Data Mining Revolution On High Dimensional Data

G SIREESHA[1], CH.RAVINDRA REDDY[2], J.V Krishna[3]

[1]M-Tech Dept. of CSE SreeVahini Institute of Science and Technology TiruvuruAndhra Pradesh

[2, 3] Assist. Professor Dept. of CSE SreeVahini Institute of Science and Technology Tiruvuru Andhra Pradesh.

Abstract:

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Keywords: - Dataset, Clustering, Big data.

1. INTRODUCTION

Data mining involves exploring and analyzing large amounts of data to find patterns for big data. Data volumes grow exponentially. Its growth is caused by the increasing number of system and people acting as data sources of textual, verbal, video and transactional information. This data contains insider information and patterns previously hidden due to lack of proper technologies. Generally, the goal of the data mining is either classification or prediction. In classification, the idea is to sort data into groups. For example, a marketer might be interested in the characteristics of those who responded versus who didn't respond to a promotion. Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources, This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. Our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) Autonomous with distributed and decentralized control,

and 3) complex and evolving in data and knowledge associations.

Our Contributions

The main contributions of this paper are the following:

1. In this dataset, we store all the information or records about any student database or hospitals etc.
2. We can update this data whenever it is changing i.e., like in student database if when a student got admitted newly or their details
3. Clustering is one of the major techniques used for data mining in which mining is performed by finding out clusters having same group of data.
4. This is used for storing of homogenous and heterogeneous data and mining of data using specified category.
5. As we know that Big data is the collection of structured(tables) , semi-structured(xml files) and unstructured data(pdf,images)
6. Here we use horizontal scalability for sharing of data.

2. RELATED WORK

On the level of mining platform sector, at present, parallel programming models like Map Reduce are being used for the purpose of analysis and mining of data. Map Reduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing, with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. In case of design of data mining algorithms, Knowledge evolution is a common phenomenon in real world systems. But as the problem statement differs, accordingly the knowledge will differ for example, when we go to the doctor for the treatment, that doctor's treatment program continuously adjusts with the conditions of the patient. Similarly the knowledge for this, we proposed and established this theory, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find.

3. PROPOSED METHOD

A HACE theorem to model Big Data characteristics. The characteristics of HACE make it an extreme challenge for discovering useful knowledge from the Big Data. The HACE theorem suggests that the key characteristics of the Big Data are

1. Huge with Heterogeneous and diverse data sources.

2. Autonomous with distributed and decentralized control.

3. Complex and Evolving in data and knowledge associations.

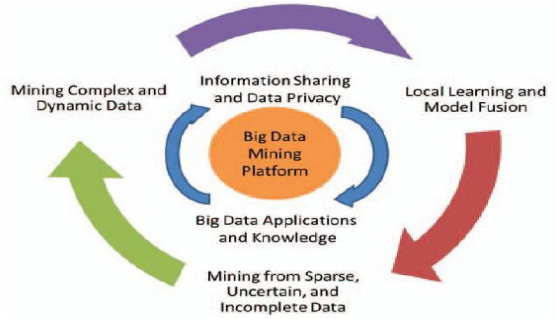

Figure 1: Block Diagram for Proposed System.

4. IMPLEMENTATION

Integrating and mining bio data

We have integrated and mined bio data from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission, and processing for information networks. We have developed methods for semantic-based data integration, automated mining with wildcards, and application problems as follows:

hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

Big Data Fast Response

We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making.

1. Designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing
2. Building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data, as well as accurately predict the trend of the data in the future; and
3. A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

Pattern matching and mining

We perform a systematic investigation on pattern matching, and mining as below points.

1. Exploration of the NP-hard complexity of the matching and mining problems.
2. Multiple patterns matching with wild cards.
3. Approximate pattern matching and mining,
4. Application of our research onto ubiquitous personalized information processing and bio information.

5. CONCLUSION

Big data is the term for a collection of complex datasets, Data mining is an analytic process designed to explore data(usually large amount of data typically business or market related also known as "big data") in, search of consistent patterns and then to validate the findings by applying the detected pattern to new subsets of data.

To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of Big Data.

REFERENCES

[[1] D. Akhawe, P. Saxena, and D. Song. Privilege separation in HTML5 applications. In Proceedings of the 21st Usenix Security Symposium, Bellevue, WA,
Aug. 2012.

[2] A. Arasu, S. Blanas, K. Eguro, R. Kaushik, D. Kossmann, R. Ramamurthy, and R. Venkatesan. Orthogonal security with Cipherbase. In Proceedings of the 6th Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, Jan. 2013.

[3] G. Ateniese, K. Fu, M. Green, and S. Hohenberger. Improved proxy re-encryption schemes with applications to secure distributed storage. In Proceedings of the 13th Annual Network and Distributed System Security Symposium, San Diego, CA, Feb. 2006.

[4] S. Bajaj and R. Sion. TrustedDB: a trusted hardware based database with privacy and data confidentiality. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pages 205–216, Athens, Greece, June 2011.

[5] A. Barth, C. Jackson, and J. C. Mitchell. Securing frame communication in browsers. In Proceedings of the 17th Usenix Security Symposium, San Jose, CA, July–Aug. 2008.

[6] A. Barth, J. Caballero, and D. Song. Secure content sniffing for web browsers, or how to stop papers from reviewing themselves. In Proceedings of the 30th IEEE Symposium on Security and Privacy, Oakland, CA, May 2009.

[7] F. Beato, M. Kohlweiss, and K. Wouters. Scramble! your social network data. In Proceedings of the 11th Privacy Enhancing Technologies Symposium, Waterloo, Canada, July 2011.

[8] D. Benjamin. Adapting Kerberos for a browserbased environment. Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Sept. 2013.

Authors Profiles

G SIREESHA M-Tech Dept. of CSE
SreeVahini Institute of Science and Technology. Tiruvuru. Andhra Pradesh.

Ch.Ravindra Reddy
Asist.Professor Sree Vahini Institute of Science and Technology Tiruvuru Andhra Pradesh
"M.TECH(CSE),MBA,MISTE"
Email id :-ravindrareddy.ch94@gmail.com

J.V Krishna
Assoc.Professor
SreeVahini Institute of Science and Technology Tiruvuru Andhra Pradesh.