DATA MINING WITH CLUSTERING ON BIG DATA FOR SHOPPING MALL'S DATASET

Fatema Jamnagarwala
Department of Civil Engineering, Sipna College of Engineering & Technology, Amravati, India
fatema.jamnagarwala1993@gmail.com

Dr.P.A.Tijare
Department of Computer Science & Engineering, Sipna College of Engineering & Technology, Amravati, India
pritishtijare@rediffmail.com

Manuscript History
Number: **IJIRAE/RS/Vol.06/Issue01/JAAE10088**
Received: 07, January 2019
Final Correction: 15, January 2019
Final Accepted: 26, January 2019
Published: **January 2019**
Citation: Fatema,Tijare (2019). Data mining with clustering on big data for shopping malls dataset. IJIRAE:: International Journal of Innovative Research in Advanced Engineering, Volume VI, 27-31. **doi://10.26562/IJIRAE.2019.JAAE10089**
Editor: Dr.A.Arul L.S, Chief Editor, IJIRAE, AM Publications, India

Abstract: Big Data is the extremely large sets of data that their sizes are beyond the ability of capturing, managing, processing and storage by most software tools and people which is ever increasing day-by-day. In most enterprise scenarios the data is too big or it moves too fast that extremely exceeds current processing capacity. The term big data is also used by vendors, may refer to the technology which includes tools and processes that an organization requires to handle the large amounts of data and storage facilities. This advancement in technology leads to make relationship marketing a reality for today's competitive world. But at the same time this huge amount of data cannot be analyzed in a traditional manner, by using manual data analysis. For this, technologies such as data warehousing and data mining have made customer relationship management as a new area where business firms can gain a competitive advantage for identifying their customer behaviors and needs. This paper mainly focuses on data mining technique that performs the extraction of hidden predictive information from large databases and organizations can identify valuable customers and predicts future user behaviors. This enables different organizations to make proactive, knowledge-driven decisions. Data mining tools answer business questions that in the past were too time-consuming, this makes customer relationship management possible. For this in this paper, we are trying explain the use of data mining technique to accomplish the goals of today's customer relationship management and Decision making for different companies that deals with big data.

Keywords: Big Data; Data mining; Knowledge discovery in databases (KDD); Classification; Clustering;

I. INTRODUCTION

Big Data itself termed as extremely large datasets that their sizes are beyond the ability of capturing, managing, and processing by most software tools and people [1]. For example, Search engines, social networking, and online Advertising, ecommerce, as well as education, healthcare, and medicine, etc. As the datasets are so long it have to face the challenges including capture, storage, [2] search, sharing, transfer, analysis [3] and visualization. The trend are becoming to larger as the data sets are useful for the additional information deliverables from analysis of a single large set of related data, as compared to separate smaller sets with identical total quantity of knowledge, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases in medical sector, combatcrime, and determine real-time roadway traffic conditions or other real-time applications [4]."

The advent of knowledge technology in varied fields of human life has crystal rectifier to the massive volumes of knowledge storage in varied formats like records, documents, images, sound recordings, videos, scientific data, and many new data formats. For better decision making, the data collected from different applications require proper mechanism of extracting useful knowledge/information from large data repositories [5]. Data mining, often called as Knowledge discovery in databases (KDD), aims at the discovery of useful information from large collections of data [6]. This Data mining using its different algorithms generate useful information from massive collection of data and that are now a day's called as Big Data.

Data mining uses different techniques such as statistical, mathematical, artificial intelligence and machine learning [1] as the computing techniques. This works as interdisciplinary subfield of computer science, and provide highly targeted information to support decision-making and forecasting for many important scientific, physiological, sociological, the military and business decision making. These are vital important fields of our life hence needing progress [3].Its prophetic power comes from distinctive style by combining techniques from machine learning, pattern recognition, and statistics to automatically extract concepts, and to determine the targeted interrelations and patterns from large databases [6].Organizations get facilitate to use their current news capabilities to find and determine the hidden patterns in databases. The extracted styles from the information are then wont to build data processing models, and may be wont to predict performance and behavior with high accuracy. There are always developments of new business culture, in that the economics of customer relationships are changing in fundamental ways. Companies are facing the need to implement new solutions and strategies that address the changes that are normally required by current market strategies.

The ideas of production and mass selling, first created during the Industrial Revolution, are being supplanted by new ideas in which customer relationships are the central business issue. Firms nowadays area unit involved with increasing client price through analysis of the client lifecycle. The tools and technologies of information deposit, data mining, and other customer relationship management (CRM) techniques afford new opportunities for businesses to act on the concepts of relationship marketing. This paper mainly focuses on use of Big data for managing customer relationships. It may seem that CRM is applicable only for managing relationships between businesses and their consumers. But the closer examination reveals that it is even more crucial for business customers. In business-to-business (B2B) environments, an amazing quantity of knowledge is changed on a daily basis .For example, transactions are a lot of various, custom contracts are a lot of numerous, and valuation schemes are a lot of sophisticated. CRM helps sleek the strategy once various representatives of merchant and customer companies communicate and collaborate. Tailored catalogues, personalized business portals, and targeted product offers will change the acquisition method and improve growth and development of each the businesses. E-mail alerts and new product data tailored to completely different roles for the patron that ends up in increase the effectiveness of the pitch. Trust and authority unit of measurement augmented if targeted tutorial reports or news unit of measurement delivered to the relevant individuals. All of these is assumed of among the benefits of CRM.

II. RELATED WORK

As in the research of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of Exabyte of data [7]. The limitations also affect Internet search, finance and business informatics [8].Data sets grow in size partly as a result of they are progressively being gathered by present information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks [9] [10]. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s [11], as of 2012, every day 2.5 Exabyte of data were created [12]. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization [13].

Big data is difficult to manage with most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers" [14].As the thought of "big data" varies looking forward to the capabilities of the organization that manages the set, and on the capabilities of the applications that sq. measures to methodology and analyze the information set in its domain."For some organizations, facing many gigabytes knowledge| of information for the primary time might trigger a necessity to rethink data management choices. For others, it ought to take tens or several terabytes before data size becomes an enormous thought." [15].

III. ISSUES OF MINING METHODOLOGY

As the big data is of much large volume and size to extract useful knowledge from it using Data mining techniques some issues and limitations are pertain to its approaches. Some another points that needs to be consider are the versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs, the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, etc [16].Above square measure all such examples that may be as dictate mining methodology decisions.

There can also some different approaches may suit and solve user's needs differently. Most algorithms assume the data to be noise-free. This is in fact a powerful assumption that is needed. Many datasets have exceptions, invalid or incomplete information, etc., which might complicate, if not obscure, the analysis process and in many of the scenarios compromise the accuracy of the results. As a consequence, data preprocessing and transformation becomes vital It is typically seen as lost time, but data cleaning, as time consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process [17].Data mining techniques ought to be able to handle noise in information or incomplete info. More than the dimensions of knowledge, the size of the search space is even more decisive for data mining techniques. This is known as the curse of dimensionality. This "curse" affects thus badly the performance of some data processing approaches that it's changing into one in every of the foremost imperative problems to resolve.

A. Challenges of Security

As the Big data taken from most of the online portals frequently contains huge amounts of personal identifiable information, personal account information, health related data, etc. therefore privacy of users is a huge concern. That needs to be secure from other large number of attributes. That creates the biggest challenge for big data from a security point of view called as protection of user's privacy [18][19]. Because of this a big data security breach will potentially affect a much larger number of people. This in itself is often a security challenge as removing distinctive identifiers may not be enough to ensure that the info can stay anonymous. When storing the data organizations will face the problem of encryption [20].Data cannot be sent encrypted by the users if the cloud has to perform operations over the info. While victimization huge knowledge a big challenge is a way to establish possession of knowledge. If the data is stored in the cloud a trust boundary should be establish between the data owners and the data storage owners [21]. An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default [22].This makes the matter of access management worse, as a default installation would go away the knowledge receptive unauthenticated users.

Following are some if the general security recommendations that can be applied to big data:

- If we are store big data in the cloud, one must ensure that provider has adequate protection mechanisms in place. We should make sure that the supplier carries out periodic security audits and agree penalties just in case that adequate security standards are not met.
- Create an adequate access control policy that allow access to authorized users only leading to secure access of data.
- Protect both the raw data and the outcome of the analytics should be adequately protected. Encryption should be used accordingly that ensure no sensitive data is leaked.
- Protect data in transit should be adequately protected to ensure its confidentiality and integrity.
- It is necessary to perform real-time security monitoring while accessing the data. As the AI is growing on increasing threat intelligence should be used to prevent unauthorized access of data.
- When producing information for big data, it should be adequately anonymised, removing any unique identifier for a user. A solution to the problem of encryption is to use "Fully Homomorphic Encryption" (FHE), which allows data stored in the cloud to perform operations over the encrypted data so that new encrypted data will be created.
- Applying adequate access control mechanisms becomes the key in protecting the data. Here the approach is to protect the information using encryption / cryptographic techniques that only allows decryption if the entity trying to access the information is authorized by an access control policy.
- Big data solutions often depend on traditional firewalls or implementations at the application layer to restrict access to the information.

B. Performance issues:

Many artificial intelligence and statistical methods exist for data analysis and its interpretation. However, these methods were often not designed for the very large data sets, for with data mining is dealing today. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large amount of data [23].Algorithms with exponential and even medium-order polynomial quality can't be of sensible use for data processing. One technique that's, sampling can be used for mining instead of the whole dataset.

However, concerns such as completeness and choice of samples may arise [24].The issue of performance additionally encompasses progressive change, and parallel programming. It is rightly said that parallelism can help resolve the size difficulty if the dataset can be divided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without necessary to re-analyze the complete dataset.

IV.APPLICATIONS OF BIG DATA

A. Applications of Big Data in Business and Organizations

Data mining is primarily used today by companies with online business and a strong consumer focus - retail, financial, communication, and marketing organizations. It permits these corporations to see relationships among "internal likewise as external". Internal factors like value, product positioning, or employees skills, and external factors such as economic indicators, competition, and customer demographics. And, it permits them to see the impact on sales, client satisfaction, and corporate profits. Finally, it also enables them to "drill down" into summary information to view detail transactional data and perform decision making helps in increasing their profit margins and growth of business. With data processing, a merchant might use location records of client purchases to send targeted promotions supported somebody's purchase history. By mining demographic information from comment or pledge cards, the merchant might develop product and promotions to attractiveness to specific client segments. For example, WalMart is pioneering large data processing to rework its provider relationships. WalMart holds from over 2,900 stores in 6 countries point-of-sale transactions and transmits this data to its massive 7.5 terabyte in data warehouse. WalMart permits quite three, 500 suppliers, to access information on their product and perform information analyses. Businesses using data processing may even see a come on investment, however conjointly they acknowledge that the quantity of prophetical models will quickly become terribly giant. Data mining are often useful to human resources (HR) departments in distinctive the characteristics of their most booming workers. Data mining could be an extremely effective tool within the catalog promoting trade. Catalogers have a chic info of history of their client transactions for various customers chemical analysis back variety of years .Data mining tools will determine patterns among customers and facilitate determine the foremost doubtless customers to retort to imminent mailing campaigns.

V. PROPOSED WORK

Customer segmentation is one amongst the foremost necessary data processing methodologies utilized in selling and CRM. It helps business firms to find the characteristics of their customers and build them derive acceptable selling activities per the knowledge discovered. This paper principally focuses of the employment knowledge of information mining techniques for analyzing the massive data generated kind on-line looking portals that's one amongst the foremost growing business lately. In data processing language the various classification and bunch algorithms ar accustomed classify info from the massive assortment of knowledge set. The bunch rule appearance for clusters that's the cluster of comparable info within the knowledge. It finds sets of cases that are additional kind of like each other than they're to cases in other sets. Here during this paper, the most technique used activity bunch that's grouping similar customers along, supported many alternative criteria. During this means it's attainable to focus on each cluster of shoppers counting on their characteristics. Bunch the client helps firms to develop acceptable selling campaigns and rating methods. For instance, it's attainable to supply a special value or free some units to a definite teams of clients or consumers at an equivalent time while not characteristic the one customer at an equivalent time. Using bunch we tend to try and analyze giant and sophisticated set of knowledge. By applying bunch technique, the analyst will break down an outsized drawback into variety of teams with common characteristics. Since every cluster provides an outline, the analyst will perceive the character of the matter higher. However the analyst should experiment with the model's variables. By making use of this techniques users will mine knowledge as, behavioral knowledge that helps one to spot teams of shoppers World Health Organization have similar shopping for behaviors. During this means it's attainable to target what customers do instead of what they're.

VI. CONCLUSION

This paper performs the study of usefulness of big data with Data mining for performing Customer Relationship Management (CRM). The study shows that CRM is an important technique that cannot be practiced in business without tracking patterns within customer data. And as identification of user behaviors is very important for business organization to perform decision making, it becomes necessary to use the advancement of Information technology to store huge amount of data and then apply the data mining technique to extract the useful pattern from these big data storage. This paper also performs the study that the size of the database for identifying customer data for CRM today can range to terabytes to zeta bytes, which makes it almost impossible to store and analyze in a traditional manner with manual system. For this reason, Big data is necessary to store this data properly and then data mining has attracted a great deal of attention to perform analysis and making some useful business decisions for CRM. Since the goal of data mining is extracting meaningful patterns and relationships from large data sets, data mining can redefine and improve customer relationships in many or all types of businesses. This paper studies the overall concepts of Big data, data mining and issues identified with growing amount of data. Finally, this paper proposes the Data mining techniques to the customer relationship management, that helps to the concept of decision making with the help of Big data collected from large business organizations.

REFERENCES

1. Efraim, T.; Jay, E. A.; Tin-Peng, L. & Ramesh, S. (2007). Decision Support and Business Intelligent Systems, Pearson Education.
2. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic. "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.
3. D. Krishna. "Big Data". http://www.irmac.ca/1011/Big%20Data%20v2.1.pdf
4. Kusnetzky, Dan. "What is "Big Data?"". ZDNet.
5. Vance, Ashley (22 April 2010). "Start-Up Goes After Big Data With Hadoop Helper". New York Times Blog.
6. "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.
7. "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.
8. Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.
9. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.
10. Edelstein, H. (1997). Data mining: Exploring the hidden trends in your data. DB2 Online
11. Magazine. Available: http://www.db2mag.com
12. Chris Rygielski, Jyun-Cheng Wang, David C. Yen, "Data mining techniques for customer relationship management", Technology in Society 24 (2002) 483–502, [online] www.elsevier.com/locate/techsoc
13. D. Ćamilović, "DATA MINING AND CRM IN TELECOMMUNICATIONS", Serbian Journal of Management 3 (1) (2008) 61 – 72.
14. Sumathi, S, Sivanandam S.N., "Introduction to Data Mining and its Applications", Springer, Berlin (2006).
15. Weiss, G. M., "Data Mining in Telecommunications", Data Mining and Knowledge Discovery, Springer Science, New York, pp.1189-1201, (2005).
16. Berry, M. and Linoff, G., "Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management" - Second Edition, Wiley Publishing Inc., Indianapolis, 2004.
17. David Floyer,"Enterprise Big-data",on Nov21,2014. http://www.reportsnreports.com/reports/288019-the-big-data-market-2014-2020-opportunities-challenges-strategies-industry-verticals-and-forecasts.html
18. http://www.linkedIn.com/The-Eye-Opening-Facts-Everyone-Should-Know_BernardMarr_LinkedIn.html
19. http://www.csc.com/business_drivers/offerings/82042-big_data_storage_solutions
20. https://www.mwrinfosecurity.com/articles/big-data-security---challenges-solutions/
21. Apache Hadoop. Available at http://hadoop.apache.org
22. http://cloud.asperasoft.com/big-data-cloud/
23. http://www.sei.cmu.edu/uls/Addressing-the-Software-Engineering-Challenges-of-Big-Data»SEI log.html.