ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

Analytics of Pollution Smart City Data using New Pattern Mining Algorithm

[1] Monika Saxena, [2] Dr. C.K. Jha,

Abstract - **Pattern mining algorithms is used to mine the useful data from the massive amount of IOT data. Mostly used data mining algorithms are classification, clustering, association rule and regression in which the classification and regression comes under supervised learning and other two in unsupervised learning. The objective is to review different techniques applied for mining the pattern by using classification, clustering and other algorithm. By applying parallel data mining algorithm in map reduce frame work. We have also implemented the algorithm for mining the frequent pattern from the datasets. Apriori and FP Growth algorithm have been implemented practically on the market basket dataset. As per the result we have concluded that Apriori is better than FP Growth in terms of time stamp. And FP Growth is better than Apriori in terms of large datasets. This paper represents the problems occur in this type of methods with little bit solution of them by new modified algorithm.**

Keywords — IoT Streaming, Map Reduce, parallel data mining.

I. INTRODUCTION

In the era of information technology, everything we are using in the everyday life is represented in form of information. Transportation, parking, traffic, pollution are some examples of hundreds of infrastructure systems with which we act every day. By using information technologies combined with communication, it becomes very easy to representall details even the tiniest parts of these fields in forms of data. Furthermore, the Internet of things (IoT) plays a very important role in connecting physical objects with electronics, software, and sensors. Based on that, smart cites have been modeled and implemented in thousands place over all the world; In these cities, all smart systems in different fields like transportation network, pollution,traffic,airlines, etc. are showed in form of numbers and strings of characters. This data is collectedto form a big data.

This data can be accessed from everywhere via internet to know some information or take a decision.Also, transport energy, health care and waste management are a good example for systems which were improved intelligently, smartly, and automatically [1][2]. The live Data is very important to know facts like knowing the status of things to take immediate decisions in various aspects of our daily life. Furthermore, it is very essential to store the historical data of these things for several past years. Thereby, new rules can be predicted; unknown behaviors can be deduced thanks to the aid of analytical algorithms. Having all these sources of big data such as emails, social sites, videos, images, blogs, sensor data which are produced daily from the infrastructural systems of various aspects of life such as transportations network, pollution, IoT, Traffic systems either for buses network, underground metrosnetwork, trains network, these massive amounts of data need very huge space of storage and very special parallel computing systems. The most important questionsarise at this point are that, how these immense amounts of big data can be stored and be processed? And how to fetch the meaningful data from millions of millions of records of data?In order to answer this question, it should be known that, not all pieces of the big data are important; a lot of them are redundant information.

Consequently, filtering the data is the primary clue for solving this problem. By distinguishing the unique information and filtering the meaningful data, it could be save storage and processing time. On the other hand, by determining the frequent patterns of data, it could helpgreatly to predict the associate rule sets that can be taken as a guide in deducing the behavior of systems in advance based on the historical data. This approach is called data and frequent pattern mining.Distributed pattern mining is one of the solution method to improvethe performance of processing the Big Data. Furthermore, it saves exabytes of storage space alongside withsaving the processing time. Not only that but also, it widely opens the door for mining thousands of rule sets that are used in predicting facts and reveal the mysterious behavior of the unknown systems[3]. Data mining processis deployed by running some parallel programming tools like SAMOA (Scalable Advanced Massive Online Analysis) orMapReduce [4]. In the next section, the frequent patterns and data mining process and its evolutions is demonstrated including the good features and the weak points to be used as a guide in approaching a new technique or method of distributed pattern mining

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

that keep on the advantages and avoid the drawbacks of the conventional techniques.

Internet of things is a machine to machine connectivity. Technologies can continuously integrate classical networks with network instrument and devices. IoT brings great challenge in order to maintain and analyze the data for future use.[3].

We have implemented parallel data mining algorithm in order to improve the speed, accuracy and quality of the algorithm which is going to be applied for data mining from ware houses.

PARALLEL AND DISTRIBUTED DATA MINNING

Sujni Paul [4] has describes the method of parallelism. Problems like memory and CPU speed limitations faced by the single processors. So to solve these problems we have parallelism algorithms. There are two approaches:
Task parallel algorithm: the work of this algorithm is to assign the portion of search space to the single processor. It is divided into two groups .first group is divide and conquer that separate the search space and allocation of each portion to the specific processor. And second is Task queue that dynamically assign the smaller portion of search space to the processor when it become available.
Data parallel algorithm: Distribute the available data to the processors that are free for allocation. Data parallel algorithms have two ways. Record based partition that assigns non overlapping sets of record to each processor. And assigning sets of attributes to each processor is attribute based portioning

MAP REDUCES MODEL:

Map reduce is the framework in which we can write applications to process large amount of data parallelly. It consists of two functions that are Map and reduce. The mapper processes the data and creates several small chunks of data. The Reducer's job is to process the data that comes from the mapper.

3. PROBLEM FORMULATION & PROPOSED SYSTEM

Distributed frequent patterns and Data mining
Data volume that is represented in size of data, Data Variety which is expressed in terms of type of data and its homogeneity, velocity that is represented in the form of data, and the value of Data which is measured in how much the meaningful of data, are the great challenges in performing any type of analysis in real time. On top of these, is the veracity of the data. The problem of frequent pattern mining is to determine all relationships among the items of data in a dataset. For example, assuming there is adatabase DB with transactions Tid1 ...TidN, calculating all patterns P that are present in at least a fraction r of the transactions when"r" refers to the minimum support which is the minimum number of repetitions of a pattern of itemset. The parameter r can be expressed either as an absolute number, or as a percentage equivalent to fraction of the total number of transactions in the database. Each transaction Tidk can be considered as a binary vector, or as a set of discrete values representing the identifiers of the binary attributes that are instantiated to the value of 1. Consider the sets of items are R and Z. The rule $R \Rightarrow Z$ is considered as a rule at minimum support r and minimum confidence c, when the following two conditions hold true:
I. The set $R \cup Z$ is a frequent pattern.
II. The ratio of the support of $R \cup Z$ to that of R is at least c.

The minimum confidence c is always a fraction less than 1 because the support of the set $R \cup Z$ is always less than that of R [5].

The main idea behind deciding the frequent patters, consists of three main sub tasks which are:scanning the database for all distinct items value in the database to build the list of candidates, making all possible itemset patterns of these candidates to build the complete set of all possible itemset patterns, and testing the frequent of each pattern throughout the whole database transactions. There are many algorithms for calculating the frequent pattern. For comparing the performance parameters of these algorithms, there are two main parameters which are the number of iterations or looping paths through the database transactions and the number of candidate sets. the less are the number of looping paths and the candidate set. The more better is the performance.These algorithms will be surveyed later in the literature review section in order to show the proses and cones in order to keep the good features and avoid the drawbacks rather than adding an improvement in the proposed technique.

The most two common algorithms for determining the frequent patterns are the Apriori algorithm which was proposed by Agrawal and Srikantin 1994 to fix the problem of mining frequent itemset[6]. And, The FP-growth method which was developed by Han et al., [7]

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

which utilizes the FP-tree data structure to store the frequency information of the transaction database. Without candidate generation like in the Apriori algorithm, FPgrowth applies a frequent divide-and-conquer method and the database projection approach to determineefficiently the frequent itemsets.

The most of frequent pattern set mining algorithms are eitherApriori-like algorithm or FP-growth-like algorithm. All of them have been used in the field of analysis of patterns. Researchersfound that these algorithms which are like the Apriori algorithm,have some drawbacks such as repeated scans of the whole transactions of the database, and candidate key generation, which further requires candidate tests. Therefore, in case of there is too large datasets or complex, the time and complexity are increased proportionally. While The FP-growth like algorithms appliesthe 'Divide and Conquer' strategy and does not require candidate key generation tests. Furthermore, it does needs only two paths of scan of the database transactions. therefore, it can be concluded that the FP-growth algorithm has a faster performance. However, the FP-Growth algorithm requires a much bigger space of storage to store the information of the nodes of the tree structure and the liking pointers and indexes. Consequently, there is a room for enhancing either the performance of determining the frequent pattern or minimizing the space of storage that is required for the mining process.

In this research we are proposing a new distributed pattern mining algorithm to overcome the performance or the space of storage related problems in parameters like minimizing the number of candidates set andminimizing the number of paths of scanning the database transactions which in turn reducing thecommunication overheads,throughput, memory usage & computational costs of I/O of Big Data Mining. The new method will be categorized as Apriori-like method. However, it is completely difference in searching techniques and forming the candidates and itemset lists. The proposed method can complete the process efficiently in just one path of scan of the database transactions with a significant reduction in the number of frequent sets and in turn reducing the number of tests for the frequent of the pattern set. Moreover, the only one path of scan, applies the binary search technique which has a complexity of N LOG N while the traditional Aproiri uses many paths of scan applying the linear search algorithm with complexity of N^2 . Furthermore, it will save the required space for a tree structure and its nodes and linking pointers which is the drawback of the FP-Growth method.

4. SIMULATION AND RESULTS

Proposed solution is simulated with new modified algorithm, Our test deployment of now able to support

1. Input/output Load
2. Traffic Control

Datasetname: pollution
dataset transactions =126,373 TIDs
no of cols = 8
Field name = Ozone
The action to be taken for preparation: field+300

Field name =Particulatematter
The action to be taken for preparation:field+600

Field name = Carbon monoxide
The action to be taken for preparation : field+900

Field name =sulfure_dioxide
The action to be taken for preparation : field+1200

Field name =nitrogen_dioxide
The action to be taken for preparation : field+1500

Field name =longitude
The action to be taken for preparation: field * 10^8 - 10*10^8

Field name =latitude
The action to be taken for preparation : field * 10^8-56 * 10^8

Field name =Timestamp
The action to be taken for preparation : convert the field to number

Table 4.1 sample of the prepared data of pollution dataset

355	682	934	1237	1543	18448532	17871060	8.01E+08
355	677	936	1234	1548	18448532	17871060	8.01E+08
351	682	935	1230	1547	18448532	17871060	8.01E+08

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

348	686	934	1230	1552	18448532	17871060	8.01E+08
347	684	932	1228	1550	18448532	17871060	8.01E+08
350	682	937	1228	1545	18448532	17871060	8.01E+08
354	682	942	1227	1546	18448532	17871060	8.01E+08
352	677	942	1228	1549	18448532	17871060	8.01E+08
355	672	941	1224	1551	18448532	17871060	8.01E+08
350	671	944	1229	1546	18448532	17871060	8.01E+08
350	672	944	1226	1550	18448532	17871060	8.01E+08
346	672	949	1224	1554	18448532	17871060	8.01E+08
351	667	944	1219	1556	18448532	17871060	8.01E+08
355	665	939	1228	1554	18448532	17871060	8.01E+08
360	664	942	1230	1556	18448532	17871060	8.01E+08
356	659	947	1228	1551	18448532	17871060	8.01E+08
353	659	945	1228	1548	18448532	17871060	8.01E+08
351	658	950	1232	1549	18448532	17871060	8.01E+08
346	661	955	1231	1549	18448532	17871060	8.01E+08
348	666	951	1232	1551	18448532	17871060	8.01E+08

The data now is ready to be processed by the new algorithm of frequent pattern mining and also to be processed by any conventional frequent pattern methods like FP-Growth or Apriori. In the next chapter there are four projects as a large examples of large scale datasets for pollution, traffic, Iot, and parking to execute the test cases and showing the results of the new algorithm and comparing the results to the other conventional methods to evaluate the performance parameters of the new technique. In these projects, the process of data preparation will be discussed in details showing the encoding and decoding processes.

SIMULATION & RESULTS:

Dataset Name Pollution data set
Transaction Count 121844 TIDs

Time of processing (ms)

Support(%)	Apriori	fpgrowth	New algorithm	FP itemsets count	Minimum support in transaction
5	859	782	1063	15	6092
1	4688	1312	5822	214	1218
0.5	6422	1093	5585	519	610
0.25	17965	1078	5948	1432	305
0.1	17816	1172	6091	3418	122
0.05	28701	1187	6982	8182	61

Evaluation method

In this research we are proposing a new distributed pattern mining algorithm to overcome the performance or the space of storage related problems in parameters like minimizing the number of candidates set andminimizing the number of paths of scanning the database transactions which in turn reducing thecommunication overheads,throughput, memory usage & computational costs of I/O of Big Data Mining. The new method will be categorized as Apriori-like method. However, it is completely difference in searching techniques and forming the candidates and itemset lists. The proposed method can complete the process efficiently in just one path of scan of the database transactions with a significant reduction in the number of frequent sets and in turn reducing the number of tests for the frequent of the pattern set. Moreover, the only one path of scan, applies the binary search technique which has a complexity of N LOG N while the traditional Aproiri uses many paths of scan applying the linear search algorithm with complexity of N^2 . Furthermore, it will save the required space for a tree structure and its nodes and linking pointers which is the drawback of the FP-Growth method.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

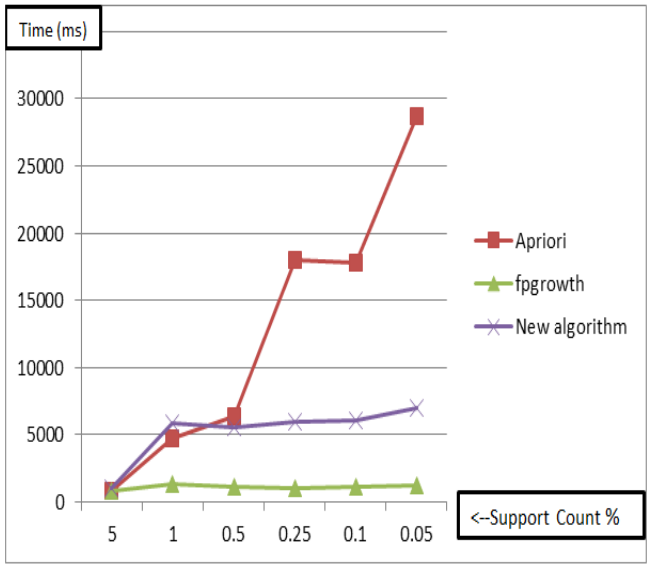

Figure 1 Time (ms) with support count after using algorithms.

6. CONCLUSION AND FUTURE WORK

In this paper we have given the review of many algorithms, methods of classification and clustering algorithm of data mining. Some applications are applicable on map reduce also. We concluded that Parallel classification algorithms are more effective on map reduce as compared to the parallel clustering algorithms. We have discussed about the advantages and disadvantages of each algorithms as shown on the table format. The new algorithm is better from apriori classical algorithm in terms of its time complexity and better than FP Growth algorithm in terms of space complexity.

REFERENCES

1. hen, Feng, et al. "Data mining for the internet of things: literature review and challenges." International Journal of Distributed Sensor Networks (2015).

2. Bin, Shen, Liu Yuan, and Wang Xiaoyi. "Research on data mining models for the internet of things." Image Analysis and Signal Processing (IASP), 2010 International Conference on. IEEE, 2010.

3. Shweta Bhatia Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 1) November 2015, pp.82-85

4. Paul, Sujni. "Parallel and distributed data mining." New Fundamental Technologies in Data Mining. InTech, 2011.

5. Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics 4 (1985): 580-585

6. Anyanwu, Matthew N., and Sajjan G. Shiva. "Comparative analysis of serial decision tree classification algorithms." International Journal of Computer Science and Security 3.3 (2009): 230-240.

7. Liu, Mingyang, Ming Qu, and Bin Zhao. "Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm." Mobile Networks and Applications 22.3 (2017): 418-426.

8. Srivastava, Anurag, et al. "Parallel formulations of decision-tree classification algorithms." High Performance Data Mining. Springer US, 1999. 237-261.

9. Wu, Gongqing, et al. "MReC4. 5: C4. 5 ensemble classification with MapReduce." ChinaGrid Annual Conference, 2009. ChinaGrid'09. Fourth. IEEE, 2009.

10. He, Qing, et al. "Parallel implementation of classification algorithms based on Map Reduce." Rough Set and Knowledge Technology (2010): 655-662.

11. Verma, Manish, et al. "A comparative study of various clustering algorithms in data mining." International Journal of Engineering Research and Applications (IJERA) 2.3 (2012): 1379-1384.

12. Lokeswari, Y. V., and Shomona Gracia Jacob. "A Comparative study on Parallel Data Mining Algorithms using Hadoop Map Reduce: A Survey." Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016.

13. Joshi, Aastha, and Rajneet Kaur. "A review: Comparative study of various clustering techniques in data mining." International Journal of Advanced Research in Computer Science and Software Engineering 3.3 (2013).

14. Chakraborty, Sanjay, and Naresh Kumar Nagwani. "Analysis and study of Incremental DBSCAN cluster algorithm." arXiv preprint arXiv: 1406.4754 (2014).

15. Kobren, Ari, et al. "An Online Hierarchical Algorithm for Extreme Clustering." arXiv preprint arXiv:1704.01858 (2017).