International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842
Issue 05, Volume 4 (May 2017) SPECIAL ISSUE www.irjcs.com

Big Data Mining For Uncertain Data

J Brundha Elci[1] , Madhavi J R[2] , Abinaya S[3] , Manjesh M[4] , Kushal G S[5]
[1] Assistant professor, [2345]UG Students
Department of CSE,
Vemana Institute of Technology, Bangalore-560034, India

Abstract ***—Data Mining aims to search for implicit, previously known and potentially useful information from data. Big Data Mining is the capability of extracting useful information from large datasets or stream of data. The existing system attempts to search the pattern of interest from probabilistic database. However, the output sometimes includes the uncertain data from existential probabilities. In many real-life applications, users may look for a tiny portion of this large search space for Big Data Mining. The proposed system reduces the search space to a greater extent as it concentrates more on the constraints by using the MapReduce model. The users are given complete freedom to express their interests by specifying their own constraints. Besides classification and clustering, anomaly detection, frequent pattern mining and association rule mining are included as the latter two analyze valuable data and helps the producer by finding the interesting or popular patterns that reveal customer purchase behavior. The algorithm proposed here greatly reduces the search space for Big Data mining of uncertain data, returning only those patterns that are interesting to the users for Big Data analytics.***

Keywords*—Big data models, Big data analytics, Frequent Patterns, Constraints, Uncertain Data*

I. INTRODUCTION

Data mining mainly deals with extracting data from the data warehouse. The data warehouse contains very large amount of data. This large section will have basically 2 sets of data. They are the interested data and uninterested data. The uninterested data must not be displayed on the screen. The interested data is expected to be on the screen. There are different tools available to reduce the space to mine the data. This tool also increases the speed of the system. In other words, the whole of the result will be displayed in less amount of time and this will take less time and will be more efficient. As the day passes the more of the data gets collected. The "uncertain" data which is asked for must be searched in already existing data plus the newly added data. The uncertain data is the data given by the user. Since the system will have the least or no guess about the users next move, the data given will be uncertain to the system. In the traditional method whole of the data warehouse is searched for the data and hence the time required was too much. The data mining was not much in scene in the beginning of the computers. As the computers have evolved, the storage area is increased. Before the whole of the database itself dint contain so much of data as is in the present case's one database.The search space is drastically increased and is the need to increase the advancement in the search tools. Big Data Mining, in brief, is the intersection of big data and data analytics.

The Big Data Mining is nothing butthe capability of extracting useful information from large datasets or stream of data. If the data mining is done on very large set of data then it can be termed as "Big Data Mining".The Big Data Mining is used in almost each and every field today.Without the use of data mining it will be very difficult to extract the required data. There are numerous search engines that are existing today. The entire search engine uses one or the other algorithm to extract the interested data.The field can be medical, banking, business, games, science and engineering and many more. This paper is organized as follows. The next section gives background and related work. In Section IV, we propose our algorithm for mining constrained frequent patterns from uncertain data using MapReduce. Conclusions are presented in Section VII.

II. LITERATURE SURVEY

Database management tools are defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. [1]Most of the presented approaches in data mining are unable to handle the large amount of data in a proper way. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools(for example data mining and statistical analysis).

The study on complexity theory of big data will aid the understanding of essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of computing models and algorithms on big data. The challenges of big data analytics are classified into four main categories.They are data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security. Data reduction, Data selection, Feature selection is an essential task especially when dealing with large datasets.A standard process is to transform the data that are semistructured and unstructured into structured and then apply data mining algorithms to extract knowledge. It is difficult to search the user interesting pattern among thousands of terabyte and hence allows users to express their interest in terms of constraints and uses the MapReduce model to mine uncertain Big data for frequent patterns that satisfy the user-specified constraints.[2]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 lead into the new era of big data. In uncertain data each transaction contains items and their existential probabilities. Existing techniques are fp growth and apriori algorithm. As implied by its name, MapReduce involves two key functions: “map” and “reduce”. One of the problems to uncover hidden knowledge from Big Data is concept where statistical properties of the attributes and their target classes shift over time resulting in less accuracy.

III. PROBLEM DEFINITION

The data mining deals with extracting the data from large set of data.But sometimes, due to inefficiency of algorithm, the system will fail to consider all the data sets present.This causes the missing of required information which is the loss to the user.This may be the result of lack of coordination among the data sets. It sometimes becomes difficult to give importance to the user’s constraint as the required data must be searched from very large amount of data.This requested data is most probably uncertain to the system. Another challenge to expose the hidden information from the data sets which makes results less useful.

IV. SYSTEM DESIGN

The fig 4.1explains the system architecture of the proposed model. The architecture has many modules. The movement of the data from one module to other module is shown in the diagram below. The foremost end is “User Accessibility” and the other end is “Data base or Data Storage”.The “knowledge base” is connected to Pattern Evaluation and Data Mining Engine. Every time the data moves from one module to another its form changes. Various processes like “Data Selection, Integration” etc are involved in the system

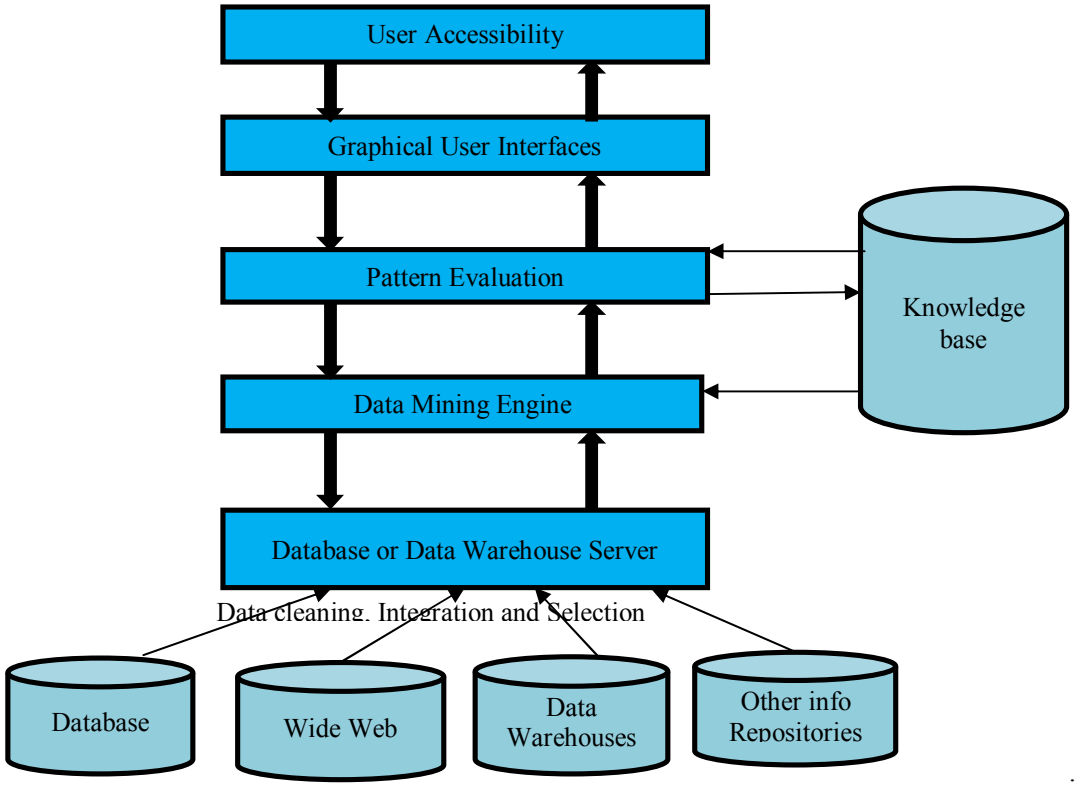

Figure 4.1: The abstract view of the system architecture

The Fig 4.2 shows the details of the system architecture. The data is mainly searched from the data sources. This data can be called as raw data. It is later subjected to the algorithm and the result is produced.It is diagrammatically explained below.

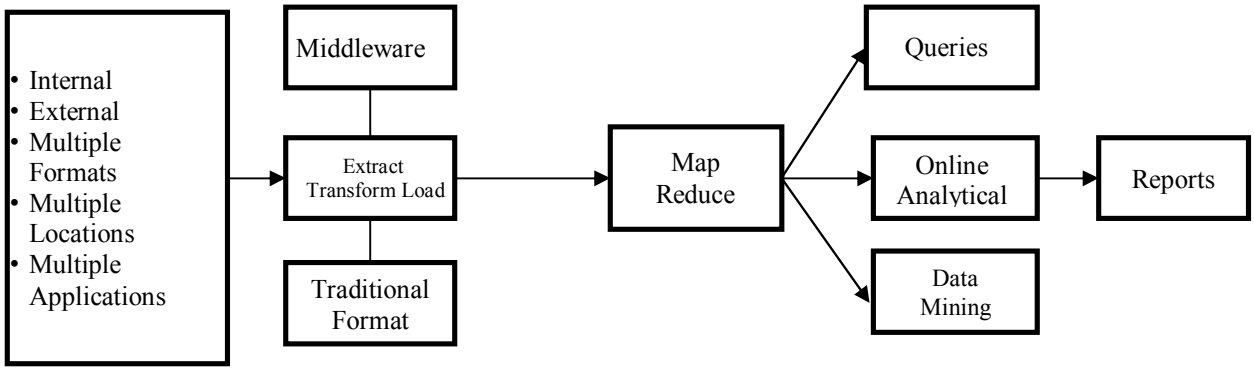

Figure 4.2: Detailed view of the system architecture

The Fig 4.3 places the use case diagram. There are mainly three actors present. They are Data Manager, Data Sources, and End User. The functionalities presented are Account Management, Data Management, Big Data Transformation, MapReduce Model and Report. The Manager manages account and data.The end user manages Data Management and Report.

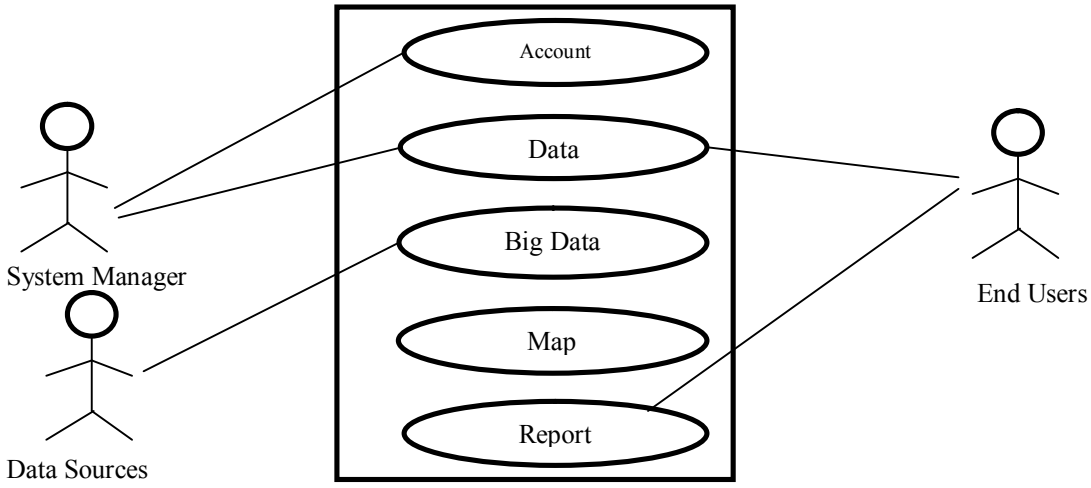

Figure 4.3: The interaction between the System and the actors.

The Fig 4.4 explains the data flow of the system. The main two persons are system manager and normal user. The system manager manages the data and account where as the normal user manages the data.

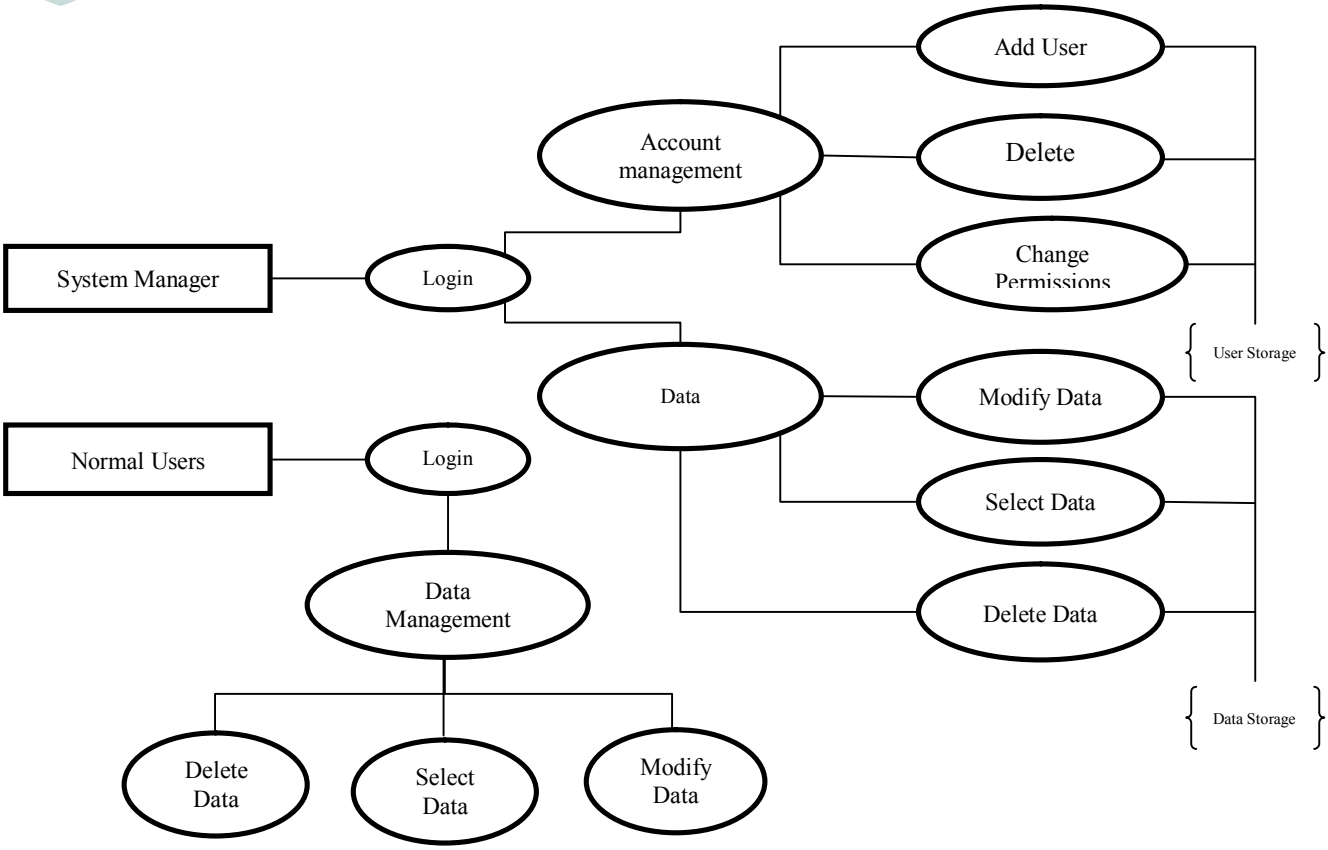

Figure 4.4: The flow of data in the system.

V. EXISTING SYSTEM

The existing system uses the tree based algorithm. The tree based algorithm is widely used. But this algorithm mines without focusing on the user's constraints. If the output produced is not matching the user constraint then the effort put to produce that line is complete waste. It consumes more time to produce both interested and uninterested data.The tree based algorithm is very traditional method. The next upcoming method is MapReduce algorithm. In MapReduce algorithm also the least required output were included in the output frame sometimes.This made the user search more and more for the accurate data. This increased the time consumption as the time was used to produced the uninterested data and also to search the required data from the whole set of output. In both the cases, the user's constraints were not given importance and this made them less efficient.

VI PROPOSED SYSTEM

In the proposed system,the search engine usesMapReduce algorithm. The MapReduce algorithm already exists but it lacks some specific features like filtering of data. Hence the result so produced will contain the uninterested data. The probabilistic database must be considered in order of their existence probability. The uninterested data, to some extent, is assumed to be among the interested data and hence it will also be displayed as the result of the search. This consumes more space of the result. It is also increases the time for producing the result. The map reduce algorithm mainly has 2 key functions. [3]They are "Map" and "Reduce". The Mapping is done by the mapper class and reducing is done by the Reducer class. The new algorithm will have the existing qualities but also will have additional features like Clustering, Classification and Anomaly Detection. Clustering means "A group of the same or similar elements gathered or occurring closely together". In the proposed system clustering is done keep all the similar kind of data in same place. The will reduce the search space and also the time required to produce the output.Classification refers to "a category into which something is put."Classification is supervised learning where as Clustering is unsupervised learning. In supervised learning, the output datasets are provided which are used to train the machine and get the desired outputs whereas inunsupervised learning no datasets are provided, instead the data is clustered into different classes.Anomaly Detection is also used together with the above both.

This increases the accuracy of the system. Anomaly Detection is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. The intersection of all these three will produce the output which will be very effective. All the unwanted results will be filtered and only the required one will be displayed. This makes the system user friendly and easily understandable.

VII. CONCLUSION

Existing system mainly focuses on obtaining the something related to the query rather than obtaining the exact result.It does not focus on the user's requirement. The algorithms used in the existing system are also traditional. In the proposed system, user is given the prime importance. The requirements of user is kept in focus while displaying the result.The result so produced is very effective and also useful.This is done with the help of basic knowledge in probabilistic database and existential probability. The probabilistic database refers to the database where required result will be present. The existential probability refers to the probability of data being present in the searched database. This results in very efficient searching.

REFERENCES

[1] "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools"- D. P. Acharjya ,Kauser Ahmed P
[2] "Review on Mining of Constraints Based Interesting Patterns from Uncertain Data Using MapReduceTechnique" - Darpan Kumari, Leena H. Patil, U. K Thakur
[3] "Reducing the Search Space for Big Data Miningfor Interesting Patterns from Uncertain Data"- Carson Kai-Sang Leung, Richard Kyle MacKinnon Fan Jiang