International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

New Distributed Approach for Frequent itemset Data mining

Nikhil Prakash Gorde and Prof. B.B.Gite

Department of Computer Engg, SAE kondhwa, Pune, Maharashtra, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

To discover the frequent itemset is very significant task in data mining. These frequent itemsets are beneficial in applications like Association rule mining as well as co-relations. To mine frequent itemsets these systems are with certain algorithms, but these are incompetent in allocating as well as balancing the load, when it comes through extreme data. Automatic parallelization is also impossible with these algorithms. To overcome these issues of present algorithms there is necessity to develop algorithm which will support the missing features, such as spontaneously parallelization, balancing as well as good distribution of data. In this paper we are with a new method to discover frequent itemsets by using MapReduce. Modified Apriori algorithm is used with HDFS environment this is called FiDoop Method. In this technique mapreduce process will work individually as well as simultaneously by using the decompose strategy. The outcome of this mapreduce method will be given to the reducers then reducers will display the outcome. In the experiment we used three diverse algorithm like basic apriori, FP Growth and our proposed modifies apriori, the system has executed in standalone machine as well as distributed environment and shown the results how proposed algorithm is better than existing algorithms.

Keywords: *Association Rules, Frequent item sets, FiDoop, MapReduce, Modified Apriori.*

Introduction

Data mining ideas and strategies can be connected in different fields similar to marketing, medicine, real estate, and customer association management, and engineering, web mining and so forth. Data mining (DM) has as its main goal; the generation of nonobvious however useful info for conclusion makers from very huge Data mining functions take in clustering, classification, prediction, and link analysis (associations). With these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. One of the most vital data mining applications is that of mining association rules to generate the knowledge which will help to the top level management or stakeholders to take an effective decision in the business organizations. This will definitely enhance the E- business in all extent. After a careful study of existing algorithms used for data mining through the research the researcher proposes efficient algorithms for mining multilevel association run the show, which searches for fascinating relationship among things in a given data set at various levels in a successful way. This will be helpful for the business specially the e-business. Various existing data mining techniques are produced displayed to infer association rule and frequently happening item sets, yet with the quick entry of time of big data customary data mining algorithm have not been able to meet huge datasets analysis requirements. There is need to improve performance and accuracy of parallel processing with minimizing execution time complexity. Also assuring the output of a computation is insensitive to changes in any one personal record. So that it will restricting privacy leaks from results. Hence, there is need to provide better frequent item set mining approach using HDFS framework with privacy Preservation techniques.

A standout amongst the most essential data mining applications is that of mining association rules to generate the knowledge which will help to the top level management or stakeholders to take an effective decision in the business organizations. This will definitely enhance the E-business in all extent.

After a careful study of existing algorithms used for data mining through the research the researcher proposes an efficient algorithm for mining multilevel association rule, which looks for intriguing relationship among items in a given data set at various levels in a compelling way. This will be helpful for the business specially the e- business.

In this modern era datasets are excessively large so only sequential algorithms are not able to compute large database and they failed to analyze data accurately and also they suffer from performance degradation. To solve this problem, a new parallel frequent item sets mining algorithm is used with Map

Reduce named as fidoop. This mechanism improves the capacity of storage and computation of problem.

Literature Survey

JW. Han, J. Pei et al [1] the frequent pattern tree stores the compressed data in a broadened prefix tree structure. The frequent patterns are put away in a compressed shape. A FP-tree based mining technique known as the FP-growth is created. The proposed algorithm helps in mining the frequent item sets without the hopeful set age. Three methods were utilized to accomplish the effectiveness of mining:

1) An extensive database is changed over into a little data structure to dodge the rehashed database checks which is said to be exorbitant.
2) It embraces a pattern frequent growth strategy to abstain from producing substantial hopeful sets which is exorbitant.
3) The mining tasks are isolated into littler task which is exceptionally helpful in lessening the hunt space. The FP-tree based mining likewise has numerous examination issues like the SQL-based FP-tree structure with high versatility, mining frequent patterns with imperatives and utilizing FP-tree structure for mining successive patterns.

As per H. Li, Y et. Al. [2] parallel FP-growth algorithm the mining task is isolated into various parcels. Every one of the segments is given to the diverse machines and each parcel is registered freely. To conquer the difficulties faced by the FP-growth algorithm like the capacity, dispersion of calculation and exceedingly costly calculation parallel FP-growth algorithm is proposed. The PFP algorithm comprises of five stages. In the initial step, the database is separated into little parts. In the second step the Mapper and the reducer are utilized to do the parallel counting. In the third step the frequent items are gathered. In the Fourth step the FP-tree is developed and the frequent item sets are mined. In the fifth step the neighborhood frequent item sets are totaled. The PFP algorithm is powerful in mining tag-tag associations and Web Page-Web Page associations which are utilized as a part of question suggestion or some other inquiry.

Removing frequent item sets from the huge database the creators showed an issue. In this paper the creators have exhibited an issue of separating the frequent items from extensive number of database. The creators discovered rules that have least value-based support besides least confidence. They have projected an calculation that deliberately assesses the thing sets for one pass. Similarly it will change between the quantity of disregards information and thing sets that are estimated. This count uses pruning framework for maintaining a strategic distance from certain thing sets. Points of interest of this calculation are that it utilizes support administration system which are not appropriate in the memory in one pass thus will move to next pass. Likewise there is no repetition [1].

This is an upgraded method to measure performance of Apriori like algorithm into MapReduce.
MapReduce is the approach which is used for parallel mining of large size data in either homogeneous or heterogeneous groups. MapReduce distributes the excessive data between map and reduce functions and it allows total utilization of resources compared to existing systems. Therefore now a days MapReduce is the popular technique for parallel mining. By taking benefit of MapReduce the authors have suggested three algorithms that are SPC, FPC, and DPC. In these algorithms they have used Apriori algorithm with MapReduce function. DPC algorithm accepts the different lengths of data dynamically, which is advantage of this algorithm. DPC shows great performance compared to other two algorithms that are SPC and FPC. Thus these three algorithms demonstrate that these calculations scale up straightly with dataset sizes.

According to Zhigang Zhang et.al. [3] The vertical format algorithm the frequent patterns are mined utilizing the algorithm Eclat. The algorithms for mining frequent patterns in flat format databases are not the same as the algorithms for mining vertical databases like the Eclat. A parallel algorithm MREclat which utilizes a map reduce system has been proposed to get the frequent item sets from the enormous datasets. Algorithm MREclat comprises of three stages. In the underlying advance, all frequent 2-item sets and their tid records are gotten from exchange database. The second step is the adjusted gathering step, where frequent 1-item sets are parceled into gatherings. The third step is the parallel mining step, where the data got in the initial step is redistributed to various computing nodes. Every hub runs an enhanced Eclat to mine frequent item sets. At last, MREclat gathers all the yield from each computing hub and arrangements the last outcome. MREclat utilizes the enhanced Eclat to process data with a similar prefix. It has been demonstrated that MREclat has high scalability and great speedup proportion.

Frequent item set mining is a critical part [12] in association rules and different other fundamental data mining applications. Be that as it may, tragically as dataset gets greater well ordered, mining algorithms neglected to deal with such unnecessary databases. The creators have proposed an adjusted parallel FP-Growth algorithm BPFP [3], an expansion of PFP algorithm [1]. FP-growth is utilized with the MapReduce worldview called as Parallel FP-growth algorithm. BPFP is accustomed to balancing the load in PFP, which upgrades parallelization and naturally this component improves execution. BPFP gives more noteworthy execution by utilizing PFPs grouping system. BPFP parallelizes the huge load with well-balanced algorithm [3].

FIUT is another strategy for mining frequent item sets. It is extremely productive strategy for FIM (frequent item set mining) named as FIUT (Frequent Item set Ultra metric Trees) [4]. It encloses two primary stages of scans of database. In the first stage it calculates the support count for all item sets in a large database. In the second stage it relates prune method & give merely frequent item sets. In the intervening time frequent one item sets are premeditated, phase two will assemble small ultra metric trees. These results will be displayed in small ultra metric trees. Benefit of FIUT is that it expels K-FIU tree speedily. FIUT has four fundamental points of interest. First, it decreases I/O overhead by examining the databases just twice. Second, decreases the searching space. Third, FIUT gives frequent item sets as yield for every expansive number of processing. So user can get just frequent item sets by using this new strategy for FIUT as each leaf provides frequent item sets for each datum trade inside the cluster [4].

It [4] uses an extended Map-Reduce Framework. A number of sub files are obtained by splitting the mass data file. The bitmap computation is performed on each sub file to acquire the frequent patterns. The frequent pattern of the general mass data file is acquired by incorporating the results of all sub files. A statistical analysis method is used to prune the insignificant patterns when processing each sub file. It has been demonstrated that the strategy is scalable and effective in mining frequent patterns in big data.

Xinhao Zhou and Yong feng Huang have proposed An Improved Parallel Association Rules Algorithm Based on Map Reduce Framework for Big Data in [5]. The proposed algorithm is contrasted and the existing conventional Apriori algorithm. The time many-sided quality of both the algorithms has been used to think about the execution of the algorithms. It has been demonstrated that the proposed algorithm is more productive contrasted with the conventional algorithm.

According to Jinggui Liao et. Al. [6] is a parallel algorithm which is executed using the Hadoop stage. The MRPrePost is an enhanced Pre-Post algorithm which uses the map reduce structure. The MRPrePost algorithm is used to discover the association rules by mining the substantial datasets. The MRPrePost algorithm has three steps. In the first step the database is isolated into the data blocks called the shards which are distributed to every specialist hub. In the second step the FP-tree is constructed. In the last step the FP-tree is mined to acquire the frequent item sets. Test results have demonstrated that the MRPrePost algorithm is the fastest.

In [7] Large datasets are mined using the Map reduce system in the proposed algorithm. Big FIM algorithm is altered to get the ClustBig FIM algorithm. ClustBig FIM algorithm provides scalability and speed which are used to get useful data from substantial datasets. The useful data can be used to settle on better decisions in the business action. The proposed ClustBig FIM algorithm has four fundamental steps. In the first step the proposed algorithm uses K-means algorithm to produce the clusters. In the second step the frequent item sets are mined from the clusters. By constructing the prefix tree the worldwide TID list are acquired. The sub trees of the prefix tree are mined to get the frequent item sets. The proposed ClustBig FIM algorithm is ended up being more productive contrasted with the Big FIM algorithm.

In [8] this paper tackles the issues of finding the unprecedented and weighted thing sets. The occasional thing set mining issue is finding thing sets whose repeat of the information is not exactly or equal to most outrageous edge. This paper reviews different system for mining occasional thing set. Finally, relative strategy for each procedure is introduced. Information Mining is described as Extraction fascinating examples or gaining from tremendous measure of information". Information digging is the framework for finding information from different perspectives and outlining into helpful information. Finding of common examples concealed in a database has a key impact in a few information mining errand. There are two anticipate sorts of models in information mining.

3. Existing System

In this section we will understand the essential concepts theory concerning Association Rule Mining (ARM) and Map Reduce technique. A thorough survey of the literatures relevant to FPM in Big Data is presented.

Association Rule Mining

Within Association Rule Mining, the frequent rules are found that term relations sandwiched between unconnected frequent items in databases, and it has two (2) main measurements: confidence values and sup-port [4]. Association rules have two fragments: antecedent (if) as well as a consequential (then). A forerunner is an item discovered in the information. A subsequent is found in mix with the precursor. These rules are made by examining information for frequent if/at that point examples and utilizing the criteria support and confidence to distinguish the most essential connections. Sign of what number of frequent items shows up in the database is called as Support. Number of times if/at that point articulations have been observed to be genuine is called as Confidence. The item set that have support esteem more noteworthy than or equivalent to a base limit support esteem, and frequent rules as the rules that have confidence esteem additional prominent than otherwise equivalent to least edge confidence esteem are called as frequent item sets. The limit esteems are

generally thought to be accessible for mining frequent item sets. ARM is tied in with discovering all rules whose support and confidence surpass the edge, least support and least confidence esteems. Association Rule Mining comprises of two primary advances: initial step is to discover all item sets with adequate supports and the second step is to create association rules by consolidating these frequent or extensive item sets. In customary framework the edge esteems are thought to be known. It is dreadfully tricky to set the threshold significance without any prior knowledge and to obtain the required results. Setting the threshold significance very high may produce very small no. of rules or else if the threshold significance is rest dreadfully low then it may produce big number of rules and it will take long time for computing the result. The outcome of these rules can be acquired in key-value pair and this output is then mapped. To map, Map and Reduce technique is used.

Parallel Frequent Pattern Mining Algorithms

The primary weakness of tree-based algorithms is that they use memory excessively. Even for relatively smaller-sized datasets, when the minimum threshold is set to a small value, the main tree can grow to billions of nodes, leading to substantial memory consumption. Hence, significant research efforts have been devoted to developing parallel implementations of frequent pattern mining algorithms of precise data, which make it feasible to construct and mine smaller-sized trees on multiple machines in parallel. Recall that each conditional tree does not have any computational dependencies on other conditional trees. This property was exploited by a number of parallel implementations of FP-growth.

4. Proposed System Overview

Problem Description

Mining has turned into a blasting subject of research in Computer Science. This fast growing phenomenal is driven by numerous reasons. First, information mining and information warehousing are to a great degree fertile with inquire about issues but then present an outrageous helpful instrument to oversee huge measure of information. Besides, information develops at an exponential rate and Internet technology has made it simple to suspect that information from everywhere throughout the world so companies and associations these days find themselves immersed with information, and anxious to extricate useful information from them to benefit their business.
Therefore, from computer science point of view, the two dominant problems in e-commerce becomes **(1) How to extract information efficiently – this is the topic of data mining algorithm;**
(2) How to be serviced and benefited by results of data mining – this is the topic of ecommerce. **This study is aimed to present a software design and implementation to solve these problems for enhancing e-commerce businesses.** The proposed work first investigate data mining approaches like Apriori, FP Tree, FIUT and find the issues of existing system, System also focus on database security like SQL injection with parallel data mining top k retrieval approach on synthetic and real time dataset in HDFS framework Implement Fidoop on our internal Hadoop cluster with multi data node.

Proposed System Architecture

The initial step of Frequent item set generation is to generate frequent 1-item sets for the given database. For this support count algorithm has been proposed which is explained in detail here, it has been shown how MapReduce is used to find frequent 1-item sets as well as to generate frequent item sets using constraints. To increase the efficiency of map reduce task a cache has been included in the map phase to maintain support count tree for calculating the frequent-1 item set of each Mapper which is shown in Figure 1. As the data in cache can be quickly fetched it reduces the total time of calculating Frequent-1 item sets, since it bypasses the shuffle, sort and the combine task of each Mapper in the original MapReduce tasks. To further increase the efficiency of generating FIM, cache is introduced so that the support count can be calculated in the cache itself. For this a Modified Map Reduce algorithm has been proposed.

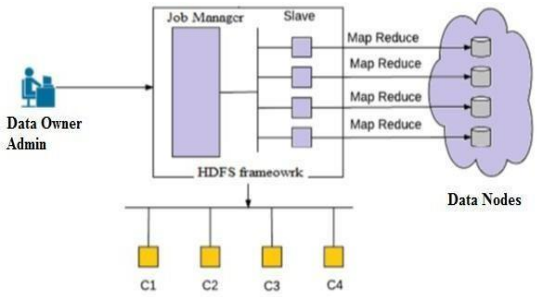

Figure 4.2: Proposed System Architecture

1. System Authentication

This is the starting page of the software in which user has to sign up to use the system. This accepts login id and password for the registered user. If a user fails to provide correct id and password then system will not allow it to be used further. This is done for security purpose.

2. File Upload with Hadoop Process By clicking "File Upload" option it allows to use the dataset for the system. Also it starts the MapReduce functionality by clicking on the "Hadoop Process" button. MapReduce will start working with this step.

3. Inventory data

By clicking on the show -data option it shows the Item code which is given to the items purchased by the user.

4. Hash Base Apriori
"Process" option allows selecting the algorithm to calculate frequent item sets. It shows two options with proposed algorithm name. By clicking on the option user can see the frequent item sets with reduced time.

5. Frequent-1-item set
This is the first MapReduce stage. By giving the minimum support value and dates user can get the values. By using Modified Apriori algorithm user will get the frequent-1 item sets.

6. Frequent-k-item set
This is the final frequent item sets in the form of frequent-k item sets. K value is given by the user.

7. Largest item set
By clicking on this option the user will get the largest items purchased by the user in a particular date.

8. Comparison graph with two methodologies This is the graph which shows the timing difference between Fidoop and base Apriori as well as FP Tree algorithm.

Hybrid Algorithm for finding frequent itemset (FP tree+ Apriori)
Hash Base FIM Algorithm (AM_ebiz)

Input: Dataset DB, Support generation denominator De, min_req_itemsmk;
Output: generate T item set
Step 1: for all (T in DBi) do
Step 2: items [] split (T)
Step 3: Create support
T = (T.count/100)*De

Step 4: Create hash table from HT= {Ti+1…….Ti+n}

Step 5: add each item occurrences with respective Ti

Suppor t D	Ti+1	Ti+ 2	Ti+ 3	--	--	--	Ti+1 0
Count	{I1}, { In}	{I1} , {In}	{I1} , {In}	--	--	--	{I1}, {In}

Step 6: Create two pair group

Support D	Ti+1	Ti+2	--	Ti+10
Count	{I1, I2}, { I1, I2}	{I1, I2}, { I1, I2}	--	{I1, I2}, { I1, I2}

Step 7: Create three pair group

Support D	Ti+1	Ti+2	--	Ti+10
Count	{I1, I2, I3}, {I1, I2, I3}	{I1, I2, I3}, {I1, I2, I3}	--	{I1, I2, I3}, {{I1, I2, I3}

Step 8: Create n pair group

Support D	Ti+1	Ti+2	--	Ti+10
Count	{I1, I2, I3, In}, {I1, I2, I3, In}	{I1, I2, I3, In}, {I1, I2, I3, In}	--	{I1, I2, I3, In} {I1, I2, I3, In}

Step 9: Apply pruning on HT till when get top k items.

Step 10: Return top-k Items from HT

Performance of Fidoop in terms of Time required in seconds with different support denominator with different dataset

Following table indicates the time taken in seconds by Proposed Algorithm Fidoop for the three different datasets of Grocery, Electronic & Sports for different support counts.

Results And Discussions

We used synthetic data look like market basket data through short frequent patterns. The further two datasets are real data, which are condensed in long frequent patterns. These data sets were frequently used in the preceding study of association rules mining. The experimental outcomes of this framework are set to the minimum provision threshold (or, proportionately, bigger data sizes) than having even yet been deliberated. These upgrade same at no implementation cost, as demonstrate by the way that our execution reaches the performance associate to other methods less time. Consequently we are taking Fidoop algorithm; it can be best algorithm to give the precise outcomes as match to present systems. Proposed system algorithm demonstrates the faster execution even for large database. We could create our own vast dataset against which to likewise run tests; yet the cost for doing as such is trifling. The information in the web documents set comes from a real domain besides thus is significant. We have used five sets of data in our experiments. Three of these sets are synthetic data T10I4D100K, T25I10D10K, and T40I10D100k. The other two datasets are real data (Groceries) which are compressed in long frequent

patterns. These data sets were live data sets for the study of association rules mining in addition to were downloaded from http://www.jbtraders.in/ .

For the proposed execution we have used three different dataset, the grocery dataset has taken from www.jbtraders.in and some electronic item base synthetic dataset has given from internet. The third dataset has taken from sport www.sports365.in. Below we have done multiple experiments which are represented in graphs.

Table 5.1: Performance of Time required in seconds with different support denominator with different dataset

Support Count	Time in Seconds	Time in Seconds	Time in Seconds
Support 5	31	29	30
Support 8	26	24	25
Support 10	22	21	20
support 15	15	16	14

Figure 5. 2 shows the result presented in Table 1 for three different item set of Grocery, Electronic & sports for various support values of 5,8,10 and 15% respectively. It shows the time required to extract the frequent item set in seconds with different dataset. Utilization is very good in applying the algorithm Fidoop for frequent item set mining with retail item set for various support values.

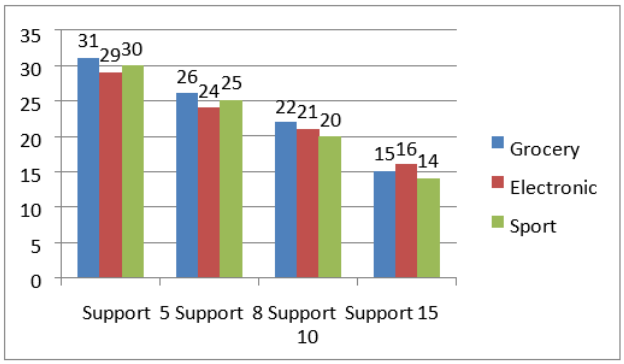

Figure 5.2: Time required in seconds with different support denominator with different dataset.

The experiment results showed using graphs exhibits that the time used for extraction of frequent item set using Fidoop decreases as the support increases

Conclusion

In this research work, the key attention is on the problem of Mining Frequent Item sets from large Data efficiently and effectively in Hadoop environment. Due to the fact that a huge number of patterns or rules are frequently found in Frequent Item set Mining, Constrained Frequent Item set Fidoop algorithm has been proposed and it is implemented in Big Data using MapReduce task in Hadoop. Most of the FPM algorithms spend half the time in generating frequent 1-item sets. A simple and easy to implement support count using proposed hash base algorithm has been proposed which has reduced the time of generating frequent 1-item sets. This algorithm can be effortlessly implanted into any of the current algorithms intended at association rule mining in order to excerpt frequent 1-item sets as well as their consistent counts. To reduce the execution time of extracting Frequent Item sets from Big Data using MapReduce, a modified Map Reduce Fidoop has been proposed. In this a cache has been included in the map phase to maintain support count tree for calculating the frequent-1 item set of each Mapper. This reduces the total time of calculating Frequent-1 item sets since it bypasses the shuffle, sort and the combine task of each Mapper in the original MapReduce tasks. Another feature of finding cumulative frequent item sets from multiple files is also added.

Future Scope

For future enhancement we can focus on parallel mining with distributed hadoop environment using slot configuration. The runtime slot allocation and ordering can maximized the resource utilization and improve the system accuracy.

References

[1]. JW.Han, J.Pei and YW.Yin, —Mining Frequent Patterns without Candidate Generation‖, International Conference on Management of Data, vol. 29(2), 2000, pp. 1-12.
[2]. H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang, —PFP: Parallel FP-growth for query recommendation‖, Proceedings of the 2008 ACM Conference on Recommender Systems, 2008, pp. 107- 114.
[3]. Zhigang Zhang, GenlinJi, Mengmeng Tang, MREclat: an Algorithm for Parallel Mining Frequent Itemsets‖, 2013 International Conference on Advanced Cloud and Big Data.
[4]. Hui Chen, Tsau Young Lin, Zhibing Zhang and JieZhong, "Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce‖, 2013 IEEE International Conference on Granular Computing.
[5]. Xinhao Zhou, Yongfeng Huang, "An Improved Parallel Association Rules Algorithm Based on MapReduce Framework for Big Data‖, 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery.
[6]. Jinggui Liao, Yuelong Zhao, Saiqin Long, MRPrePost-A parallel algorithm adapted for mining big data‖, 2014 IEEE Workshop on Electronics, Computer and Applications.
[7]. SheelaGole, Bharat Tidke, — Frequent Item set Mining for Big Data in social media using ClustBigFIM algorithm‖, International Conference on Pervasive Computing.
[8]. Siddique Ibrahim S P, Priyanka R, —A Survey on Infrequent Weighted Item set Mining Approaches‖ , 2015, IJARCET, Vol.4, pp. 199-203.
[9]. SurendarNatarajan, SountharrajanSehar, —Distributed FP-ARMH Algorithm in Hadoop Map Reduce Framework‖, 2013 IEEE.
[10]. Xiaoting Wei, Yunlong Ma , Feng Zhang, Min Liu, WeimingShen, Incremental FP-Growth Mining Strategy for Dynamic Threshold Value and Database Based on Map Reduce‖, Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.
[11]. M.-Y. Lin, P.-Y.Lee, and S.-C. Hsueh, "Apriori- based frequent item-set mining algorithms on MapReduce," in Proc. 6th Int. Conf. Ubiquit. Inf. Manage. Commun. (ICUIMC), Danang, Vietnam, 2012.
[12]. L. Liu, E. Li, Y. Zhang, and Z. Tang, "Optimization of frequent Item-set mining on multiple-core processor," in Proc. 33rd Int. Conf. Very Large Data Bases, Vienna, Austria, 2007
[13]. Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao proposed "Mining Frequent Patterns without Candidate Generation: A Frequent- Pattern Tree Approach".
[14]. Assaf Schuster, Ran Wolff, define a approach "Communication Efficient Distributed Mining of Association Rules "