association Rule Discovery

Supermarket shelf management – Market-basket model:

- ❑ **Goal:** Identify items that are bought together by sufficiently many customers
- ❑ **Approach:** Process the sales data collected with barcode scanners to find dependencies among items
- ❑ **A classic rule:**
  - ➢ If someone buys diaper and milk, then he/she is likely to buy beer
  - ➢ Don’ t be surprised if you find six-packs next to diapers!

3.1.1 The Market-Basket Model

- ❑A large set of **items(项)**
  - ➢e.g., things sold in a supermarket, bread, coke…

- ❑A large set of **baskets(购物篮)**
  - ➢Each basket is a **small subset of items(多个项的集合, 称作项集)**
  - ➢e.g., the things one customer buys on one day

Input:

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

- ❑Want to discover **association rules(关联规则)**
  - ➢People who bought {x,y,z} tend to buy {v,w}
    - 永辉、中百仓储…

Output:

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

3.1.2 Applications – (1)

- **Items** = products; **Baskets** = sets of products someone bought in one trip to the store
- **Real market baskets:** Chain stores keep TBs of data 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 diapers and raise the price of beer
  - Need the rule to occur frequently, or no $$’s
- **Amazon’s people who bought *X* also bought *Y***

3.1.2 Applications – (2)

- **Baskets** = sentences; **Items** = documents containing those sentences
  - Items that appear together too often could represent plagiarism (剽窃, 文档抄袭)
  - Notice items do not have to be “in” baskets

- **Baskets** = patients; **Items** = drugs & side-effects
  - Has been used to detect combinations of drugs that result in particular side-effects
  - **But requires extension:** Absence of an item needs to be observed as well as presence

3.1.2 More generally applications

- **A general many-to-many mapping (association) between two kinds of things**
  - But we ask about connections among “items” , not “baskets”

- **For example:**
  - Finding communities in graphs (e.g., Twitter)

Outline

- **First: Define**
  - Section 3.2: Frequent itemsets, Association rules
- **Then: Algorithms for finding frequent itemsets**
  - Section 3.3: Finding frequent itemsets
  - Section 3.4: A-Priori algorithm
  - Section 3.5: PCY algorithm
  - Section 3.6: 2 refinements
  - Section 3.7: Frequent Itemsets in <= 2 Passes

Section 3.2: Frequent itemsets, Association rules

Content

1 Frequent itemsets

2 Association rules

3 Compacting Frequent Itemsets Output

3.2.1 Frequent Itemsets

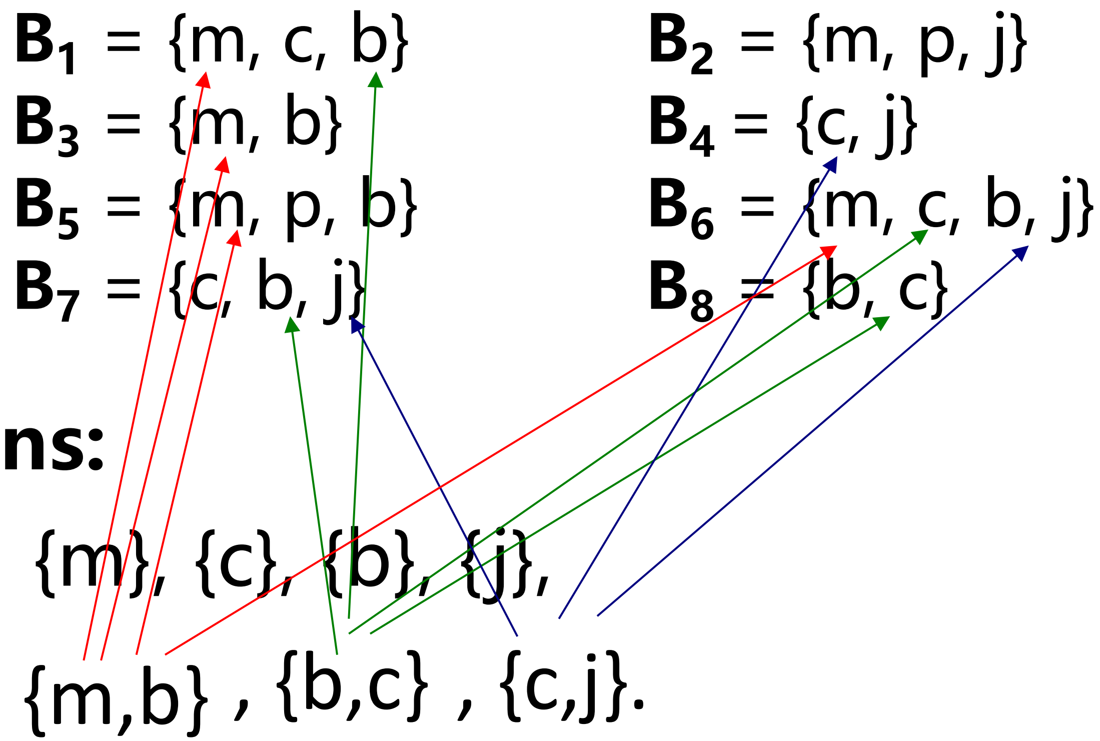
- **Simplest question:** Find sets of items that appear together "frequently" in baskets
- ***Support* (支持度)** for itemset ***I*:** Number of baskets containing all items in ***I***
  - (Often expressed as a fraction of the total number of baskets)
- Given a ***support threshold s***, then sets of items that appear in at least ***s*** baskets are called ***frequent itemsets*** **（频繁项集）**

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Support of {Beer, Bread} = 2

3.2.1 Example: Frequent Itemsets

- **Items** = {milk, coke, pepsi, beer, juice}
- **Find frequent itemsets, support threshold** = 3 baskets



$\mathbf{B_1}$ = {m, c, b} $\mathbf{B_2}$ = {m, p, j}

$\mathbf{B_3}$ = {m, b} $\mathbf{B_4}$ = {c, j}

$\mathbf{B_5}$ = {m, p, b} $\mathbf{B_6}$ = {m, c, b, j}

$\mathbf{B_7}$ = {c, b, j} $\mathbf{B_8}$ = {b, c}

Note: m for milk,
c for coke, p for pepsi,
b for beer, j fo r juice

- **Ans:**

{m}, {c}, {b}, {j},

{m,b} , {b,c} , {c,j}.

3.2.2 Association Rules

- **Association Rules(关联规则):** If-then rules about the contents of baskets
  - $\{i_1, i_2,\ldots,i_k\} \rightarrow j$ means: “if a basket contains all of $i_1,\ldots,i_k$ then it is ***likely*** to contain j “
  - **In practice, there are many rules, we only want to find significant/interesting ones!**
- ***Confidence* (置信度, 可信度)** of this association rule is the probability of j given $I = \{i_1,\ldots,i_k\}$:

$$conf(I \rightarrow j) = \frac{support(I \cup j)}{support(I)}$$

3.2.2 Interesting Association Rules

- **Not all high-confidence rules are interesting**
  - The rule $X \rightarrow \boldsymbol{milk}$ may have high confidence for many itemsets X, because milk is just purchased very often (independent of X) and the confidence will be high
- **Interest (兴趣度)** of an association rule $I \rightarrow j$, difference between its confidence and the fraction of baskets that contain j:

$$\boldsymbol{Interest}(I \rightarrow j) = \boldsymbol{conf}(I \rightarrow j) - \Pr[j]$$

  - Interesting rules are those with high positive or negative interest values (usually above 0.5): {diapers}->beer, {coke}->pepsi

3.2.2 Example: Confidence and Interest

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}

Note: m for milk,
c for coke, p for pepsi,
b for beer, j fo r juice

- **Association rule: {m, b} →c, how much interest does it have?**
  - **Confidence** = 2/4 = 0.5
  - **Interest** = |0.5 – 5/8| = 1/8
    - Item *c* appears in 5/8 of the baskets
    - Rule is not very interesting!

$$conf(I \rightarrow j) = \frac{support(I \cup j)}{support(I)}$$

$$Interest(I \rightarrow j) = conf(I \rightarrow j) - \Pr[j]$$

3.2.2 Finding Association Rules

- **Problem: Find all association rules with support ≥ s and confidence ≥ c**
  - **Note:** $support(I \rightarrow j) = support(I)$. Support of an association rule is the support of the set of items on the left side
- **Hard part: Finding the frequent itemsets!**
  - If $\{i_1, i_2,\ldots, i_k\} \rightarrow j$ has high support and confidence, then both $\{i_1, i_2,\ldots, i_k\}$ and $\{i_1, i_2,\ldots,i_k, j\}$ will be "frequent".
  - Why?

$$conf(I \rightarrow j) = \frac{support(I \cup j)}{support(I)}$$

$$Interest(I \rightarrow j) = conf(I \rightarrow j) - \Pr[j]$$