A Study of Current State of Work and Challenges in Mining Big Data

Kaushika Pal
Assistant Professor, Sarvajanik College of Engineering and Technology, Surat
Email: kaushikapal@scet.ac.in
Dr. Jatinderkumar R. Saini
Associate Professor and Director I/C, Narmada College of Computer Application, Bharuch
Email: saini_expert@yahoo.com

--**ABSTRACT**--

Big data is term for any collection of data sets very huge and complex that it becomes difficult to process using traditional data processing applications. With data mining software tools they are difficult to manage. Big data mining has the capability of extracting useful information from large data sets which was not earlier possible due to its complexity. The Big Data Mining challenge is one of the thrilling prognoses in coming years. In this paper, we discuss current status of Mining Big data and challenges in mining big data in coming years

Keywords: **Big Data, Garbage Mining, Data Management, Data Mining, Knowledge Discovery**

1. Introduction

Huge amount of data are generated and collected from various sources like sensors, devices etc. all are in different formats from connected or independent application [8]. This data has to be processed, investigated, stored and understood. Considering internet data the web pages indexed by Google were One million in 1998, One billion in 2000 and one trillion in 2008 [9].

This expansion of data has resulted due to use of internet facilities by various sites which allow public to store their data on web, example social networking sites like twitter, Face book, etc, which allows public to expand the already huge web volume.

Smartphones are now highly connected to internet and use and store data on web and thus increasing web volume. Smartphones becoming sensory gateway to get real time data on people from different traits, the vast amount of data that mobile carrier process have gone beyond call data record based processing for billing purpose only. Smartphones are the real producer of big data, and it is up to us how we can utilize that data to change our lives. Now BILLIONs of smartphones is in use generates data [10]. Every text, every phone call, every search every email and every picture or video you upload or share is stored [10]. If we consider each smartphone user will generate about 60 gigabytes of data each year, times the six billion devices (only mobile phones), we generate and store more than 335 Exabyte of information every year with Smartphone's alone [10]. Data created via smartphones can be put to good use. Smartphone usage patterns helped researchers in Africa determine where malaria outbreaks were occurring and where the affected people went [10]. This information can be used to determine where to best distribute medicines more efficiently. This is the power of big data analysis which has a positive impact on humanity. Soon, as mobile devices are used more frequently to purchase goods and services, the information produced will be mined to determine what are your interests, where you go to shop, and even what brand of coffee you like, so advertisers and others can identify your wants and desires. In all this, we are facing a significant challenge in leveraging the vast amount of data, including challenges in System Capabilities, Algorithmic design, Business Models [8]

Each day Google has more than 1 billion inquiries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day.[11] The data produced is estimated to be in terabytes, and is growing around 40% every year [11]. Large amount of data is also generated by smartphones and big companies are starting to look carefully to this data to find useful patterns to improve user experience.

We need new tools and new algorithm to deal with all this huge amount of data. While working with Big Data 7 V's have to be considered for Big Data Management [14]

Volume: 100 terabytes of data are uploaded daily to Facebook [13]; Wal-Mart handles 1 million customer transactions every single hour [12]. The massive use of internet brings torrent of social media updates, sensor data from devices and a burst of e-commerce, means that every industry is flooded with data, which can be extremely valuable, if it can be used to retrieve important information.

Variety: 90% of data generated is amorphous coming in all shapes and forms-the data is generated from geo-spatial, tweets, photos and videos uploading on social networking sites, which can be analysed for content [13]

Velocity: Velocity' refers to the increasing speed at which this data is created, and the increasing speed at which the data can be processed, stored and analysed [13]

Value: The probable value of Big Data is huge. The value lies in severe analysis of accurate data, and the information and understandings this provides.

Variability: Variability refers to data whose meaning is constantly changing. There are changes in the structure of data and how users want to interpret that data.

Veracity: Although we agree with importance of Big Data, the data is of no use if it's not accurate. Big Data Veracity refers to the biases, noise and abnormality in data. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems.

Visibility: Data from different sources should be visible to the technology stack making up Big Data. Certain data which are crucial are available but not visible to Big Data.
All these are assets which demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

2. Literature Review

Big Data mining infrastructures and the experience of doing analytics at twitter shows that due to the current state of the data mining tools, it is not simple to perform analytics. The preparation works consumes most of the time, using application of data mining methods. Converting or using these preliminary models of data mining into strong solutions is time consuming and increase complexity [1].

Mining heterogeneous information networks is a new and promising research frontier in Big Data mining. It considers interconnected, various different types of data, including the relational database data, as heterogeneous information networks. These semi-structured heterogeneous information network models leverage the rich semantics of typed nodes and links in a network and can uncover surprisingly rich knowledge from interconnected data [2].
Mining Big Data in Real Time discusses the challenges in structured pattern classification. The classification methods mostly deal with vector data. To apply them to graph pattern classification can be converted into vectors of attributes. Each and every attributes indicates the presence or absence of sub patterns. Attributes are created for every frequent sub patterns. The number of such sub patterns can be very large, feature selection process is used to select a subset of these frequent sub patterns maintaining approximately the same information [7].
Data Mining with Big data had drawn our attention on challenges with mining big data at three levels dealing with data, model, and system. At the Data Level as Big Data are stored at different locations in variety of forms and volumes which keeps on growing, an efficient computing platform will have to take distributed large-scale data storage into consideration for computing. The diversity of information sources and varied data collection environments, frequently result in data with missing or indecisive values. Further privacy concerns, unnecessary data, and error can be introduced into the data, to create altered data copies. Developing a safe and effective information sharing protocol is a major challenge. The Model level faces the challenge to create global models by combining already discovered patterns to form a unifying view. At the system level the challenge is to consider complex relationships between varieties of data sources, patterns, models. Linking unstructured data in complex relationship to form useful pattern is also a great challenge [3]

3. Applications of big data:

- Business: expands customer intelligence, improves operational efficiencies, customer personalization.
  To gain deep customer requirements one need strong personal connections and give customized services if possible which will drive more sales.
- Managing demands in the market
  By capturing external market and retailer data in real time to sense, evaluate, and answer to demand indicators faster than ever before.
- Fraud detection
  By analysing certain abnormal pattern from various data sources, fraud can be detected in financial transaction, health insurance etc.

These applications will allow people to have better services, better costumer experience

4. Challenges of Mining Big Data

Mining Big data has opened many new challenges and opportunities. Existing data mining techniques face great difficulties when they are required to handle the unprecedented heterogeneity, volume, speed, privacy, accuracy and trust coming along with big data and big data mining [5]. Significant challenges in leveraging the vast amount of data, including challenges in (1) System (2) Algorithm (3) Business model [4].
Key issues and challenges are heterogeneity, volume, speed, accuracy and trust, privacy crisis and garbage mining [5]

Variety and Heterogeneity: Different sources generate Big Data leading to great variety or heterogeneity of big data. The data from different sources possesses different types and representation forms, and is greatly interrelated and inconsistently represented. Mining such a dataset is great challenge and the degree of complexity is not imaginable before we sincerely get there. Heterogeneity in big data also means that it is an obligation to accept and deal with structured, semi-structured, and even entirely unstructured data concurrently. While structured data can be managed well with database systems, semi-structured data may partially managed, but unstructured data definitely will not. Both semi-structured and unstructured data are normally stored in files. The heterogeneity of big data means a new opportunity of unveiling, previously impossible, hidden patterns or knowledge dwelt at the intersections within heterogeneous big data.

Scalability: The extraordinary volume requires high scalability of its data management and mining tools. Following approaches if exploited properly, may lead to remarkable scalability required for future data and mining systems to manage and mine the big data: (1) cloud computing allows elasticity, which, combined with parallel computing architectures, bears the hope of realizing the needed scalability for dealing with the volume challenge of big data; (2) advanced user interaction support that facilitates prompt and effective system-user interface. Big data mining straightforwardly implies extremely time-consuming navigation in a massive search space, and prompt feedback from users must be beneficially exploited to help make early decisions, adjust mining strategies on the fly, and narrow down to smaller but promising subspaces.

Velocity/Speed: The capability of fast accessing and mining big data is highly essential – processing/mining of a task must be finished within a definite period of time, otherwise, the processing/mining results becomes less valuable or even worthless. Speed is also relevant to scalability – conquering or partially solving anyone helps the other one.

The speed of data mining depends on two major factors: data access time and the efficiency of the mining algorithms. Exploitation of advanced indexing schemes is fruitful to overcome speed concern. Multidimensional index structures are especially useful for big data. However design of new and more efficient indexing schemes is much desired, but remains one of the greatest challenges to the research community.

Controlling the probable parallelism in the access and mining algorithms can improve the speed of big data access and mining. The elasticity and parallelism support of cloud computing are the most promising facilities for boosting the performance and scalability of big data mining systems. The MapReduce framework allows users to define two tasks, map and reduce, to process data entries in parallel [6] The MapReduce parallel computing model is applicable to only limited class of data-intensive computing problems. Therefore, design of new and more efficient parallel computing models besides Map Reduce is needed.

Accuracy, Trust, and Provenance: With big data, the data sources are of many different origins, not all well-known, and not all confirmable. As a result, the accuracy and trust of the source data quickly become a serious concern, which also affects the mining results.

Privacy Crisis: Data privacy has been always an issue. The concern has become extremely serious with big data mining that often requires personal information in order to produce relevant/accurate results such as location-based and personalized services. Also, with the huge volume of big data such as social media that contains incredible amount of highly interrelated personal information, each bit of information can be mined out. Every transaction regarding our daily life is being pushed to online and leaves a trace there: we comminute with friends via email, instant message, blog, and Facebook; we do shopping and pay our bills online; credit card companies hold our confidential identity information; your payroll office has your personal information; everyone so far has the righteous sense of protecting your confidential personal information, but the possibility of unintended leaking cannot be ruled out once and forever, and no leaking today does not guarantee impermeable tomorrow. As time goes, every piece of your personal information will be scattered here or there. Everyone would easily gain the privilege of using powerful tools to extract your confidential information.

Evaluation and prevention of privacy violation during knowledge mining are two related concerns that call for serious investigation and novel solutions.

Garbage Mining: In the big data epoch, the amount of data produced and inhabited on the World Wide Web keeps increasing at an amazingly fast pace. In such an environment, data can quickly become obsolete, contaminated, and inadequate. In addition, there are data created as junks like junk emails. For having a relatively clean cyberspace and clean World Wide Web, attentions and research efforts are required. Cyberspace cleaning is not an easy task because of at least two foreseeable reasons: garbage is hidden, and there is an ownership issue.

Garbage mining is a serious research topic, different but related to big data mining – for the sake the sustainability of our digital environment, “mining for garbage” (and cleaning it) is as important as “mining for knowledge” (the canonical sense of data mining). This is especially so in the new era of big data. Garbage definition remains one of the greatest challenges

5. Conclusion

Big data is growing faster and will grow rapidly in coming years. It would be a great challenge to find useful information from ocean of data. Data mining algorithm will not be suitable as the challenges are new. The challenges are much higher than what have been discussed in this paper. Researchers teams need to work on each challenge but the challenge will be growing as Big Data will grow. But Big Data mining will help us to extract and use that information that was never ever done before.

References

[1] Jimmy Lin and Dmitriy Ryaboy, Scaling Big Data Mining Infrastructure: The Twitter Experience, SIGKDD Explorations Volume 14, Issue 2 (2012)

[2] Yizhou Sun and Jiawei Han, Mining Heterogeneous Information Networks: A Structural Analysis Approach, ACM SIGKDD Explorations Newsletter, Volume 14, Issue 2, pp. 20-28 (2012)

[3] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, Data Mining with Big Data, IEEE TRANSACTIONS

ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1 (2014)
[4]http://bigdata-mining.org/bigmine14/ JMLR Workshop and Conference Proceedings, KDD 2014
[5] Dunren Che, Mejdl Safran, and Zhiyong Peng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities (2013)
[6] Dean, J., Ghemawat, S.: MapReduce: a Flexible Data Processing Tool. In: Communication of the ACM, vol. 53, no. 1, pp.72-77 (2010).
[7] Albert Bifet, Mining Big Data in Real Time , Informatica 37 (2013) 15–20
[8] Wei Fan, Albert Bifet, Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations Volume 14, Issue 2 (2013)
[9] A. V. N. S. Jyothirmayee, Dr. G. Sreenivasula Reddy, K. Akbar, Understanding Big Data & DV^2 law, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 7 (2014)
[10]http://www.computerworld.com/article/2473730/smartphones/smartphones--big-data--storage-and-you.html Smart phones, Big Data, Storage and you (2013)
[11] http://albertbifet.com/big-data-mining/
[12] http://wikibon.org/blog/big-data-statistics/
[13] http://dataconomy.com/seven-vs-big-data/
[14] http://mbitm.uts.edu.au/feed/7-vs-big-data

P***rof. Kaushika Pal*** is MCA from VNSGU, Surat, Gujarat, India. Currently she is Assistant Professor at Sarvajanik College of Engineering and Technology, Surat. She has several research papers in National and International journals. Her research areas are Data Mining, Big data, Software Testing, Software engineering.

Dr. Jatinderkumar R. Saini is Ph.D. from VNSGU, Surat. He secured First Rank in all three years of MCA and has been awarded Gold Medals for this. Besides being University Topper, he is IBM Certified Database Associate (DB2), IBM Certified Associate Developer (RAD), BULATS Certified (C1), CPT Certified (C1) and LASSIB White Belt Certified. Associated with more than 50 countries, he has been the Member of Program Committee for more than 50 International Conferences (including those by IEEE) and Editorial Board Member or Reviewer for more than 30 International Journals (including many those with Thomson Reuters Impact Factor and also those published by Springer). He has more than 55 research paper publications and nearly 20 presentations in reputed International and National Conferences and Journals. He is member of ISTE, IETE, ISG and CSI. Currently he is working as Associate Professor and Director I/C at Narmada College of Computer Application, Bharuch, Gujarat, India. He is also Director (Information Technology) at Gujarat Technological University, Ahmedabad (GTU)'s A-B Innovation Sankul.