International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Mining the Big Data and its challenges

[1]**Mahesh S Nayak**
[1] Research and Development Centre
Bharathiar University,
Coimbatore – 641 046
mnayak67@yahoo.com
[2] **Dr. M. Hanumanthappa**
[2] Professor, Department of Computer Science & Applications,
Bangalore University, Bangalore.
hanu6572@hotmail.com
[3] **B R Prakash**
[3] Assistant Professor, Department of MCA,
Sri Siddhartha Institute of Technology, Tumkur.
brp.tmk@gmail.com

Abstract: *Big Data is a new term used to identify the datasets that are large in size and complexity. In this paper provide the overview of the process for data mining for big data. We address the current issues and challenges in big data mining process compared to the traditional data mining, the. Big Data Mining is one of the most excited research challenges in coming years. One of the key issues raised by data mining technology is not a business or technological one, but a social one. Other issues are that of data integrity, Analytics Architecture, Evaluation, Visualization and Distributed mining are deliberated.*

1. Introduction

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter, Weibo, etc., that allow users to create contents freely and amplify the already huge Web volume. Furthermore, with mobile phones becoming the sensory gateway to get real time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR (call data record)-based processing for billing purposes only.

It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, to buses, railway stations and airports) are all loosely connected. Trillions of such connected components will generate a huge data ocean, and valuable information must be discovered from the data to help improve quality of life and make our world a better place. For example, after we get up every morning, in order to optimize our commute time to work and complete the optimization before we arrive at office, the system needs to process information from traffic, weather, construction, police activities to our calendar schedules, and perform deep optimization under the tight time constraints. In all these applications, we are facing significant challenges in leveraging the vast amount of data, including challenges in (1) system capabilities (2) algorithmic design (3) business models.

Scalability is at the core of the expected new technologies to meet the challenges coming along with big data. The simultaneously emerging and fast maturing cloud computing technology delivers the most promising platforms to realize the needed scalability with demonstrated elasticity and parallelism capacities. Numerous notable attempts have been initiated to exploit massive parallel processing architectures [1].

Google's novel programming model, MapReduce [2], and its distributed file system, GFS (Google File System) [3], represent the early groundbreaking efforts made in this line. From the data mining perspective, mining big data has opened many new challenges and opportunities. Even though big data bears greater value (i.e., hidden knowledge and more valuable insights), it brings tremendous challenges to extract these hidden knowledge and insights from big data since the established process of knowledge discovering and data mining from conventional datasets was not designed to and will not work well with big data. The cons of current data mining techniques when applied to big data are centered on their inadequate Scalability and parallelism. In general, existing data mining techniques encounter 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. Improving existing techniques by applying massive parallel processing architectures and novel distributed storage systems, and designing innovative mining techniques based on new frame works/platforms with the potential to successfully overcome the aforementioned challenges will change and reshape the future of the data mining technology.

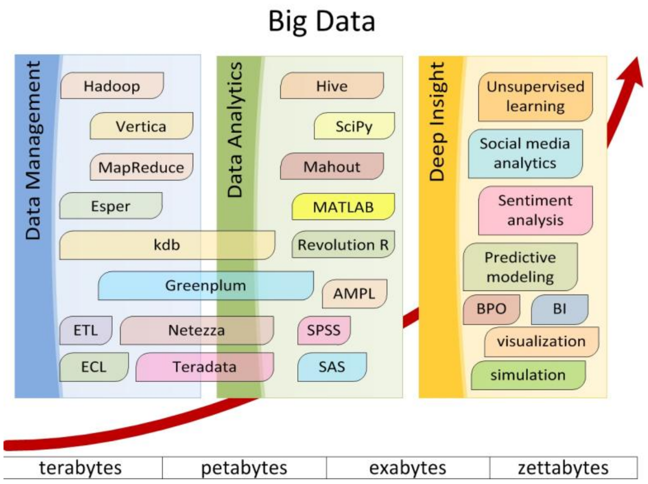

Figure 1: Big Data Management, Data Analytics, Deep insight

2. Data Mining

Knowledge discovery (KDD) is a process of unveiling hidden knowledge and insights from a large volume of data [4], which involves data mining as its core and the most challenging and interesting step (while other steps are also indispensable) . Typically, data mining uncovers interesting patterns and relationships hidden in a large volume of raw data, and the results tapped out may help make valuable predictions or future observations in the real world. Data mining has been used by a wide range of applications such as business, medicine, science and engineering. It has led to numerous beneficial services to many walks of real businesses – both the providers and ultimately the consumers of services. Applying existing data mining algorithms and techniques to real-world problems has been recently running into many challenges due to the inadequate scalability (and other limitations) of these algorithms and techniques that do not match the three Vs of the emerging big data. Not only the scale of data generated today is unprecedented, the produced data is often continuously generated in the form of streams that require being processed and mined in (nearly) real time. Delayed discovery of even highly valuable knowledge invalidates the usefulness of the discovered knowledge. Big data not only brings new challenges, but also brings opportunities – the interconnected big data with complex and heterogeneous contents bear new sources of knowledge and insights. Big data would become a useless monster if we don't have the right tools to harness its "wildness". Current data mining techniques and algorithms are not ready to meet the new challenges of big data. Mining big data demands highly scalable strategies and algorithms, more effective preprocessing steps such as data filtering and integration, advanced parallel computing environments (e.g., cloud Paas and IaaS), and intelligent and effective user interaction. Next we examine the concept and big data and related issues, including emerging challenges and the (foregoing and ongoing) attempts initiated on dealing with big data.

3. Big Data Mining

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" [5]. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [6] . However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [7]. The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [8] in his invited talk at the KDD BigMine'12 Work shop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, and Twitter are starting to look carefully to this data to find useful patterns to improve user experience. Alex ’Sandy’ Pentland in his ’Human Dynamics Laboratory’ at MIT, is doing research in finding patterns in mobile data about what users do, and not in what people says they do [9]. We need new algorithms, and new tools to deal with all of this data. Doug Laney [10] was the first one in talking about 3 V’s in Big Data management:

- **Volume:** there is more data than ever before; its size continues increasing, but not the percent of data that our tools can process

- **Variety:** there are many different types of data, as text, sensor data, audio, video, graph, and more

- **Velocity:** data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time

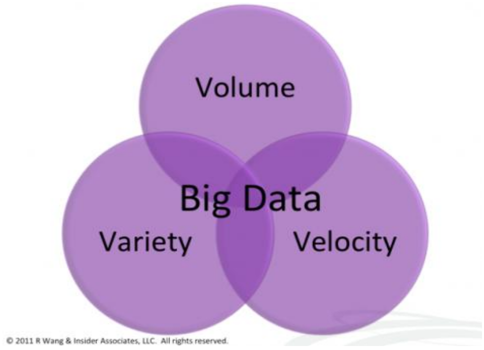
Volume
Big Data
Variety
Velocity
© 2011 R Wang & Insider Associates, LLC. All rights reserved.

Figure 2: The three Vs (volume, variety, and velocity)

Nowadays, there are two more V’s:

- Variability: there are changes in the structure of the data and how users want to interpret that data
- Value: business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

Gartner[11] summarizes this in their definition of Big Data in 2012 as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

There are many applications of Big Data, for example the following [12]:

- **Business:** costumer personalization, churn detection
- **Technology:** reducing process time from hours to seconds
- **Health:** mining DNA of each person, to discover, monitor and improve health aspects of every one
- **Smart cities:** cities focused on sustainable economic development and high quality of life, with wise management of natural resources.

These applications will allow people to have better services, better costumer experiences, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before [13].

4. Important challenges in Big Data Mining

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal with in the years to come:

Analytics Architecture. It is not clear yet how an optimal architecture of an analytics system should be constructed to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system as Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general and extensible, allows ad hoc queries, minimal maintenance,debug gable.

Evaluation: It is important to achieve significant statistical results, and not be fooled by randomness. As Efron explains in his book about Large Scale Inference, it is easy to go wrong with huge data sets and thousands of questions to answer at once. Also, it will be important to avoid the trap of a focus on error or speed as Kiri Wagstaff discusses in her paper “Machine Learning that Matters”.

Distributed mining: Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

Time evolving data: Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first. For example, the data stream mining field has very powerful techniques for this task.

Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where we don’t lose anything, or sampling where we choose data that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are losing information, but the gains in space may be in orders of magnitude. Use core sets to reduce the complexity of Big Data problems. Core sets are small sets that provably approximate the original data for a given problem. Using merge-reduce the small sets can then be used for solving hard machine learning problems in parallel.

Visualization: A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to find user-friendly visualizations. New techniques and frameworks to tell and show stories will be needed, as for example the photographs, info graphics and essays in the beautiful book"The Human Face of Big Data".

Hidden Big Data: Large quantities of useful data are getting lost since new data is largely untagged file-based and unstructured data. The 2012 IDC study on Big Data explains that in 2012, 23% (643 Exabyte) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

5. Conclusion

The big data movement has energized the data mining, knowledge discovery in data bases and associated software development communities, and it has introduced complex, interesting questions for researchers and practitioners. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area, still in its infancy. we have discussed the challenges that in our opinion, mining evolving data streams will have to deal during the next years. We have outlined new areas for research. These include structured classification and associated application areas as social networks. Our ability to handle many Exabyte 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 is going to continue growing.

REFERENCES

1. Berkovich, S., Liao, D.: On Clusterization of big data Streams. In: 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York (2012)
2. Beyer, M.A., Laney, D.: The Importance of 'Big Data': A Definition. Gartner (2012)
3. Madden, S.: From Databases to big data. IEEE Internet Computing 16(3), 4–6 (2012)
4. Shmueli, G., Patel, N.R., Bruce, P.C.: Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, 2nd edn. Wiley & Sons, Hoboken (2010)
5. Ghoting, A., Kambadur, P., Pednault, E., Kannan, R.: NIMBLE: a Toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on MapReduce. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, pp. 334–342 (2011).
6. B R Prakash, Dr. M. Hanumanthappa: Issues and Challenges in the Era of Big Data Mining International Journal of Emerging Trends & Technology in Computer Science Volume 3, Issue 4, July - August 2014 ISSN 2278-6856.
7. Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., Hellerstein, J.M.: Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. VLDB Endowment 5(8), 71–727 (2012)
8. Borkar, V.R., Carey, M.J., Li, C.: big data Platforms: What's Next? ACM Crossroads 19(1), 44–49 (2012)
9. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four Degrees of Separation. CoRR, abs/1111.4570, 2011.
10. 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.
11. B. Efron. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.
12. U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http:// big-data-mining.org/keynotes/#fayyad, 2012.
13. D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In SODA, 2013.
14. Agrawal, D., Bernstein, P., Bertino, E., et al.: Challenges and Opportunities With big data A Community White Paper Developed by Leading Researchers Across the United States (2012), http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf.