Volume 6, Issue 1, January 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

www.ijarcsse.com

Research Paper

Available online at: www.ijarcsse.com

Review on Mining of Constraints Based Interesting Patterns from Uncertain Data Using MapReduce Technique

Darpan Kumari, Leena H. Patil, U. K Thakur
Computer Science Engineering, RTM Nagpur University,
Maharashtra, India

Abstract- Mostly Data mining algorithms finds the interesting patterns of data from transactional database where the data is precise.But in order to get the required frequent pattern of users from data which is not precise rather it is uncertain then mining of data is not smooth.This problem has increased if we go searching of data from huge amount of data like finding users interesting patterns from big data where data are in thousands of terabytes. In this paper,we propose an algorithm that (i) allows users to express their interest in terms of constraints and (ii) uses the MapReduce model to mine uncertain Big data for frequent patterns that satisfy the user-specified constraints.

Keywords- MapReduceModel; programming skill for Big Data mining; Big Data analysis; Searching and mining Big Data; Frequent Pattern; Constraints; Uncertain data.

I. INTRODUCTION

As we know very well that Data Mining is a domain which helps in extracting useful information from data.Some data Mining techniques like clustering,Classification in real life applications such as grouping similar people based on their interest or grouping similar people based on their constraints properties.Association rule and frequent pattern technique of data minig helps super market owner to know the purchase behavior of customer's pattern for any item. As the technology advances ,the Big Data information explosion is mainly due to the vast amounts of data generated by social media platform,data input from omni-channels,various mobile devices,user generated data,multi-media data and so on. This leads us into the new era of *Bigdata*[10],which refer to interesting high-velocity, highvalue,and/or high-variety data with volumes beyond the ability of commonly-used software to capture, manage, and process within a tolerable elapsed time. Hence, new forms of processing data are needed to enable enhanced decision making, insight, process optimization, data mining and knowledge discovery.This results into big data analytics[9],[11] to mine and analyze Big data allows us to continuously or iteratively explore, investigate, and understand past business performance so as to gain new insight and drive science or business planning.

A. Uncertain Dataset

In uncertain data each transaction contains items and their existential probabilities[3],[4],[5] .The uncertainty of such expected item can be expressed in terms of existential probability [3],[4]. In uncertain data, each item in a transaction is associated with probability that indicate the possibility that the item exists in the transaction. Normally we These types of set of data items is called uncertain dataset. Figure.1 shows an example of transactional dataset of Precise and uncertain data.

II. EXISTING TECHNIQUES AND ALGORITHM

A. FP Growth

It involves two main jobs:-(a)Constructing the FP-tree which capturtes the transactional values from precise data of databases.(b)Then growing frequent patterns will be eliminated.Then these frequent patterns will be sorted in adecreasing order.

B. Apriori Algorithm

In this algorithm a database consists of several transactions which contains items.Any pattern can be said as frequent if its support i.e the frequency or the number of occurrences is equal or greater than the predefinedsupport threshold also called as minsup.Actually it first produces the set of candidate patterns of size 1.Afterwards it counts the support for each candidate pattern in 1st set containing all frequent pattern of size1.From first set another set of different candidate patterns of sizer2 is generated.

III. LITERATURE REVIEW

Author in this paper [2] focuses all the existing data mining algorithms for searching information interesting patterns from transactional databases of precise data.But when the data are uncertain items in each transaction of these probabilistic databases of uncertain data are usually associated with existential probabilities, which shows the possibility of items to be present in the transaction.

In comparision with precise data, the search space mining from uncertain data is much larger due to the presence of the existential probabilities. This problem is worsened in Big Data era. Furthermore, in some real-life applications, users can take interest in little portion of this large search space for Big data mining. So by providing opportunities for users to express their interest for interesting patterns mining.But mostly existing data mining algorithms returns number of patterns out of which only some are interesting. In this paper,we propose an algorithm that (i) allows users to express their interest in terms of constraints and (ii) uses the MapReduce model to mine uncertain Big Data for frequent patterns that satisfy the user-specified constraints. By exploiting constraints properties, our algorithm minimizes the search space for Big data mining of uncertain data, and returns required requested patterns for Big data analytics.

In order to use the Map reduce model several algorithms have been used to mine information from a large space.An important Big data mining and analytics task is Frequent pattern mining which mine the frequently occurring items with consideration of parallel and distributed computing [8] on large clusters or grids of nodes (i.e.,commodity machines), which consist of a master node and multiple worker nodes. As implied by its name, MapReduce involves two key functions: “map” and “reduce”.One of the probem to uncover hidden knowledge from Big Data is concept drift where statistical properties of the attributes and their target classes shift over time, making the trained model less accurate.for example as people’s preferences for products change over time We have examined three considerable single tree that is VFDT,ADWIN and ioVFDT.iOVFDT has good performance for both synthetic and real-world concept-drift data streams.

Author in this paper [1] focus on its introduction, mining of frequent patterns has been the subject of study. Generally, they focus on improving algorithmic efficiency for finding frequent patterns. Most of these studies find patterns from traditional transaction databases, in which the items of each transaction is definitely known and precise. However, there are many real-life situations in which we are uncertain about the content of transactions. To manage these situations, we propose a tree-based mining algorithm to efficiently find frequent patterns from uncertain data, where each item in the transactions is associated with an existential probability. Experimental results show the efficiency of our algorithm over its non-tree-based counterpart.

Author in this paper[6] proposed to extract useful data in real-time, the information technology (IT) world is coping with big data problems. In this paper, we present implementation details and performance results of ReCEPtor,our system for “online” Association Rule Mining (ARM) over big and fast data streams. Specifically, we added Apriori and two different FP-Growth algorithms inside EsperComplexEvent Processing (CEP) engine and compared their performances using Last FM social music site data. Our most important findings shows that online ARM can generate (1)more unique rules, (2) with higher throughput, and (3) (lower latency) than offline rule mining. In addition, we have found many interesting and realistic musical preference rules such as “George Harrison□Beatles”. We demonstrate a sustained rate of ~15K rows/sec per core. We hope that our findings can help on the design and implementation of other fast data analytics systems in the future.

Author in this paper [5] focuses on frequent pattern mining is one of the most important research topics for many real life applications in the area of data mining. Frequent item set comes from association rule mining that uses to find association rules of items in large transactional database. Many existing algorithm to mine for the frequent itemset from static transaction database that is definite known and precise. In this paper we studied on uncertain data and fuzzy association rule mining approach. The main contribution of this paper is first we presented are view on existing method for finding frequent patterns form uncertain data. Second a new approach is proposed for finding frequent patterns from uncertain data, and third, the experiments are carried out to evaluate the performance of proposed approach for uncertain data. In this approach we have used fuzzy concept and originate the frequent patterns. The experimental results from the survey demonstrate that this proposed approach for valuable frequent pattern is good for uncertain data mining.

First We take uncertain data , and then preprocesses it by using some data mining technique. Then users are asked to feed some input values for getting frequent pattern from the dataset. Reverse apriori algorithm will be used for frequent pattern.

A. Algorithm

U- Reverse Apriori

Apriori Algorithm has few drawbacks for association rule mining such like; the iterations involved to reduce the minimum support until it finds the required number of rules with the given minimum confidence. The traditional approach can be improved by overriding some trade-off phases and discarding the unwanted objects and fields from the association analysis. Much attention and analysis are required for apriori for inefficiencies in various types of applications.

The proposed methodology uses reverse Apriori algorithm where we backtrack a database in order to find maximum number of frequent patterns and step by step will derive the corresponding association rules. In contrary to Apriori, this approach begins with the highest frequency of occurrences of gathered attributes of database transaction. These collective attributes are compared against the minimum support for the associated rule and is selected for next step.

B. Finding Frequent Itemsets Using Reverse- Apriori Algorithm

This approach is bottom-up which works entirely opposite to apriori algorithm. In this approach, first find out the pattern by making all possible pairs of itemset and reject the items which does not satisfy the user defines minimum threshold called minimum support minsupp. and evaluate a maximum possible limit of number of items in the dataset thereby generating a huge bulk of frequent item sets satisfying a user specified minimum support.It will gradually

decrease the simultaneously frequent item set till it gets a set of possible frequent item sets. Let DS= (a, b, c, d) are the set of items where a,b,c,d belongs to the transaction T. the pairs are said to be conjunctive if (a,b) E is user defined support. A pattern P is said to be frequent if minSupp (P) is greater than or equal to a minimum support threshold, denoted as minsupp.On the contrary, disjunctive patterns are those which contains all the different and irrelevant pairs of sets and therefore should be rejected as they are outliers Disjunctive if (a,b)! E user defined support.For example in generalized terms, let's consider a transaction based on a super market which contains a huge set of items and their occurrence frequency. It has been zeroed in user defined support on milk-made items. Considering a transaction that has all the possibilities of items being paired, Now this transaction consists of all the items ranging from–T = {bread, onion, banana, butter, toothpaste, cheese, egg, pasteurized milk, peas, wafers,
biscuits ....}

Now user defined support is to bakery products then it's not an intelligent step to take sample combination of all the item sets one by one and then generate candidate-1 item sets and so on. Thus what can be done here is that it will just take only those items which seem to lie in this category of user defined support and that is bakery made products.Thus the conjunctive pattern will contain only those products which fall into this specified range. And the rest of the items are considered as disjunctive patterns since they do not fall under the category of selection and therefore needs to be discarded.

Conjunctive sets= {bread,biscuit,pastries…}Disjunctive sets = {bread, egg, toothpaste, wafers…}.The reverse Apriori is then applied which works faster than the existing Apriori algorithm.

IV. FIGURE

TID	ITEMS
T1	A,B,C
T2	B,D
T3	A,B,D

(a)

TID	I1	I2
T1	76%	85%
T2	67%	24%

(b)

Figure 1:-(a) Precise and (b) Uncertain Dataset

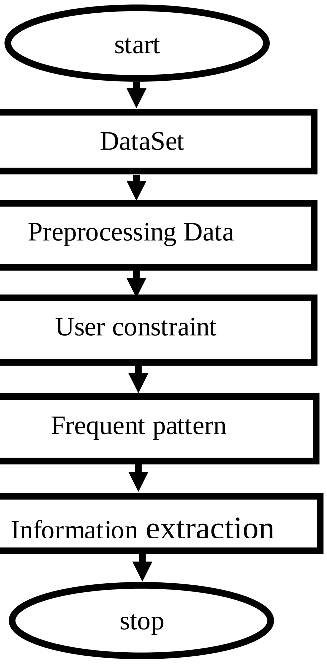

Figure2:- Dataflow Diagram of Proposed Research Methodology

V. CONCLUSION

Big Data mining is the application of data mining techniques to discover users pattern from huge amount of data, in order to understand and better serve the needs of Web-based applications.Big data mining process can be divided into two main parts.Also discovering frequent pattern minning with association rule requires a very efficient algorithms and some of algorithms have been proposed upto now,but those algorithms seemed to inefficient reases computation cost.Thus Apriori and reverse apriori suggested for frequent itemset mining. In reverse apriori approach, first find out the pattern by making all possible pairs of itemset and reject the items which does not satisfy the user defines minimum threshold called minimum support minsupp and evaluate a maximum possible limit of number of items in the dataset thereby generating a huge bulk of frequent item sets satisfying a user specified minimum support. It will gradually decrease the simultaneously frequent item set till it gets a set of possible frequent item sets.The main advantage of reverse apriori is it can reduce the time for execution of apriori algorithm. it is viable to collect a heavy collection of frequent itemset and then reducing it to give lesser scans.

ACKNOWLEDEMENT

I thank (Leung et al.,2014,Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data) for assistance with (MapReduceTechnique,Dataset,User,UncertainData,Constraint,BigDataProcessing,Frequent Pattern).I also thank L.H Patil and U.K Thakur for their comments that greatly improved the manuscript.

REFERENCES

[1] Carson Kai-Sang Leung, Christopher L. Carmichael "Efficient Mining of Frequent Patterns from Uncertain Data" Seventh IEEE International Conference on Data Mining – Workshops DOI 10.1109/ICDMW.2007 IEEE

[2] Carson Kai-Sang Leung,Richard Kyle MacKinnon,Fan Jiang "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data" 2014 IEEE International Congress on Big Data 978-1-4799-5057-7/14

[3] C.K.-S. Leung & F. Jiang, "Frequent pattern mining from time-fading streams of Uncertain data," in DaWaK 2011 (LNCS 6862), pp. 252–264.

[4] C.K.-S. Leung & S.K. Tanbeer, "PUF-tree: A compact tree structure for frequent pattern Mining of uncertain data," in PAKDD 2013 (LNCS7818), pp. 13–25.

[5] D.S. Rajput, R.S. Thakur, G.S. Thakur "Fuzzy Association Rule Mining based Frequent PatternExtraction from Uncertain Data"978-1-4673-4805-8/12 2012 IEEE

[6] E. ¨O lmezo˘gullari& I. Ari, "Online association rule mining over fast data," in IEEE Big Data Congress 2013, pp. 110–117

[7] H. Yang & S. Fong, "Countering the concept-drift problem in big datausingiOVFDT," in IEEE Big Data congress 13, pp. 126-132.

[8] M.J. Zaki, "Parallel and distributed association mining: a survey," IEEE Concurrency, 7(4):14–25, Oct.–Dec. 1999.322.

[9] P. Agarwal, G. Shroff, & P. Malhotra, "Approximate incremental bigdataharmonization," in IEEE Big Data Congress 2013, pp. 118–125.

[10] S. Madden, "From databases to big data," IEEE Internet Computing, 16(3): 4–6, May–June 2012.

[11] Yang & S. Fong, "Countering the concept-drift problem in big data using iOVFDT," in IEEE Big Data Congress 2013, pp. 126–132.