Data Analysis of E-Commerce Websites

Prity A. Choudhary[1] Prof. M. S. Chaudhari[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]PBCOE Nagpur India

Abstract— In this paper, we presented algorithm for mining top-k ranked association rules where k is the number of association rules user wants to mine. We have taken dataset of Ecommerce website for analysis of products, which are used in association rule mining, then made evaluation for association topkrules algorithm. Also studied TopK Rules optimized algorithm and compare with our proposed algorithm. We got better performance of proposed algorithm than of TopKRule algorithm.

Key words: Association Rule Mining, Top-K Rules, Support, Confidence, Hadoop, Big Data, Frequent Closed Itemsets, Mapreduce

I. INTRODUCTION

Products of Ecommerce website are arranged in categories with large number of datasets. Data mined and storage format are irrespective. Applicability is the most important rather than storage. Association rules mines to identify patterns that are frequently purchased together.

Drawbacks of Association rules are, as it produces number of support and confidence. To overcome this, Top-k association rules are derived by user, Fibonacci Heap sort algorithm is used.

MapReduce Algorithm always uses distributed data and work parallel in distributed system. Algorithm has interfaces and it is independent of the backend technology. Hadoop Distributed file system is a Java-based file system that provides scalable and reliable data storage, and designed to span large clusters of commodity servers.

In this paper, we address the problem of association rule mining by proposing an algorithm named TopKRules unified. This latter utilizes a new approach for generating association rules named "rule expansions" and several optimizations. An evaluation of the algorithm with datasets commonly used in the literature shows that TopKRules has excellent performance and scalability.

Moreover, results show that TopKRules is an advantageous alternative to classical association rule mining algorithms for users who want to control number of association rules generated. The rest of the paper is organized as follows. Section 2 presents the related work and issues. Section 3 represents proposed system, the association rules, its advantage & disadvantages; describes TopkRules. Section 4 presents Result analysis and evaluation. Finally, Section 5 presents the conclusion.

II. RELATED WORK

Some papers describe the various issues and related work. Philippe Fournier-Viger, Cheng-Wei Wu and Vincent S.[1], studied the Apriori and Association rules deeply and they found Algorithms generates no. Of sequential rules need long execution time and huge memory consumption. To address this issue, they proposed TopkRules, an algorithm to discover the top-k rules having the highest support, where k is set by the user.

To generate rules, TopKRules relies on a novel approach called rule expansions and also includes several optimizations that improve its performance. They concluded results show that TopKRules has excellent performance and scalability, and that it is an advantageous alternative to classical association rule mining algorithms when the user wants to control the number of association rules generated**.**

Drashti B Patel & Reema Patel[2] studied that users need to select the parameter values of minimum confidence and minimum support in classical algorithm. So for these extremely large numbers of rules are generated and it suffers from the long execution time and more memory required. They proposed the algorithm which generate rules exactly how much user wants. This proposed algorithm gives better performance and scalability than others and it is advantageous to classical algorithms when the user wants to control the number of association rules generated.

Luo Fang et.al[3], proposed a Frequent item generation is a key approach in association rule mining. The Data mining is the process of generating frequent item sets that satisfy minimum support. Efficient algorithms to mine frequent patterns are crucial in data mining. Since the Apriori algorithm was proposed to generate the frequent item sets, there have been several methods proposed to improve its performance. But they do not satisfy the time constraint

JongWook Woo [4] proposes Apriori Map/Reduce Algorithm and illustrates its time complexity and shows that the algorithm gains much higher performance than the sequential algorithm as the map and reduce nodes get added. The item sets produced by the algorithm can be adopted to compute and produce Association Rule for market analysis.

They develop the code following the algorithm on Hadoop frame, which practically proves that the proposed algorithm works. Besides, the algorithm should be extended to produce association rule.

III. PROPOSED SYSTEM

The main procedure of Top K association is to find interesting associations between items in a transaction database.

A transaction database is a set of transactions, where each transaction is a set of items (an itemset).

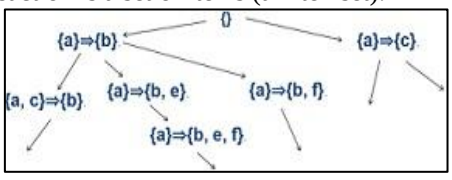

Fig. 1: Left-Right Expansion

In fig 1, Top-k find larger rules by recursively scanning the database for adding a single item at a time to the left or right part of each rule (these processes are called left and right expansions)

An evaluation of the algorithm shows that it gives better scalability and good performance. Algorithm takes transactional database as an input and also minimum confidence as an input. Then it will generate k number of association rules as in output. The algorithm will work as follows. Algorithm takes three parameters as inputs that are minimum confidence value, k value and transactional database.

Two more variables are N & E.

N is used to store the top k sequential rules. E is used to store the rules to be expanded right or left.Rules are expanded on the basis of condition. Initially minimum support is set to 1, N and E will be NULL. Then dataset will be scanned one time to store sequential ids of each items in Tids(i).Then algorithm will generate frequent item and its rule which satisfy minimum confidence and minimum support value. Condition will be checked that if no. of rules generated till now is greater or equal to value of k or not.

If N is greater than k then a rule having lowest minimum support will be removed from our rule list.

If N is less than k then first check rule should not be exist already. If yes already exist then rule will not added to N, if no then the rule will added in both N and E. After this the algorithm will select a rule with highest min_sup from the E and perform left and right expansion.

Algorithm is based on the concept of pattern growth also takes less space and time. In this method rules of size 1 are generated like a=>b. Now, the method is based on expansion technique the rules are right expanded a=>b,c & a=>b,d and checked for the minimum support. Then rules are left expanded like a,b=>c & a,d=>c and checked for minimum support.

In this way 1*1 rules are expanded to generate larger rules and then those larger rules are expanded to generate much larger rules.

IV. Result Analysis

We run algorithm by varying value of k. Value of k is taken different as 5, 7, and 10 respectively. And also we have taken minconf value as 0.1 to 0.6. As value of k increases, the execution time and memory required both increases linearly.

Datasets Average	No. of transaction	No. of distinct items	Transaction size
Chess	3196	75	37
Connect	67557	129	43
Mushroom	8416	128	23
Pumsb	49046	7116	74

Table 1: Dataset Characteristics

This evaluation has been shown below in the table and its respective chart. First as shown below we have results of dataset amazon dataset with different value of k. This table 2 and figure 1 shows memory requirement is linearly increasing with value of k.

Memory(mb)			
Min Conf	Dataset=200 0 k=5	Dataset=500 0 k=7	Dataset=10000 k=10
0.1	04.66	05.95	16.32
0.2	04.17	13.82	09.71
0.3	18.64	06.13	10.72
0.4	18.60	17.59	19.11
0.5	18.81	19.18	19.10
0.6	17.75	17.45	18.57

Table 2: Memory Consumption

Execution Time			
Min Conf	Dataset=200 0 k=5	Dataset=500 0 k=7	Dataset=10000 k=10
0.1	93	78	187
0.2	46	93	125
0.3	951	63	124
0.4	858	921	1997
0.5	890	952	1950
0.6	951	905	1967

Table 3: Execution Time

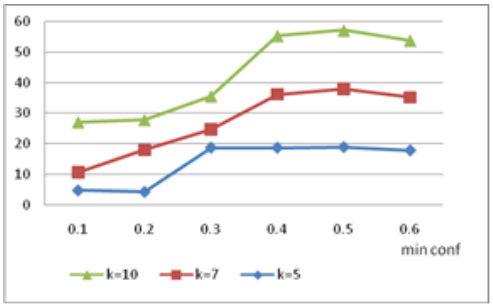

Fig. 2: Memory for Topk rules

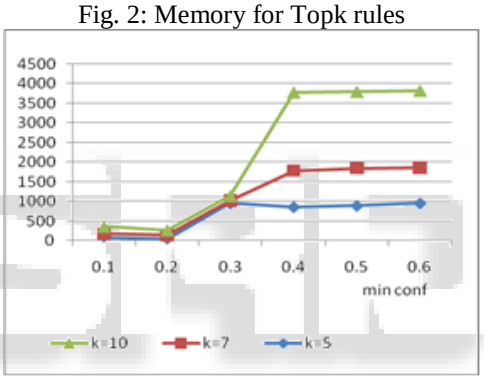

Fig. 3: Execution time

V. Conclusion

The two different algorithms after combining improves the efficiency of system in terms of time and iteration. The Apriori and Association algorithm applied on large uncertain database to obtain frequent sequential patterns. Since database is large uncertain, the large number of patterns are generated which is difficult to study hence there is need of top k rule mining algorithm for obtaining precise results. TopKRules algorithm generate number of rules and is very efficient and not time consuming.

References

[1] Philippe Fournier-Viger, Cheng-Wei Wu 87and Vincent S., "Mining Top-K Association Rules", Canadian Conference on Artificial Intelligence, 2012.

[2] Drashti B Patel,Reema Patel, "Technique for mining top k-association rules", IJIRT-2015, Volume 1 ,Issue 12.

[3] Luo Fang, Qiu Qizhi ,"The Study on the Application of Data Mining Based on Association Rules", International Conference on Communication Systems and Network Technologies 2012 pp.477-480

[4] Jongwook Woo, "Apriori-Map/Reduce Algorithm", Computer Information Systems Department California State University Los Angeles, CA

[5] Serenko, James Hayes, "Investigating the functionality and performance of online shopping bots for E-commerce", Int. J. Electronic Business, Vol. 8, No. 1, 2010
[6] Othman Yahya, Osman Hegazy, Ehab Ezat, "An Efficient Implementation of Apriori Algorithm Based On Hadoop-MapReduce Model", International Journal of. Reviews in Computing, Vol 12, 31st December 2012.
[7] Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters.",Proc. of the 6th Symposium on Operation Systems Designing and Implementation, 2004.
[8] Xin Yue ,Yang Zhen Liu ,Yan Fu, "MapReduce as a Programming Model for Association Rules Algorithm on Hadoop" ,Information Sciences and Interaction Sciences (ICIS), 3rd International Conference on July 2010
[9] Ahmad Tasnim & Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications", International Journal of Computer Applications (0975 – 8887) Volume 69– No.8, May 2013
[10] Michael R. Baye, John Morgan, Patrick Scholten , "The Value of Information in an Online Consumer Electronics Market", Journal of Public Policy and Marketing, 2003.
[11] Manisha Girotra, Kanika Nagpal, "Comparative Survey on Association Rule Mining Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 84 – No 10, December 2013
[12] Jalpa Mehta, Jayesh Patil, Rutesh Patil, Mansi Somani, Sheel Varma, "Sentiment Analysis on Product Reviews using Hadoop "International Journal of Computer Applications (0975 – 8887) Volume 142 – No.11, May 2016
[13] Jiawei Han and Micheline Kamber. Data Mining, Concepts and Techniques. Morgan Kaufmann, 2001

