Lesson 9

Frequent Itemsets and Association Rule Mining

Frequent Itemset

- Refers to a set of items that frequently appear together, for example, Python and Big Data Analytics when the students of computer science frequently chose these subjects for in-depth studies
- Frequent Itemset (FI) refers to a subset of items that appears frequently in the datasets

Frequent Itemset Mining (FIM)

- Refers to a data mining method which helps in discovering the itemsets that appear frequently in a dataset
- Finding a set of students who frequently show poor performance in semester examinations

Frequent Itemset Mining (FIM)

- Is a Frequent subsequence mining
- A sequence of patterns that occurs frequently
- For example, purchasing a football follows purchasing of sports kit

Frequent substructure Mining (FIM)

- Refers to finding different structural forms, such as graphs, trees or lattices, which may be combined with itemsets or subsequences
- Provides the knowledge of important pairs of items that occur much more frequently than the items bought independently

FIM Algorithm

- A technique to extract knowledge from data
- Extracts on frequently occurring entities, events, …
- Finds the regularities in data

FIM Algorithm

- Specifies a given minimum frequency threshold for considering an itemset as frequent
- The extraction generally depends on the specified threshold

FIM Algorithm

- Is preceding step to the association rule learning (mining) algorithm
- For example, customers of supermarkets, mail order companies and online shops use FIM to find a set of products that are frequently bought together (association)

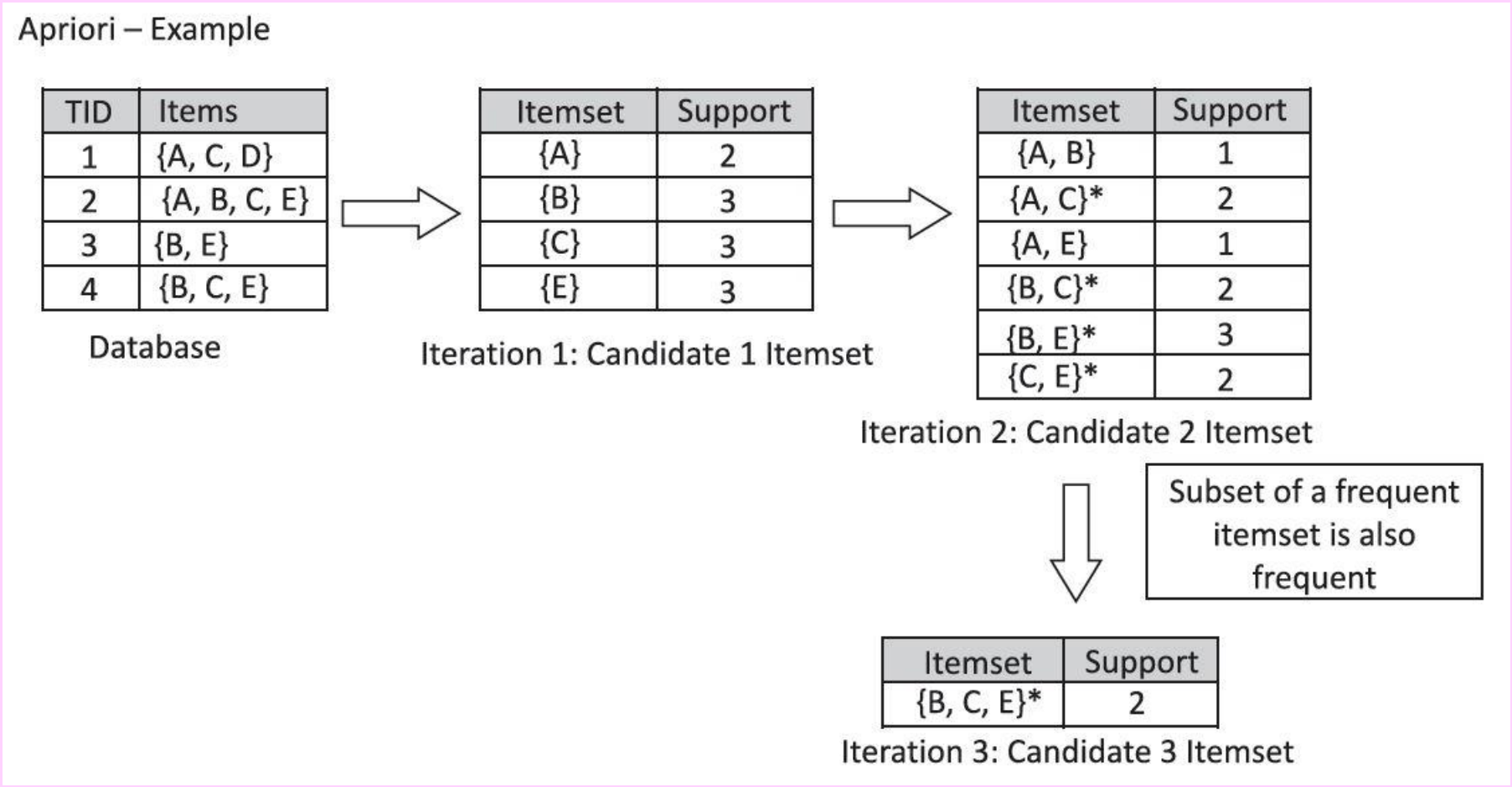
Apriori principle

- Suggests if an itemset is frequent, then all of its subsets must also be frequent
- For example, if itemset {A, B, C} is a frequent itemset, then all of its subsets {A}, {B}, {C}, {A, B}, {B, C} and {A, C} must be frequent.

… Apriori principle

- On the contrary, if an itemset is not frequent, then none of its supersets can be frequent. (Superset means a set consisting of the members which includes the itemsets in the subsets)
- This results into a smaller list of potential frequent itemsets (FIs) as the mining progresses.

Figure 6.8:
Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset.



Steps 1 and 2

- Step 1 (Database): Assign TIDs for the subsets. TID means Term ID, for example, TID is 3 for {B, E} in figure
- Step 2 (Iteration 1): Find for each itemset A, B, C, … number of TIDs supporting (including) that itemset, for example {B} is in the three TIDs

Step 3

- Step 3 (Iteration 1): Find for each pair of combinations of itemset A, B, C, … number of TIDs supporting (including) that itemset, for example {B} in three TIDs. For example, {B, E} has support in three TIDs (2, 3 and 4)

Step 4

- Step 4 (Apply Apriori Principle): {B, E} is frequent 3 times.

C is also present three times in {A, C}, { B, C}, { C, E}

Thus, Aprori principle of subset of an FI is also an FI.

Algorithm Result

- Iteration 1: Candidate 1: {B}, {C} and {E}
- Iteration 2: Candidate 2: {B, E}, {C}; each three times in TIDs
- Final: {B, C, E} is frequent items set, whose subsets: {B, E} and {C} are the candidates of interest
-

FI Analysis Applications

- Improvement of arrangement of products in shelves and on catalog pages
- Marketing and sales promotion

FI Analysis Applications

- Planning of products that a store should stock up
- Support cross-selling (suggestion of other products) and product bundling..

Association Rule

- FIM method has been widely used in many application areas for discovering interesting relationships which are present in large datasets
- The objective is to find uncovered relationships using some strong rules

Association Rules for frequent itemsets

- Mahout includes a ‘parallel frequent pattern growth’ algorithm
- The method analyzes the items in a group and then identifies which items typically appear together (association)

Formal statement of the association rule finding problem

- Let $\mathcal{I} = \{I1, I2, \ldots, Id\}$ be a set of d distinct attributes, also called literals
- Let $\mathcal{T} = \{t1, t2, \ldots, tn\}$ be set of n transactions and contain a set of items such that $\mathcal{T} \subseteq \mathcal{I}$
- $\subseteq$ means a subset of and $\subset$ means proper (strict) subset

Formal statement of the association rule finding problem

- An association rule is an implication of the form, $X \rightarrow Y$, where X, Y belong to sets of items called itemsets ($X, Y \subset I$), and X and Y are disjoint itemsets ($X \cap Y = \emptyset$).

Explanation

- ∩ means intersection
- ∅ means disjoint (no common members).
- Here, X is called antecedent, and Y consequent.

Association Rule Form

- if ( ) then ( ) form (Condition)
- ‘If’ part is called antecedent
- ‘Then’ part is called consequent (Result)

Association Rule Form

- If-then rules about the contents of baskets: $\{p1, p2 \ldots, pk\} \rightarrow q$ *means*,
- "If a basket contains all of *p1, p2…, pk* then it is likely to contain q."

Application Consumer Behaviour

- . If people tend to buy two products (say A and B) together, then the buyer of product A is a potential customer for an advertisement of product B

Applications of Association Rules

- Market Basket Model
- Analysis examples: knowledge discovery about co-occurrence of items. to derive the strength of association between pairs of product items.
- Amazon sells more than 12 million products and can store hundreds of millions of baskets.

Medical Analytics

- Medical analytics: Market basket analysis can be used for conditions and symptom analysis. This helps in identifying a profile of illness in a better way.

Web usage analytics

- Association rules can be exploited to learn about:
- Website browsing of visitor's behavior,
- Developing website structure by making it more effective for visitors
- Improving web marketing promotions.

Summary

We learnt:

- Frequent Itemsets Mining
- Apriori Algorithm
- Association Rule
- If ( ) Then ( ) form
- Antecedent and Consequent (Condition and result)
-

Summary

We learnt:

- Applications of Frequent Itemsets
- Market Basket Analysis
- Medical Analytics
- Website Visitors Analytics

End of Lesson 9 on Frequent Itemsets and Association Rule Mining