Priya Paradhi
4822

Assignment 5

Page :
Date : / /

Q.1) Explain Apriori algorithm with example.

1) It is powerful algorithm for mining frequent itemsets for Boolean association rules.

2) Name of the algorithm is based on the priority it uses i,e Apriori.

- Apriori property based on fact that it uses prior knowledge of frequent itemset properties.
- This property uses iterative approach as level wise search.
- At any level k itemsets are used to explore (k+1) itemsets.
- At first steps whole database is scanned & count of each individual item is found. Assume minimum support.
- Consider those items which satisfy minimum support. Set of such frequent itemset is found.
- The resulting set is denoted as L1 (level 1)
- L1 is used to find L2 (L2 is the frequent itemsets 2).
- L2 is used to find L3 & the process continues till no more frequent k-itemsets can be found.
- Every time database has to be scanned for find Lk frequent itemset.

Apriori property:

- Any subset of a large itemset must be large.
- OR All non empty subsets of a frequent itemset must also be format frequent.
- Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore k+1-itemsets.
- initially, scan DB once to get frequent 1-itemset.
- Generate length (k+1) candidate itemsets from length k frequent itemsets.
- Test the candidates against DB.
- Terminate when no frequent or candidate set can be generated.

Page:
Date: / /

- Apriori Pruning Principle: If there is any itemset which is infrequent, its superset should not be generated / tested

Method:

L_k denotes the set of frequent k-itemsets: Large itemset

C_k is the superset of L_k : candidate for large itemset.

- Apriori Algorithm is a two step process is followed consisting of join & prune actions to generate L_k from L_{k-1}.
  - Join step: Apriori assumes that items within a transaction or itemset are sorted in lexicographic order.

The candidate set C_k is generated by taking the join $L_{k-1} \times L_{k-1}$, where members of L_{k-1} are joinable if their first k-2 items are in common. This ensures that no duplicates are generated.

- Prune step: To reduce the size of C_k, Apriori property is used as follows —
  - Any (k-1) itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1) subset of a candidate k-itemset is not in L_{k-1}, the candidate cannot be frequent and can be removed from C_k.
- The count of each candidate in C_k is used to determine L_k.

Algorithm Apriori-generate (L_k):

1. for each itemset l_1 in L_k
2. for each itemset l_2 in L_k
3. If k-1 elements in l_1 and l_2 are equal.
4. $c = l_1 \times l_2$
5. add c to C_{k+1}
6. for each k subset s of c.
7. If s does not belong to L_k then

8. delete c
9. break.

The Apriori Algorithm:

C_k: Candidate itemset of size k
L_k: frequent itemset of size k
L_1 = {frequent items};
1 : For ($k=1$; $L_k != \phi$; $k++$) do
2. begin
3. C_{k+1} = Apriori_generate(L_k)
4. for each transaction t in database do
5. increment the count of all candidates in C_{k+1}
6. that are contained in t
7. L_{k+1} = candidates in C_{k+1} with min_support
8. end
9. return $\cup_k L_k$;

Example;
Supmin = 2.

Tid	Items
10	A, B, E
20	B, E
30	B, C
40	A, B, D
50	A, C
60	B, C
70	A, C
80	A, B, C, E
90	A, B, C

Page:
Date: / /

Solution:

Tid	Items
20	B, E
30	B, C
40	A, B, D
50	A, C
60	B, C

L_1

C_1 — 1st scan →

Itemset	sup
{A}	6
{B}	7
{C}	6
{D}	1
{E}	3

L_1 →

Itemset	sup
{A}	6
{B}	7
{C}	6
{E}	3

$L_1 \times L_1$ →

Itemset	sup
{A}	6
{B}	7
{C}	6
{E}	3

Itemset	sup
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, E}	3
{C, E}	4

2nd scan →

Itemset	sup
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, E}	3

L_2

Itemset	sup
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, E}	3

$L_2 \times L_2$ →

$C_3 = \{\{A, B, C\}, \{A, B, E\}, \{A, C, E\}, \{B, C, E\}\}$

The 2-item subsets of {A, B, C} are {A, B}, {B, E}, {A, E} which are all in L_2.

The 2-item subsets of {B, C, E} are {B, C}, {C, E} & {B, E}, {C, E} are not in L_2

Remove {B, C, E}

Page :
Date : / /

Q3

Itemset
{A, B, C}
{A, B, E}
{A, C, E}
{B, C, E}

3rd scan →

Itemset	sup
{A, B, C}	2
{A, B, E}	2

Itemset
{A, B, C}
{A, B, E}

L3 × L3 →

$C_4 = \{\{A, B, C, E\}\}$

The 3 - item subsets of {A, B, C, E} are {A, B, C}, {B, C, E}, {A, C, E} & {A, B, E}, {B, C, E} and {A, C, E} are not in L_3

Remove {A, B, C, E}

Thus, C_4 is empty & algorithm terminates.

Q.2) Explain FP growth tree.

- 1) This algorithm is an improvement to the Apriori method.

2) A frequent pattern is generated without the need for candidate generation.

3) FP Growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree.

4) This tree structure will maintain the association between the itemsets.

5) Input:

- D, a transaction database;
- min_sup, the minimum support count threshold.

output: The complete set of frequent patterns.

Q.3) Write application of Data mining.

Applications of data mining:

List of areas where data mining is widely used –

1) Financial Data analysis.
2) Retail Industry
3) Telecommunication Industry.
4) Biological Data analysis.
5) Other scientific Applications.
6) Instrusion Detection.
7) Fraud detection.
8) Health & medicine.

seen 03/06/22