ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Data Mining Techniques For Big Data

Ahmed UnnisaBegum, Mohammed Ashfaq Hussain , Mubeena Shaik

Lecturer, Jazan University, Kingdom of Saudi Arabia
Research Scholar, Acharya Nagarjuna University, India

ABSTRACT: Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems to identify future patterns. Big Data is a term used for any data that is large in quantity. It is used to refer to any kind of data that is difficult to be represented using conventional methods like Database Management Systems or Microsoft Excel. In this paper we are discussing the characteristics applications of Big Data processing model and Big Data revolution, from the data mining view.

KEYWORDS: Data mining, Bid Data, Data Security, Data availability.

I. INTRODUCTION

Data mining refers to the activity of going through big data sets to look for relevant or appropriate information. The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected[2]. Decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business. Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, where individual workers send specific queries for information to database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific working year[1]. The importance of Big Data does not mean how much data we have but what would you get out of that data. We can analyze data to reduce cost and time, smart decision making.

II. DATA MINING AND BIG DATA

Data mining, also known as data discovery or knowledge discovery, is the process of analyzing data from different viewpoints and resulting it into useful information. This information is used by businesses to increase their revenue and reduce operational expenses. The software programs used in data mining are amongst the number of tools used in data analysis.The software enables users to analyze data from different point of views, classify it and make a summary of the data trends identified[3][4]. Technically, data mining involves the process of discovering patterns or relationships in large areas of related databases.The actual data mining task is the automatic or semi-automatic analysis of large datasets. This is done to assist in the extraction of previously unknown and unusual data patterns. These include detecting abnormalities in records, cluster analysis of data files and sequential pattern mining. Database techniques like spatial indices are commonly used in these processes.
After these processes, the patterns can be seen as the summary of the input data and can be used in further analysis like predictive analytics or machine learning. For instance, multiple groups of data can be identified through data mining steps. This is the process of analyzing larger data sets with the aim of uncovering useful information. Examples of this information include market trends, customer preferences, hidden patterns and unknown correlations. The analytics findings usually lead to new revenue opportunities, improved operational efficiency, more efficient marketing and other business benefits[2]. Companies often rely on big data analytics to assist them in making strategic business decisions. Big data analytics enable data scientists, predictive modelers and other professionals in the analytics field to analyze large volumes of transaction data. They can also use big data analytics to analyze data which might not have been discovered by conventional business programs.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

III. CHALLENGES TO HANDLE BIG DATA

The programmers have to take decisions due to large availability of raw and complex data.An organization can collect, store, and analyze these large datasets in a number of ways. The Business can even use robust big data tools to store, access, and manage the structured and unstructured data collected from various sources in a faster and more efficient way. There are few challenges to address when handling big chunks of data. Some challenges listed below:

A. Handling a Large Amount of Data

The large availability of data makes the difficulty is making decisions.Data that enterprises can access has been increased exponentially from last several years. They have data for everything, right from what a consumer likes, to how they react, to a particular scent, to the amazing restaurant that opened up in Italy last weekend.
This data exceeds the amount of data that can be stored and computed, as well as retrieved. The challenge is not so much the availability, but the management of this data[5]. Along with rise in unstructured data, the availability of data is in multiple formats such as video, audio, social media, smart device data etc. Some of the newest ways developed to manage this data are a hybrid of relational databases combined with NoSQL databases. An example of this is MongoDB, which is an inherent part of the MEAN stack. There are also distributed computing systems like Hadoop to help manage Big Data volumes.

B. Data Security

In increasing of data, the major issue is to secure the data. Many organizations claim that they face trouble with Data Security. This happens to be a bigger challenge for them than many other data-related problems. The data that comes into enterprises is made available from a wide range of sources, some of which cannot be trusted to be secure and compliant within organizational standards.They need to use a variety of data collection strategies to keep up with data needs. This in turn leads to inconsistencies in the data, and then the outcomes of the analysis[6]. This data is made available from numerous sources, and therefore has potential security problems. You may never know which channel of data is compromised, thus compromising the security of the data available in the organization, and giving hackers a chance to move in. Now it is essential to introduce Data Security best practices for secure data collection, storage and retrieval.

1. Data Complexity

With the huge updating in data in every second, organizations need to be aware of handling it too. For example, if a retail company wants to analyze customer behaviour, real-time data from their current purchases can help. There are Data Analysis tools available for the same – Veracity and Velocity. They come with ETL engines, visualization, computation engines, frameworks and other necessary inputs.It is important for businesses to keep themselves updated with this data, along with the “stagnant” and always available data. This will help build better insights and enhance decision-making capabilities.

2. Shortage of Skilled Resources

There is a shortage of skilled Big Data professionals available at this time. This has become mentioned by many enterprises seeking to better utilize Big Data and build more effective Data Analysis systems. There is a lack experienced people and certified Data Scientists or Data Analysts available at present, which makes the “number crunching” difficult, and insight building slow.Again, training people at entry level can be expensive for a company dealing with new technologies. Many are instead working on automation solutions involving Machine Learning and Artificial Intelligence to build insights, but this also takes well-trained staff or the outsourcing of skilled developers.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

IV. DATA MINING TECHNIQUES TO HANDLE BIG DATA V.

Data mining techniques have been around for many years in combination with data warehouses, and have now taken on greater prevalence with the advent of Big Data. Data analytics and the growth in both structured and unstructured data has also prompted data mining techniques to change, since companies are now dealing with larger data sets with more varied content. Additionally, artificial intelligence and machine learning are automating the process of data mining[7]. Some of the techniques which are listed below:

A. Association

Association makes a correlation between two or more items to identify a pattern. For instance, a supermarket could determine that customers often purchase whipped cream when they buy strawberries and vice versa. Association is often used at point-of-sale systems to determine common tendencies among products. "It's a very simple method, but you'd be surprised how much intelligence and insight it can provide—the kind of information many businesses use on a daily basis to improve efficiency and generate revenue," according to technology company Galvanize. Application areas include physical organization of items, marketing and the cross-selling and up-selling of products.

B. Classification: Multiple attributes can be used to identify a particular class of items. Classification assigns items into target categories or classes to accurately predict what will occur within the class. Several industries use classification with customers. For instance, a banking company could use a classification model to identify loan applicants as low, medium or high credit risks. Other organizations classify current and target audiences into age and social groups for marketing campaigns.

C. Clustering: "Clustering is the method by which like records are grouped together," according to Alex Berson, Stephen Smith and Kurt Thearling in the book Building Data Mining Applications for CRM. "Usually this is done to give the end user a high level view of what is going on in the database." Seeing object groupings can help businesses in areas like marketing segmentation. Clustering can be used in this example to subdivide a market into subsets of customers[7][8]. Each subset can then be targeted with a specific marketing strategy based on the attributes of the cluster, such as buying patterns for customers in one cluster vs. another cluster.

D. Decision Trees: Decision trees are used to categorize or predict data. A decision tree starts with a simple question that has two or more answers. Each answer leads to a further question that is used to classify or identify data that can be categorized, or so that a prediction can be made based on each answer.The graphic of a decision tree represents how a cell phone provider might classify customers who combine, or those who don't renew their phone contracts. The authors of Building Data Mining Applications for CRM offer some interesting takeaways for the graphic. It divides the data into each branch without losing any of the data. For instance, the total number of records in a parent node is equal to the sum of the records contained in its two children.

E. Sequential Patterns: Sequential patterns identify trends or regular occurrences of similar events. This data mining technique is often used to understand user buying behaviors. Many retailers use data and sequential patterns to decide on the products they display[6]. "With customer data you can identify that customers buy a particular collection of products together at different times of the year," according to IBM. "In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history."

VI. CONCLUSION

In this paper we discussed the basic data mining techniques to handle Big data. Big data is evolving with the exponential rise in data availability. It is important for the organisations to work around these challenges and gain advantages over their competition with more reliable insights,. Thus, more studies should be addressed towards big data analytics to manage business with huge amount of data.

REFERENCES:

[1] S. San M. Negnevitsky, N. Hatziargyriou, "Applications of Data Mining and Analysis Techniques in Wind Power Systems", 42440178X/06/$20.00 ©2006 IEEE.
[2] Krioukov, Andrew, "Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities." IEEE Data Eng. Bull. 34.1 (2011): 3-11.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

[3] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.
[4] Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36, 4, pp. 1165-1188.
[5] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), "Big Data Framework" 2013 IEEE International Conference on 13-16 Oct. 2013, 1494-1499.
[6] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intelligence And Analytics: From Big Data To Big Impact, Big Data Analytics An Oracle White Paper", MIS Quarterly vol. 36 no. 4, pp. 1165-1188/December 2012.
[7] Anastasia, February 2015. Big data and new product development. Entrepreneurial Insights http://www.entrepreneurial-insights.com/big-data-new-product-development/, Accessed on June 15, 2015.
[8] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24.