CSE 601:DATA MINING & BIOINFORMATICS

Association Analysis

Pradeep Singh Bisht	50247429	pbisht2@buffalo.edu
Ashutosh Ahmad Alexandar	50248859	alexanda@bufalo.edu
Dilip Reddy Gaddam	50248867	dilipred@buffalo.edu

Apriori Algorithm:

Apriori Algorithm is data mining algorithm which is used to generate frequent itemsets and association rules from frequent itemsets, over transactional databases.

Key Principle of Apriori Algorithm:

If an itemset is frequent then all the subsets of that itemset are frequent. so, we can use this principle to generate large itemsets from shorter frequent item sets.

Implementation of Apriori Algorithm:

We have implemented the Apriori algorithm in 2 steps:

1. Generation of frequent itemsets.

2. Rule generation from frequent itemsets.

Step-1 (Frequent itemset generation):

1.Generate frequent itemset of length 1, by traversing the whole database and count the attributes. If count divided by total number of rows is greater than minsupport the attribute is considered to be the frequent itemset of length 1.

2. Generate frequent itemsets of length 2 using frequent itemsets of length 1 by making combinations of length 1 frequent itemsets and calculating the support of each combination. From these candidate sets, we prune(delete) the ones which have support less than minsupport. We then proceed to generate candidate sets of length n+1, and pruning them, until there are no newer frequent item sets being formed.

3. While generating frequent itemset of length n, we use 2 frequent itemsets of length n-1 and combine these two itemsets if first n-2 elements of both the itemsets are same. We prune the itemsets of length n by calculating the support for itemset and comparing with minsupport.

Step 2 (Rule generation):

1.After generating the frequent itemsets we use these frequent itemsets for rule generation. Every frequent itemset is divided into two sub-sets (head, body) which are not equal to empty sets and calculate the confidence for rules. If confidence rule is greater than minconfidence then the association rule is saved.

2. If the length of frequent item is n. There are 2^n-2 possible rules which can be generated from each frequent itemset.

Calculating Confidence:

If (X,Y,Z) is a frequent itemset and rule generated is X,Y→Z.

Then confidence of rule is support(X,Y,Z) divided by support(X,Y).

If confidence of rule is greater than minconfidence then rule is stored as association rule.

Output:

Min Support = 30%

Number of length-1 frequent Itemset: 196

Number of length-2 frequent Itemset: 5340

Number of length-3 frequent Itemset: 5287

Number of length-4 frequent Itemset: 1518

Number of length-5 frequent Itemset: 438

Number of length-6 frequent Itemset: 88

Number of length-7 frequent Itemset: 11

Number of length-8 frequent Itemset: 1

Total Number of all ItemSet:12879

Min Support = 40%

Number of length-1 frequent Itemset: 167

Number of length-2 frequent Itemset: 753

Number of length-3 frequent Itemset: 149

Number of length-4 frequent Itemset: 7

Number of length-5 frequent Itemset: 1

Total Number of all ItemSet:1077

Min Support = 50%

Number of length-1 frequent Itemset: 109

Number of length-2 frequent Itemset: 63

Number of length-3 frequent Itemset: 2

Total Number of all ItemSet:174

Min Support = 60%

Number of length-1 frequent Itemset: 34

Number of length-2 frequent Itemset: 2

Total Number of all ItemSet:36

Min Support = 70%

Number of length-1 frequent Itemset: 7

Total Number of all ItemSet:7

Association rules based on Templates:

All results are based on Support: 50% and Confidence: 70%

Template:1

Query	Count
RULE Has ANY of [G59_Up]	26
RULE Has NONE of [G59_Up]	91
RULE Has 1 of [G59_Up, G10_Down]	39
HEAD Has ANY of [G59_Up]	9
HEAD Has NONE of [G59_Up]	108
HEAD Has 1 of [G59_Up, G10_Down]	17
BODY Has ANY of [G59_Up]	17
BODY Has NONE of [G59_Up]	100
BODY Has 1 of [G59_Up, G10_Down]	24

Template:2

Query	Count
Size(RULE)>= 3	9
Size(HEAD)>=2	6
Size(HEAD)>=1	117

Template:3

Query	Count
HEAD HAS ANY OF [G10_Down] OR BODY HAS 1 OF [G59_Up]	24
HEAD HAS ANY OF [G10_Down] AND BODY HAS 1 OF [G59_Up]	1
HEAD HAS ANY OF [G10_Down] OR SizeOf(BODY) >= 2	11
HEAD HAS ANY OF [G10_Down] AND SizeOf(BODY) >= 2	0
SizeOf(HEAD) >= 1 OR SizeOf(BODY) >= 2	117
sizeOf(HEAD) >= 1 AND SizeOf(BODY) >= 2	3