ASSOCICATED RULE MINING

Pallav Shukla

2154638

CSCM35

BIG DATA AND DATA MINING

Introduction

In the following report we will discuss about association rule mining algorithm. It is a process for discovering relations between item sets in large database, rules determine about why the items are interconnected and how they are connected to each other [1] . We can see relation in figure 1 [5]. We will be applying apriori algorithm through brute force approach. In apriori we find the most frequent item-sets in the complete database and then we apply joining and pruning reiteratively [3]. Item-sets occurring above the minimum threshold value are known as frequent item-sets [2]. Minimum support count is the value which is used to eliminate items from the frequent itemset [4]. In brute force approach for every value of the subset created from different combinations we create the confidence and support to choose the rule which has the maximum support number. After finding out the association we further went ahead and found out the rules with their support numbers. Then we only choose the interesting rules that is the rules with maximum support number. We also tried the Fp growth approach for the datasets.

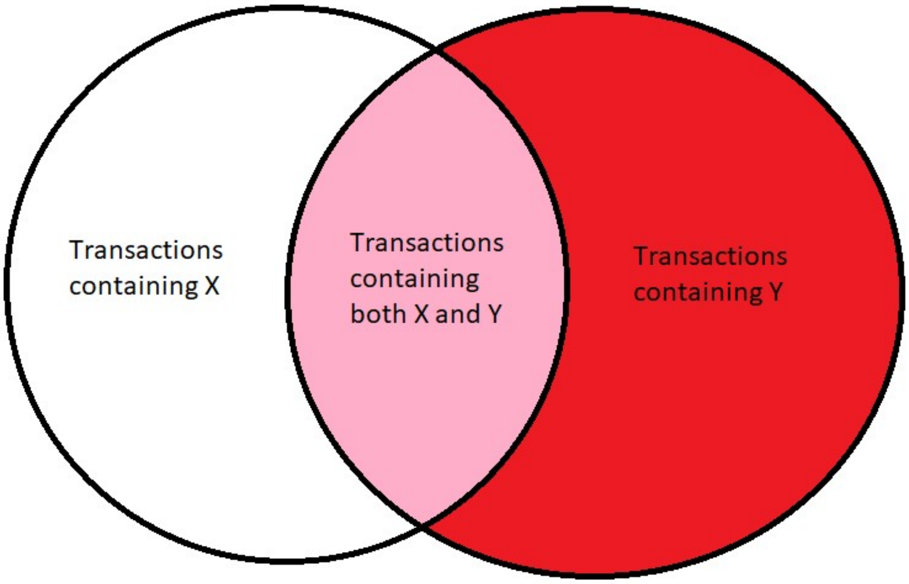
Transactions containing X
Transactions containing both X and Y
Transactions containing Y

Figure 1 : A Venn Diagram to show the associations between item-sets X and Y of a dataset.

Experiment and Discussion

Our experiment was based on association rule mining we received 2 datasets, one was from a grocery store, and one was from an online retail store. This data set which was small and has only 20 rows we proceeded through reading the comma separated dataset with a csv reader and then split the result and attached it to a list. Of course, this was done using comma as a delimiter. Once we had the data set ready in form of a list, we started applying brute force algorithm on it. For applying the same we First found out the frequency of each unique element in the whole data set We store this result in a dictionary which was then traverse one by one to fetch out the element name and then make subsequent combination of the sets. There were multiple sets created for which we calculated the support by using counter subtraction on a dictionary. Subtraction on a dictionary was possible by using the counter which subtracted the respective values of the keys from the two dictionaries. This implies that the key should be the same for the two dictionary to have some operation on them. We also used issubset to find out the particular subsets in hole of the data sets so that we can calculate the support for them. This was done for all C1, C2, C3, c4, c5 where the number denotes the maximum number of elements in a set. We then proceed to use the Fp growth algorithm on the same data set of grocery store. In this approach we use the library Fp growth , which itself is based on FP tree. FP Tree is a representation in a compressed form. The tree keeps track of relations between item sets. Here we take each item set and map it to a path in tree one by one. Concept is again that frequently occurring items shall have better chances of sharing items. This approach traverse is the tree recursively to get the frequent pattern. The frequent patterns which are generated from this approach are added for growing the tree which is also called pattern growth. There are two stages. Stage one we create the tree and stage 2 we mine the tree and get conditional FP trees.

Conclusion

In this approach we found out that a priority algorithm on its own cannot be used for a large data set because the running time of issubset is definitely not something which can be utilized on a real time data set for example when we try to apply it on online retail data set it was taking too much time to compute. And therefore, when we used the FP growth algorithm the results were much faster. We notice that while reading the data from a local repository was faster as compared to when we fetched the database from a Google drive

References

1. Upload.wikimedia.org. (2022). Retrieved 11 April 2022, from https://upload.wikimedia.org/wikipedia/commons/c/c0/Association_Rule_Mining_Venn_Diagram.png.

2. *Complete guide to Association Rules (2/2)*. Medium. (2022). Retrieved 11 April 2022, from https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84#:~:text=Frequent%20itemsets%20are%20the%20ones,above%20a%20minimum%20threshold%20%E2%80%94%20minsup.

3. https://www.softwaretestinghelp.com/apriori-algorithm/#:~:text=Apriori%20algorithm%20is%20a%20sequence,is%20assumed%20by%20the%20user.

4. *Sign in to your account*. Canvas.swansea.ac.uk. (2022). Retrieved 11 April 2022, from https://canvas.swansea.ac.uk/courses/24880/files/2762119?module_item_id=1466192.

5. Upload.wikimedia.org. (2022). Retrieved 11 April 2022, from https://upload.wikimedia.org/wikipedia/commons/c/c0/Association_Rule_Mining_Venn_Diagram.png.