Frequent Pattern Analysis

Data Mining:
Data Mining Methods
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Apply techniques for frequent pattern mining and explain how they work. Identify association and correlation.

Apriori Algorithm: Challenges

- Multiple scans of the whole dataset
- A huge number of candidates
- Support counting of all candidates

- Can we make it more efficient?

Possible Improvements

➢ Partitioning

- A freq. itemset must be freq. in at least one partition

➢ Sampling

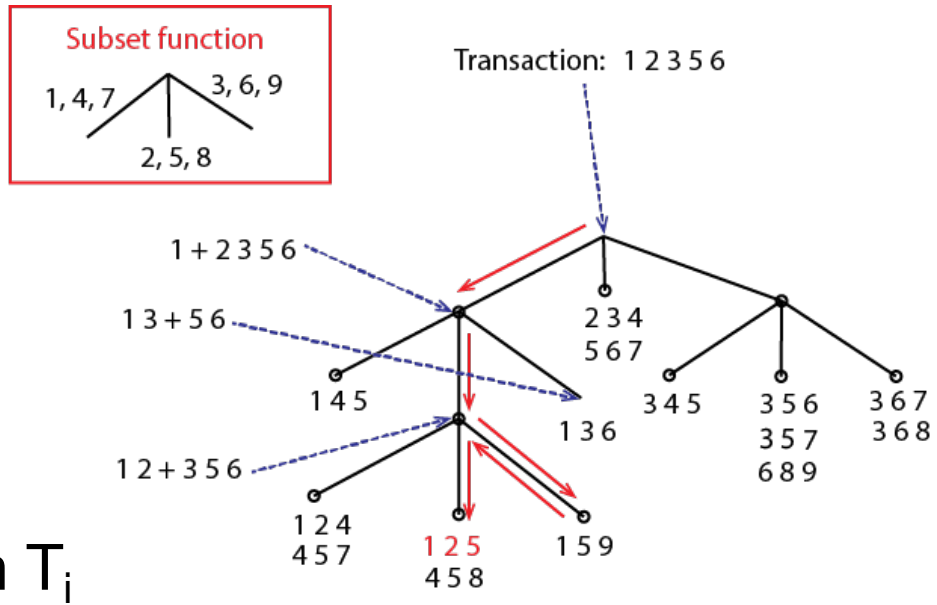
- Sampled subsets are likely to contain freq. itemsets

➢ Transaction reduction

- Remove T_i if it doesn't contain any freq. k-itemset

Support Counting using Hash-tree

- **Subset function**
  - Item-specific branching
- **Leaf nodes**
  - Candidate k-itemsets
- **Transaction T_i**
  - All possible k-itemsets in T_i



Vertical Data Format

- Horizontal data format
  - E.g., $T_i = \{A, C, E\}$
- Vertical data format
  - E.g., $t(X) = \{T_2, T_5, T_9\}$, $t(Y) = \{T_2, T_3, T_5, T_{10}\}$
- Vertical intersection
  - E.g., $t(XY) = t(X) \cap t(Y) = \{T_2, T_5\}$

Analogy: Search Engine
- documents
- keywords
- inverted index

Avoid Candidate Generation

➢ Limitations of Apriori algorithm

- Multiple scans, candidate generation, support counting
- Can we avoid candidate generation?

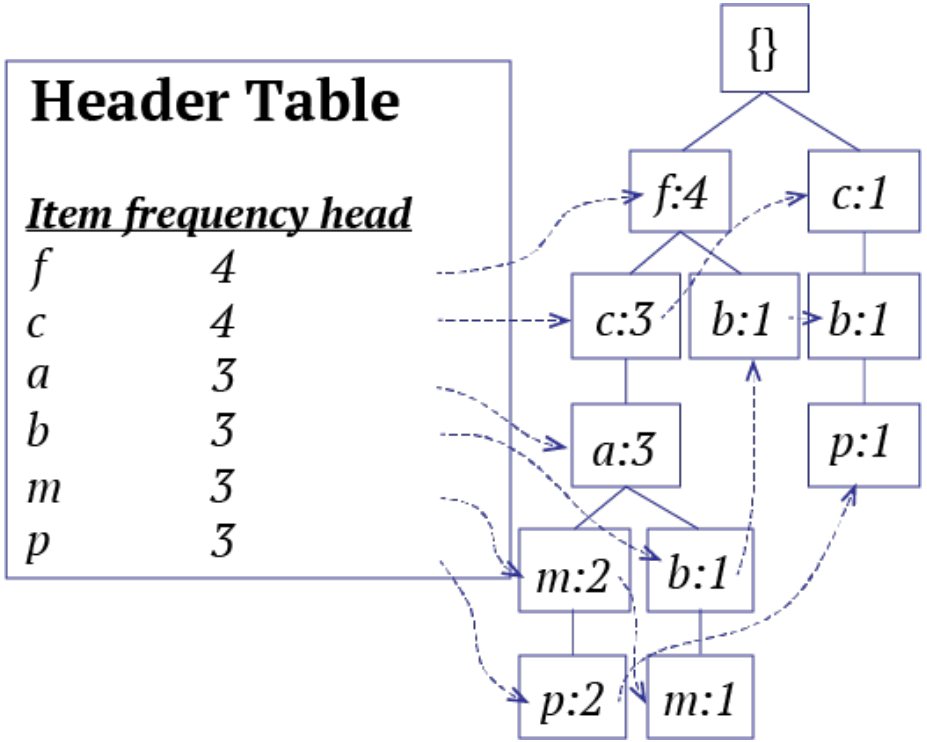
➢ FP-growth algorithm

- Intuition: if d is freq. in DB | abc, then abcd is freq.
- Grow patterns using local freq. items
- Find freq. itemsets without candidate generation

FP-tree Construction

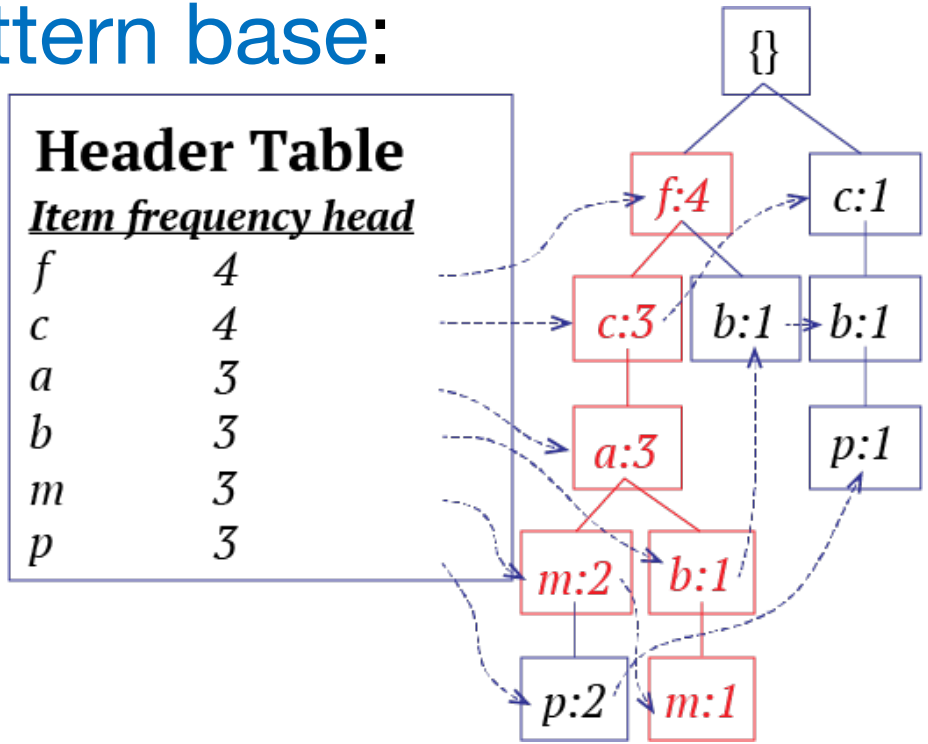
- min_sup = 0.6

	Items	Sorted
1	{a, c, d, g, f, m, p, x}	{f, c, a, m, p}
2	{a, b, c, f, m, o x}	{f, c, a, b, m}
3	{b, f, h, j, o, w}	{f, b}
4	{b, c, k, p, s}	{c, b, p}
5	{a, c, e, f, m, n, p, x}	{f, c, a, m, p}



Conditional FP-tree

- m-conditional pattern base:
  fca:2, fcab:1
  - m
  - fm, cm, am
  - fcm, fam, cam
  - fcam



Association Rule

- Given list of transactions, itemsets X, Y
- Association rule: X => Y
  - Support: P(X U Y)
  - Confidence: P(Y | X)
  - Minimum support, minimum confidence

Association Rule Example

- min_sup = 0.5, min_conf = 0.6
- Freq. itemsets
  - B: 4, E: 5, …
- Association rules
  - B => E (sup = ? , conf = ?)
  - E => B (sup = ? , conf = ?)

Tid	Items
1	A, B, C, E
2	A, D, E
3	B, C, E
4	B, C, D, E
5	B, D, E

Correlation

- Numerical attributes: correlation coefficient

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

- Nominal attributes: chi-square test

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

Correlation Rule

- A => B [support, confidence, correlation]
  - If A occurs, is B more (or less) likely to occur?
  - P(B) vs. P(B | A)
- Measure of dependent/correlated events
- lift = 1: independent
- lift > 1: positive
- lift < 1: negative

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

Correlation Example: Lift

$$lift(biking, skiing) = \frac{500/1000}{(700/1000)\times(600/1000)} = 1.19$$

$$lift(biking, not\ skiing) = \frac{200/1000}{(700/1000)\times(400/1000)} = 0.71$$

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

	skiing	not skiing	sum (row)
biking	500	200	700
not biking	100	200	300
sum (col)	600	400	1000

Correlation Example: X^2

$$e_{biking,skiing} = \frac{700 \times 600}{1000} = 420$$

X^2 = 127 > 10 (correlated)
o_{bs} = 500 > e_{bs} = 420 (positive)

$$\chi^2 = \frac{(500-420)^2}{420} + \frac{(200-280)^2}{280} + \frac{(100-180)^2}{180} + \frac{(200-120)^2}{120} = 127$$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

	skiing	not skiing	sum (row)
biking	500 (420)	200 (280)	700
not biking	100 (180)	200 (120)	300
sum (col)	600	400	1000

Other Correlation Measures

$$all\_conf(A,B) = \frac{sup(A \cup B)}{max\{sup(A), sup(B)\}} = min\{P(A|B), P(B|A)\}$$

$$max\_conf(A, B) = max\{P(A \mid B), P(B \mid A)\}$$

$$Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$$

$$cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{sup(A \cup B)}{\sqrt{sup(A) \times sup(B)}}$$

$$= \sqrt{P(A|B) \times P(B|A)}.$$

Correlation Analysis

- Multiple measures to consider
- Null transaction (i.e., not A and not B)
  - Null-variant: lift, X^2
  - Null-invariant: all_conf, max_conf, Kulc, cosine
- Imbalance ratio

$$IR(A,B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

Frequent Pattern Analysis

- Patterns: itemset, sequence, structure
- Rules: association, correlation, gradient
- Dimensions, levels: single, multiple
- Values: binary, categorical, quantitative
- Metarule-guided mining

- $P_1 \wedge P_2 \wedge \ldots \wedge P_x \Rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_y$