BIG DATA MINING

VIT University,Vellore

Yash Gupta-yashgupta.2013@vit.ac.in(SCOPE)

Mustafa Bohra-mustafa.bohra2013@vit.ac.in

Mastan Patel- Mmpatel.chandpatel2013@vit.ac.in

Jeet Kumar-jeet.kumar2013@vit.ac.in

Abstract-Everyday there are tons of transaction and data is transferred in huge amount. But this kind of data is difficult to maintain through the traditional data analysis. So the big data comes into action. But the main question is the creation of a platform which can analyse such data efficiently and finding out which mining algorithm to use to find the hidden patterns in that data. To discuss with this issue the papers starts with the introduction about Big Data and Mining and the algorithms which can be used, followed by review of researches done in past and how it can be useful in the future .

Keywords—Big Data, Data Mining, Risk Analysis,Patterns

I.INTRODUCTION

Every year we have been observing a dramatic increase in our ability to gather knowledge from varied sensors, devices, in several formats, from freelance or connected applications. This vast data has outpaced our capability to process, analyze, store and perceive these datasets. Consider the data on the internet. The internet pages indexed by Google were around 100 thousand in 1998, however quickly reached one billion in 2000 and have already exceeded one trillion in 2008. In 2016 it is around 1.3 trillion .And will be reaching approximately 2 trillion in year 2019 .This fast growth of data is speed up by the dramatic increase in acceptance of social networking applications, like Facebook , Twitter, etc., that permit users to form contents freely and amplify the already vast internet volume. Moreover , with mobile phones are becoming the sensory entryway to get real time knowledge on people from completely different aspects, the immense quantity of information that mobile carrier will doubtless method to enhance our existence has considerably outpaced our past call data record based process for charge functions solely. It may be foretold that Internet of things (IoT) applications can raise the size of data to new level. Folks and devices from home occasional machines to cars, to buses, railway stations and airports, area unit all loosely connected. Trillions of such connected elements can generate an enormous knowledge cloud, and valuable data should be discovered from the information to assist improve quality of life and create our world a much better place. For instance, once we tend to rise each morning, so as to optimize our commute time to figure and complete the optimization before we tend to attain workplace, the system has to process data from traffic, weather, construction, police activities to our calendar schedules, and perform deep optimization below the tight time constraints. All told these applications, we tend to face important challenges in investment the vast quantity of data as well as challenges in 1.system capabilities 2.recursive style 3.business models. This paper tends to introduce Bigdata processing and its applications in Section 2. Risk involved in section 3.Work of Researchers in this field in section 4.Future Work in section 5 and Conclusion in 6.

II.BIG DATA MINING AND APPLICATION

Big Data Mining is the procedure of examine vast information sets to discover concealed examples, obscure market trends, correlations, business information, customer inclinations and other helpful info. The mining can prompt to more successful showcasing, better client benefit, upper hands, enhanced operational effectiveness, over opponent associations, new income openings and different business benefits. The essential objective of enormous information mining is to bolster organizations settle on better business choices by prescient modelers,enabling information researchers, and other mining experts to mine vast volumes of exchange information, and also different types of information that might be unexploited by traditional business knowledge programs.[2] That could incorporate Internet click stream data,Web server logs and online networking substance and content from client emails, social organize action reports, and overview reactions, machine information captured, mobile-telephone call detail records by sensors associated with the IoT.[3]Huge Data Mining is essentially

utilized today by organizations with a solid buyer centre — retail, money related, correspondence, and showcasing associations, to penetrate down into their value-based information and decide evaluating, client inclinations and item situating, affect on deals, consumer loyalty and corporate benefits. With information mining, a retailer can utilize purpose of offer records of client buys to create items and advancements to engage particular client segments. Some utilization of enormous information are in Research Analysis, Corporate Surveillance, Fraud Detection, CRM.[2]

III.RISK INVOLVED

Mining Big Data includes chance in information obtaining: If information gained from source is not bonafide , ordering it utilizing calculation is as a part of vain. Henceforth, information source must be genuine to maintain a strategic distance from error in information investigation. Information pre-processing is additionally a hazard on the grounds that in procedure of expelling boisterous information, we may evacuate valuable information which will prompt to loss of information. Hazard Reduction: Data securing danger can be evaded by utilizing legitimate wellsprings of information to decrease odds of flawed information. Pre-processing danger can be evaded by utilizing proficient calculation or a device to evacuate uproarious information.[19]

IV.RESEARCH WORK

A. *Mining Large Streams of User Data for Personalized Recommendations by Xavier Amatriain (Netflix).*

The paper gives an till date overview of the use of data mining approach for recommendation and personalization. And also they have discussed about the various other data and machine learning techniques .Netflix is a platform which provides users to watch shows online or stream live videos. For the last 5 years netflix has been adopted by huge amount of people. So the data being received by them is unstructured and is very large. The reason for Netflix to be great hit because it uses a recommender system. Recommender systems are the sub-class of information system the predicts the shows or movies according to the user's previously watched shows.[7]Approach to recommendation system:

Here they have discussed about the various collaborative filtering (CF) algorithms which can be used to build one. The main assumption of these method are that people have same interest as their historical preferences and share similar taste in future. [16] The k-Nearest Neighbour (KNN) algorithm was the most favoured approach to CF, since it transparently captured this assumption of like-mindedness it operates by finding, for each user (or item), a number of similar users (items) whose profiles can then be used to directly compute recommendations.[4]Alternate approach to CF, content based approach (CBA)which identifies similarities between items based on the features inherent in the items themselves.CBA has a benefit over CF that it does not require historical data. There is another type of approach called Hybrid RS which is the combination of the above two approach. In practise most system use this type of approach.[4]Data Mining Methods in Recommender systems. A data mining task typically consists of 3 step, carried out in succession:

-Data Preprocessing

-Data Modeling

-Result Analysis

In this section, it describes some other models which can be used in designing the system likePrincipal Component Analysis,Decision Trees,Bayesian classifiers ,Artificial Neural Networks ,Support Vector Machines. Clustering approaches such as k-means can be used as a pre-processing. In next section they discussed about how they improved their system. And how there system became the best recommender system among the others.[5]The improved system comprised of 200hr of work and 107 different algorithm implementation. Testing of the system, bucket testing-is a slight variation from the traditional scientific process.

1.Start with a hypothesis: Algorithm/feature/design X will increase member engagement without service and ultimately member retention.

2.Design a test: Develop a solution or prototype. Think about issues such as dependent & independent variables, control, and significance.

3.Execute the test: Assign users to the different buckets and let them respond to the different experiences.

4.Let data speak for itself: Analyze significant changes on primary metrics and try to explain them through variations in the secondary metrics.[10]Collection and Management of data:

-As their website has a review option in which the user can provide the feedback of the show he/she has watched.

-Every day they receive millions of new ratings from members.

-Each item in catalog has rich metadata such as actors, director, genre, parental rating, or reviews

-tapping external data such as box office performance or critic reviews to improve our features

-Social data became the latest source of personalization features. Social data may include the social network connections themselves as well as interactions, or activities of connected nodes.[14]

B.Application of Big Data in Data Mining by SMITHA T, MCA, M.Phil, (PhD), V. Suresh Kumar, M.Tech CS

Big data is large quantity of data from many different sources which may be static or continuously generating in real time. The static sources include medical data, Simulation data, and business data of the fiscal year. Whereas the real time data is generated continuously from various social media applications like Facebook, Twitter, Instagram or astronomical data, weather reports etc. There are mainly 4 characteristics which need to be considered while handling big data. These are:-

1. Volume- which is the vast amount of data which is generated every second.

2. Velocity-that is how fast the data is being generated.

3. Variety-The different form or type of data. It may be structured or unstructured; real time or static; different format of data like text, images or videos etc.

4. Veracity-The validity of the data. Its inconsistencies, errors and completeness need to be checked.[6]

Conventional tools which are used will not be able to extract different information from these data, and moreover they will also not be able to handle the continuous large amount of data generated. Therefore we need new kind of technology, platform which can capture significantly large amount of incoming data so that it can be processed, analyzed, visualized, stored and shared.Moreover connection and correlation of these data has to be found. This can be done using data mining.

In Data Mining, different data repositories have to be included so that it can manage any form of data. Data mining techniques are used in object relational system, so that it can be used to find patterns or trends in these objects. For example, sales record of previous years of a large E-commerce company can be used to find the buying trends over the years. Similarly geographical databases are used for environmental and ecological planning,astronomical data is used to predict paths of different objects moving through space. There is also a spatiotemporal database that changes with time from which information can be mined.[17]There are also different types of data mining systems which perform various techniques.

Classification system- These are used to classify different types of data to create data classes that can be distinguished. They can then be used to predict the class of unknown data. Basically training the machine to create a model by giving it data and predicting classes of new data.

Evolution analysis- These types of systems are used in identifying changes in data over period of time and creating a model. They are used to predict the future changes which may occur using this model. Used in stock markets, E-commerce industry etc.

Outlier analysis- These are used to identify the data which do not follow certain trend or pattern which most of them seem to follow. They can be used to detect exceptional or fraudulent data.[12]

Cluster analysis- Different data are grouped together based on their similarity and no labels are used in training data sets. Then rules are formed from these clusters. These methods consist of portioning methods, hierarchical methods, density based methods etc.[13]

There are many new tools developed to handle big data. Hadoop MapReduce is upcoming programming model. It is a batch query processor and can run an ad hoc query for whole data set to get the results in a transformative sensible way[7]. It does this in two steps. First, queries are divided into sub-queries and assigned to different nodes which run in parallel to process it. Second, these results are assembled and then delivered. Similarly Oracle has introduced the total solution for the scope of enterprise which requires Big Data. Oracle Big Data Appliance is a tool to integrate optimized hardware and extensive software into its database to endure big data challenges.[17]

Data Mining has to be performed in Big Data to identify current trends and patterns in case of businesses; for better operation performance; increasing customer base even to predict calamities using geographical data.

C.Mining Big Data in Real Time by Albert Bifet .Yahoo! Research Barcelona,Catalonia.

Nowadays, the quantity of data that is created every two days is estimated to be 5 extra bytes. This amount of data

is similar to the amount of data created from the dawn of time up until 2003Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring and traffic management, log records and many more. In fact, all data generated can be considered as in data stream mining. we are interested in three main dimensions:[12]

Accuracy-Amount of space necessary. The time required to learn from training examples and to predict

New problems-A new important and challenging task may be the structured pattern classification problem. Patterns are elements of sets endowed with a partial order relation. Examples of patterns are item sets, sequences, trees and graphs Most standard classification methods can only deal with vector data, A way to deal with a structured output classification problem is to convert it to a multi label classification problem, where the output pattern y is converted into a set of labels representing a subset of its frequents sub patterns. Therefore, data stream multi-label classification methods may offer a solution to the structured output classification problem.[9]

New applications-A future trend in mining evolving data streams will be how to analyze data from social networks and micro-blogging applications such as Twitter. Micro-blogs and Twitter data follow the data stream model. The main Twitter data stream that provides all messages from every user in real time is called Firehose and was made available to developers in 2010. This streaming data opens new challenging knowledge discovery issues. Twitter's search engine received around 600 million search queries per day, and Twitter received a total of 3 billion requests a day via its API. It could not be clearer in this application domain that to deal with this amount and rate of data, streaming techniques are needed. Sentiment analysis can be cast as a classification problem where the task is to classify messages into two categories depending on whether they convey positive or negative feelings. Mining techniques to build classifiers for sentiment analysis, we need to collect training data so that we can apply appropriate learning algorithms. a significant advantage of Twitter data is that many tweets have author-provided sentiment indicators, changing sentiment is implicit in the use of various types of emoticons. Smiley's or emoticons are visual cues that are associated with emotional states. They are constructed using the characters available on a standard keyboard, representing a facial expression of emotion. Hence we may use these to label our training data. When the author of a tweet uses an emotion, they are annotating their own text with an emotional state. Such annotated tweets can be used to train a sentiment classifier.[20]

New techniques-A way to speed up the mining of streaming learners is to distribute the training process onto several machines. Hadoop MapReduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes [7]. The step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job. Apache S4 is a platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time. Storm from Twitter uses a similar approach. Ensemble learning classifiers are easier to scale and parallelize than single classifier methods. They are the first, most obvious, candidate methods to implement using parallel techniques.[16]

We discussed the challenges that evolving data streams will have to deal during the next years. These include structured classification and associated application areas as social networks. Our ability to handle many Exabyte's of data across many application areas in the future will be crucially dependent on the existence of a rich variety of datasets, techniques and software frameworks. There is no doubt that data stream mining offers many challenges and equally many opportunities as the quantity of data generated in real time increases.[17]

V.FUTURE WORK

Huge information is critical in business world. The hugeness of huge information is gotten from the information gathered through different information sources.[13] Beforehand business world absolutely depended on organized information gathered and put away in a customary database. Information gathered from online networking and Internet gives 90% unstructured information in different arrangements like XML records, word documents, PDF, content, email, recordings, pictures and so forth which is imperative to take business choices. Investigation of these information can give new bits of knowledge for administrators. It will give more precise, pertinent and nitty gritty information permitting top level administration to investigate execution changeability, valuable patterns, concealed behavioral examples and plan activity plan to hold existing clients and pull in new one by minimizing dangers. New item, administrations,

techniques and plans of action can create from examination of enormous information which will enhance the personal satisfaction.[11]

VI.CONCLUSION

Hence Big Data is the next big thing and will continue to be for next few decades. New ways of handling and using enormous of data will have to be found, as the data is going to be much diverse and large. Implementation of data mining in big data gives us much more development in research department as well as in the technical. Many opportunities as well as challenges will occur as the Big Data becomes 'bigger'. This paper presents some current techniques being used and future possible work which can be done.

REFERENCES

[1]https A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter streaming data. In Proc 13th International Conference on Discovery Science, Canberra, Australia, pages 1-15: Springer, 2010.

[2] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[3] R. Ahmed, G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks", *Knowledge and Information Systems*, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[4] A. Bifet, G. Holmes, B. Pfahringer, and E. Frank. Fast perceptron decision tree learning from evolving data streams. In PAKDD, 2010.

[5] J. Gama. Knowledge discovery from data streams. Chapman & Hall/CRC, 2010.

[6] B. Liu. Web data mining; Exploring hyperlinks, contents, and usage data. Springer, 2006.

[7] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In ICDM Workshops, pages 170-177, 2010.

[8] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1- 2):1-135, 2008.

[9] A. K. Choudhary, J. A. Harding and M. K. Tiwari, "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge", Journal of Intelligent Manufacturing, Volume 20, Number 5, 501-521, 2008.

[10] Arijay Chaudhry and Dr. P.S.Deshpande. Multidimensional Data Analysis and Data Mining, Black Book.

[11] Smitha.T,Dr.V.Sundaram, "Knowledge Discovery from Real Time data base using data mining technique", International journal of Scientific Research and Publication, (IJSRP) vol 2, issue 4, April 2012, ISSN 2250-3153, pp .74-76.

[12] Smitha.T,Dr.V.Sundaram, "Comparative Study of Data Mining Algorithms for High Dimensional Data Analysis"- published in International journal of Advances in Engineering & Technology (IJAET) ISSN2231-1963 173 in vol 4, issue2, sept 2012 PP 15-20.

[13] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Companies, Incorporated, 2011.

[14] S. M. Weiss and N. Indurkhya. Predictive data mining: a practical guide. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[15] J. Gama. Knowledge Discovery from Data Streams. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.

[16]U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling.. http://big-data-mining.org/keynotes/#fayyad, 2012

[17]C. C. Aggarwal, editor. Managing and Mining Sensor Data. Advances in Database Systems. Springer, 2013.

[18]Intel. Big Thinkers on Big Data, http://www.intel.com/content/www/us/en/bigdata/

big-thinkers-on-big-data.html, 2012.

[19]E. Letouze. Big Data for Development: Opportunities & Challenges. May 2011.

[20]N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2013.