Data Mining (CS 451)

JAN-MAY’ 18

PRAGYA VERMA

RECAP

- Frequent Itemset Mining Methods
  - Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation
  - Generating Association Rules from Frequent Itemsets
- Improving the Efficiency of Apriori

CONTENT

- A Pattern-Growth Approach for Mining Frequent Itemsets
- Mining Frequent Itemsets Using the Vertical Data Format
- Pattern Evaluation Methods

A Pattern-Growth Approach for Mining Frequent Itemsets

- The keys drawbacks associated with Apriori algorithm are:

1. It needs to generate a huge number of candidate sets
2. It may need to repeatedly scan the whole database and check a large set of candidates by pattern matching.

- Can we design a method that mines the complete set of frequent itemsets without such a costly candidate generation process?
- An interesting method in this attempt is called frequent pattern growth, or simply FP-growth.

Pattern-Growth Approach (Contd.)

- Frequent pattern growth, or simply FP-growth, adopts a divide-and-conquer strategy.
- First, it compresses the database representing frequent items into a frequent pattern tree, or FP-tree, which retains the itemset association information.
- It then divides the compressed database into a set of conditional databases, each associated with one frequent item or “pattern fragment” and mines each database separately.
- For each “pattern fragment”, only its associated datasets need to be examined.

Mining Frequent Itemsets Using the Vertical Data Format

- Both the Apriori and FP-growth methods mine frequent patterns from a set of transactions in TID-itemset format (i.e., {TID: itemset}), where TID is a transaction ID and itemset is the set of items bought in transaction TID.
- This is known as the horizontal data format.
- Alternatively, data can be presented in item-TID_set format (i.e., {item: TID_set}), where item is an item name, and TID_set is the set of transaction identifiers containing the item.
- This is known as the vertical data format.

Pattern Evaluation Methods

- Strong rules may not necessarily be interesting.
- Correlation between different itemsets can be examined in order to find out whether the association rules are interesting or not.
- $lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$
- $all\_conf(A,B) = \frac{\sup(A \cup B)}{\max\{\sup(A), \sup(B)\}} = \min\{P(A|B), P(B|A)\}$
- $\max\_conf(A,B) = \max\{P(A|B), P(B|A)\}$

Next Class

- Introduction to Clustering