International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 4, April-2014, pp: (24-28), Impact Factor: 1.147, Available online at: www.erpublications.com

An Improved Approach for Mining Frequent Patterns

Patel Harshit[1], Prof. Jayesh Chaudhary[2]
[1]Student (ME-CE), Sarvajanik College of Engineering & Technology, Gujarat, India
[2]Asst. Prof., Sarvajanik College of Engineering & Technology, Gujarat, India

Abstract: Generating association rules is a huge problem in data mining. An association rule among datasets is a sub problem of frequent itemset mining. Important is required time for generating frequent itemset. To increase the efficiency of mining frequent itemsets more and more approaches and techniques were developed. In this paper for generating frequent itemsets we developed new approach is Improved Approach Frequent (IAF). Also flow of approach, Example of approach and result of that analysis with wine dataset. Our approach is compare with SaM algorithm and time required for generating itemset is less.

Keywords: Data mining, Frequent itemset.

Introduction

Frequent patterns play important role in many data mining tasks which try to find interesting patterns from datasets, like correlations, association rules, classifiers, clusters, sequences and others which generating association rules. The original motivation of generating association rules came from required to analyze so called supermarket transaction data, to examine customer which product purchase together. Association rules describe how often items are purchased together. For example, an association rule “beer → chips (60%)” states that four out of five customers that bought beer also bought chips. These rules can be useful for making decisions of product promotions, store layout pricing, and others. Apriori algorithm is less successful for market based analysis in which transactions are large but frequent items generated is small in number [1].

Since introduction in 1993 by Argawal. The frequent itemset and association rule mining problems have received a great deal of attention. To solve mining problems efficiently more number of paper published for presenting new algorithms and improvements on existing algorithms. To exploiting customer behaviour and make correct decision leading to analyze huge amount of data [2]. For example, an association rule “beer, chips (60%)” states that four out of five customers that bought beer also bought chips. These rules can be useful for decisions making for promotions, store layout, product pricing and others [4].

In this paper, we explain the description of frequent itemset mining algorithms. Next, we describe a new approach for generating frequent itemsets using Improved Approach Frequent (IAF). Also we describe example of new approach and its result analysis. Finally, conclude and future scope of the paper.

Problem Description

In data mining, frequent itemset is acknowledged because there is more application like correlation, association rules based on frequent patterns, sequential patterns tasks. In frequent pattern itemsets finding association rules are important as other task for data mining. The major difficulty in frequent pattern mining is result of large number of patterns. As the minimum threshold becomes lower, an exponentially large number of itemsets are generated. So pruning is unimportant in mining and it becomes important topics in mining frequent patterns. Therefore, the goal is optimize the process of finding frequent patterns which is scalable, efficient and get important patterns. [3].

Methodology

In a large transactional database multiple items are there so the database surely contains various transactions which contain same set of items. Thus by taking advantage of these transactions trying to find out the frequent itemsets and prune off the candidate itemsets whose node count is lower than min support using new approach, result in efficiently execution time.

Sampling method is a popular method in computational statistics; two important terminologies related to it are population and sample. The population is defined in keeping with the objectives of the study; a sample is a subset of population. Usually, when the population is large, if the sample is scientifically chosen, it can be used to represent the population, because the sample reflects the characteristics of the population from which it is drawn.

Usually, in data mining, the population is large, so the sampling method is appropriate. As in given example, suppose that the sample S data in Table 2 is a carefully chosen sample of some population P in Table 1.

Table 1: Population Data

TID	List of item_IDs
1	i_1, i_2, i_5
2	i_2, i_4
3	i_2, i_3
4	i_1, i_2, i_4
5	i_1, i_3
6	i_2, i_3
7	i_1, i_3
8	i_1, i_2, i_3, i_5
9	i_1, i_2, i_3
10	i_1, i_2, i_5
11	i_2, i_4
12	i_2, i_3
13	i_1, i_2, i_4
14	i_1, i_3
15	i_2, i_3
16	i_1, i_3
17	i_1, i_2, i_3, i_5
18	i_1, i_2, i_3

Using sampling method can save much time, if the sample is carefully chosen, the sample can represent the population, and then the table that comes from the sample can represent that comes from the population, the 2-itemsets with high frequency in sample's table are liable to be the one with high frequency in population's table.

Table 2: Sample Data

TID	List of item_IDs
1	i_1, i_2, i_5
2	i_2, i_4
3	i_2, i_3
4	i_1, i_2, i_4
5	i_1, i_3
6	i_2, i_3
7	i_1, i_3
8	i_1, i_2, i_3, i_5
9	i_1, i_2, i_3

Procedure

1.) Carefully draw a sample S from the population P, usually by random sampling.

2.) To deal with the sample S to get a table, denoted as table HS.

3.) Rank the table HS with respect to the frequency of column content in order to make the column address with high frequency lie in the former and that with low frequency the latter, then we get a new table HSR.

4.) Based on HSR, to deal with the rest sample of the population P, i.e. P − S, when finished, get a table denoted as HP.

5.) Obtain frequent itemsets according to predetermined minimum support count.

Example

Take the data in Table 1 as an example, we will show how procedure works.

- Draw a sample S from the population P, shown in Table 2.
- To deal with the sample S to get a table, denoted as table HS, shown in Table 3.

Table 3 : Table H_S

Address	(1, 2)	(1, 5)	(2, 5)	(2, 4)	(2, 3)	(1, 4)	(1, 3)	(3, 5)
Count	4	2	2	2	4	1	4	1
Content	{i1, i2} {i1, i2} {i1, i2} {i1, i2}	{i1, i5} {i1, i5}	{i2, i5} {i2, i5}	{i2, i4} {i2, i4}	{i2, i3} {i2, i3} {i2, i3} {i2, i3}	{i1, i4}	{i1, i3} {i1, i3} {i1, i3} {i1, i3}	{i3, i5}

- Rank the table HS with respect to the frequency of column content in order to make the column address with the high frequent content lie in the former and that with low frequent content the latter, get a new table HSR, shown in Table 4.

Table 4 : Table H_{SR}

Address	(1, 2)	(2, 3)	(1, 3)	(1, 5)	(2, 5)	(2, 4)	(1, 4)	(3, 5)
Count	4	4	4	2	2	2	1	1
Content	{i1, i2} {i1, i2} {i1, i2} {i1, i2}	{i2, i3} {i2, i3} {i2, i3} {i2, i3}	{i1, i3} {i1, i3} {i1, i3} {i1, i3}	{i1, i5} {i1, i5}	{i2, i5} {i2, i5}	{i2, i4} {i2, i4}	{i1, i4}	{i3, i5}

- Based on HSR,To deal with the rest sample of the population P, i.e. P − S, when finished, get a table denoted as HP, shown in Table 5.

Table 5 : Table H_P

Address	(1, 2)	(2, 3)	(1, 3)	(1, 5)	(2, 5)	(2, 4)	(1, 4)	(3, 5)
Count	8	8	8	4	4	4	2	2

Content	{i1, i2}{i1, i2}
	{i1, i2}
	{i1, i2}
	{i1, i2}
	{i1, i2}
	{i1, i2}
	{i1, i2}
		{i2, i3}{i2, i3}
		{i2, i3}
		{i2, i3}
		{i2, i3}
		{i2, i3}
		{i2, i3}
		{i2, i3}
			{i1, i3}{i1, i3}
			{i1, i3}
			{i1, i3}
			{i1, i3}
			{i1, i3}
			{i1, i3}
			{i1, i3}
				{i1, i5}{i1, i5}
				{i1, i5}
				{i1, i5}
					{i2, i5}{i2, i5}
					{i2, i5}
					{i2, i5}
						{i2, i4}{i2, i4}
						{i2, i4}
						{i2, i4}
							{i1, i4}
							{i1, i4}
								{i3, i5}{i3, i5}

- Obtain frequent itemsets according to predetermined minimum support count. If we set support count as 6, we find that 2-itemset {i1, i2}, {i2, i3} and {i1, i3} are frequent.

Result Analysis

In our experiments we choose hepatitis dataset with different properties, to prove the efficiency of the algorithm. In the hepatitis dataset, 155 number of records and 19 number of columns. Table 6 shows the dataset from the UCI repository of machine learning databases.

Table 6: The characteristics of Dataset

Itemset	Number of Records	Number of Columns
Hepatitis.data.txt	155	19

As a result of the experimental study, revealed the performance of improved approach with the SaM algorithm. The run time is the time to mine the frequent itemsets. The experimental result of time is shown in Table 7 reveals that the improved approach outperforms the SaM approach. The experimental result is also shown in Figure 1.

Table 5.2: Execution Time for SaM and Improved Approach using Hepatitis dataset

Support (in %)	Total Execution time in second	
	SaM	Improved Approach
40	0.79	0.53
50	0.16	0.11
60	0.09	0.05

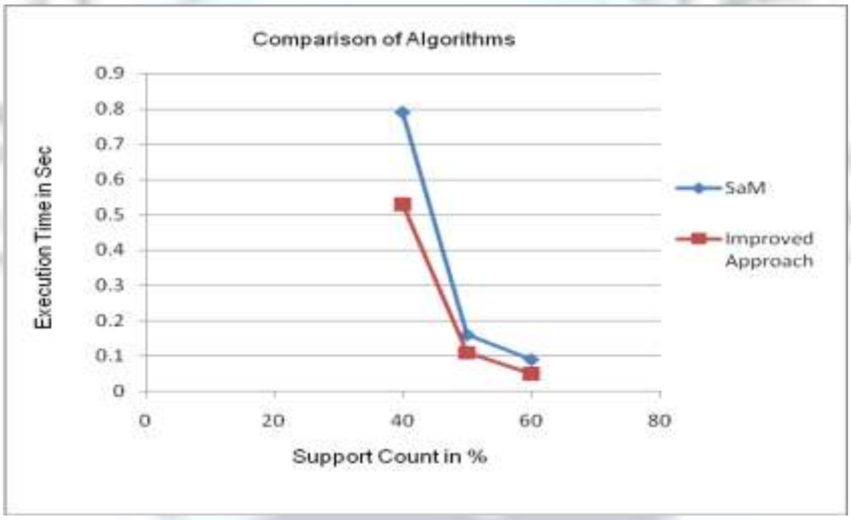

Figure 1: Total Execution Time for SaM and Improved Approach using Hepatitis dataset

As it is clear from the comparison new procedure performs well for the low support value for the Hepatitis dataset which contains 155 transactions and 19 numbers of columns. But at the higher support its performance small reduces compare to SaM algorithm. Difference between execution time of improved approach and SaM are decreases in later stages.

Conclusion

By considered the following factor for creating our improved approach, which are the time consumption, these factor is affected by the approach for finding the frequent itemsets. Work has been done to develop an improved approach which is an improvement over SaM algorithm.

For hepatitis dataset the running time consumption of our new scheme outperformed SaM. Whereas the running time of improved approach performed well over the SaM on the collected dataset at the lower support level and also running time of improved approach performed well at higher support level. Thus it saves much time and considered as an efficient method as proved from the results.
We are use some constraints by user input for reduce total execution time for mining frequent itemsets. We are developing application based on this method.

References

[1]. Agrawal.R and Srikant.R, Fast algorithms for mining association rules, In Proc. Int’l Conf. Very Large Data Bases (VLDB), 487–499 (1994).

[2]. Raorane A.A., Kulkarni R.V. and Jitkar B.D., Association Rule – Extracting Knowledge Using Market Basket Analysis, Res. J. Recent Sci., 1(2), 19-27 (2012).

[3]. Pramod S, O. P. Vyas, Survey on Frequent Item set Mining Algorithms, In Proc. International Journal of Computer Applications (0975 - 8887), 1(15), 86–91 (2010).

[4]. R. Agrawal, T. Imielienski, and A. Swami, Mining Association Rules between Sets of Items in Large Databases, Proc. Conf. on Management of Data, 207–216 (1993).

[5]. Shrivastava Neeraj and Lodhi Singh Swati, Overview of Non-redundant Association Rule Mining, Res. J. Recent Sci., 1(2), 108-112 (2012).

[6]. Raorane A.A., Kulkarni R.V. and Jitkar B.D., Association Rule – Extracting Knowledge Using Market Basket Analysis, Res. J. Recent Sci.,1(2), 19-27 (2012).

[7]. Borgelt C., Efficient Implementations of Apriori and Eclat, In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (2003).

[8]. Kumar, A.V.S., Wahidabanu, R.S.D., A Frequent Item Graph Approach for Discovering Frequent Itemsets, In Proc. Conf. Advanced Computer Theory and Engineerin, 952-956 (2008).

[9]. Borgelt C., SaM: Simple Algorithms for Frequent Item Set Mining, IFSA/EUSFLAT 2009 conference (2009).

[10]. Hai-Tao He, Hai-Yan Cao, Rui-Xia Yao, Jia-Dong Ren, Chang-Zhen Hu, Mining frequent itemsets based on projection array, Machine Learning and Cybernetics (ICMLC), 454-459 (2010).