Implementing Frequent Itemset Mining By advanced Distributed Approach Using Matrix-Based Pruning

Srinivasa Rao Divvela, V Sucharita,

Abstract: Mining of frequent itemsets is a very crucial process in the mining of associate rules. Significant challenges are being encountered in this era where big data has drawn its own circle by shaping around space and time factors. A distributed procedure for the job is what that best be defined. So we increment our previous version with an enhanced version by the implementing a distributed approach which is better that FP growth. Differentiating of the existing and proposed algorithm is done using the practical valuable data that is available.

Keywords: big data, FP growth, distributed approach, frequent itemset.

I. INTRODUCTION

Frequent itemset mining is a crucial and important part of associate rule mining. Mining of frequent itemsets is a very crucial process in the mining of associate rules. Significant challenges are being encountered in this era where big data has drawn its own circle by shaping around space and time factors. A distributed procedure for the job is what that best be defined. The use of these techniques is more essential and crucial for the development of finding of the frequent and rare items for the procedure and are very useful in the mining aspects. The incorporation of bigdata is a major break through for the procedural techniques.

II. LITERATURE SURVEY

Revised Manuscript Received on June 29, 2019.
Srinivasa Rao Divvela, Research Scholar Department of CSE KL University Lakreddy Bal Reddy College of Engineering Vaddeswaram, India
Dr V Sucharita, Professor Department of CSE Narayana Engineering College Gadur, India

Revised manuscript received on June 29, 2019.

Srinivasa Rao Divvela, Research Scholar Department of CSE KL University Lakreddy Bal Reddy College of Engineering Vaddeswaram, India
Dr V Sucharita, Professor Department of CSE Narayana Engineering College Gadur, India

We incorporate the matrix_based pruning algorithm procedure with the big data spark framework such that we can enhance the efficiency and scalability of the procedure. The main factor is the acquiring of the Boolean vector for each item in the algorithm of a given database and calculation of the 2-itemset.
Implementing Frequent Itemset Mining By advanced Distributed Approach Using Matrix-Based Pruning

matrix, by using them the reduction of amount of candidate itemset is done.

Step 1

Fig: Matrix Prunning

IV. IMPLEMENTATION

The implementation of the above algorithm for single system primarily based on spark is said in detail. In order to lessen thememory usage, hashmap is used in our program to save the boolean vector for every frequent 1 itemset and all of the common 2-itemset, in preference to saving the 2-itemset matrix m.

Step 1: achieve the boolean vectors for all common 1-itemset, then produce all the common 2-itemset based totally on these boolean vectors.

Step 2: get all common (k + 1)-itemset with the aid of the use of common ok-itemset(k ≥ 2).

With the help of these two steps we hence further improvise and improve the efficiency of the mining procedures on the framework of apache spark.

Step 2

V. RESULTS

Here are the detailed reports of the attributes of the data sets that have been used in the process of the implementation of the distributed procedure we have proposed in this paper. Here the comparison of the previous PFP and the present implemented distributed procedure with respect to their runtime along with the detailed info of the datasets and data used in given in the tabular form for better and easy understanding of the process.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Support degree</th>
<th>PFP (Spark 1.3)</th>
<th>DFIMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>T10H4D100K</td>
<td>0.01</td>
<td>64s</td>
<td>25s</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>35s</td>
<td>17s</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>37s</td>
<td>14s</td>
</tr>
<tr>
<td>T40H1D100K</td>
<td>0.01</td>
<td>321s</td>
<td>180s</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>316s</td>
<td>81s</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>250s</td>
<td>67s</td>
</tr>
</tbody>
</table>
VI. CONCLUSION

The main hindrance and lacking in the map reducing approach is the repeated scanning but after overcoming the over scan problem we have the power and space to deal with. A distributed procedure for the job is what that best be defined. So we increment our previous version with an enhanced version by the implementing a distributed approach which is better that FP growth. Differentiating of the existing and proposed algorithm is done using the practical valuable data that is available. By using the matrix_based pruning and spark we overcome the problems and we have evaluated it on a standard dataset to ensure its genuinely on a whole. We have proved that the process proposed by us shows great amount of scalability and improves efficiency of the system. When there is a high support degree our proposed procedure shows most effective results. It can be improved further by incorporation of optimization techniques.

REFERENCES

1. A distributed frequent itemset mining algorithm using Spark for Big Data analytics, Feng Zhang1, Min Liu1, Feng Gui1, Weiming Shen2, AbdallahShami, Yunlong Ma1, Springer Science+Business Media New York 2015
3. Neutrosophic_associationrule mining algorithm for big data analysisM Abdel-Basset, M Mohamed, F Smarandache-Symmetry, 2018 - mdpi.com