Research Article | Impact Factor: 0.621 | ISSN: 2319-507X
Sharayu Sangekar, IJPRET, 2014; Volume 2 (9): 793-799 | IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

DATA MINING OF COMPLEX DATA WITH MULTIPLE, AUTONOMOUS SOURCES

SHARAYU S. SANGEKAR[1], PRANJALI P. DESHMUKH[2]

1. M. E. Scholar, department of CSE, P. R. Pote (Patil) College of Engineering, Amravati, India.
2. HOD, P. R. Pote (Patil) College of Engineering, Amravati, India.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: Data mining is an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining soft- ware tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it.. We present in this paper, a broad overview of the topic, its current status, related work, and forecast to the future.

Keywords: Big Data, data mining, decentralized control, autonomous sources, HACE theorem.

PAPER-QR CODE

Corresponding Author: MS. SHARAYU S. SANGEKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Sharayu Sangekar, IJPRET, 2014; Volume 2 (9): 793-799

INTRODUCTION

Data Mining is the process of discovering interesting knowledge, such as Patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Due to the wide availability of huge amounts of data in electronic forms, and the imminent need for turning such data into useful information and knowledge for broad applications including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years . Researchers view data mining as an essential step of knowledge discovery process consists of an iterative sequence of the following steps such as data cleaning, data integration, data selection, data transformation, pattern evaluation. Data Mining is the extraction, predictive information from large database is knowledge of discovery using sophisticated blend of technique from a traditional statistics, artificial intelligence and computer graphics. To best apply the advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools.

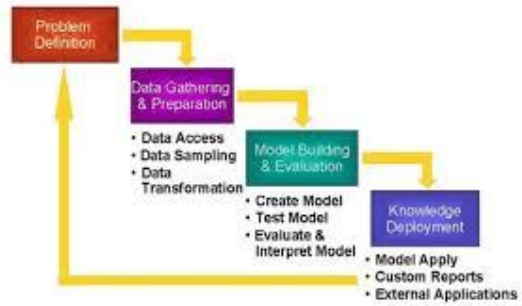

Fig.1: Data mining process

Many data mining tools currently operate outside the warehouse requiring extra steps for extracting, importing and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Fig.2: Big data

The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions".

2. ARCHITECTURE OF DATA MINING AND RELATED WORK:

Data mining techniques are the result of long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time.

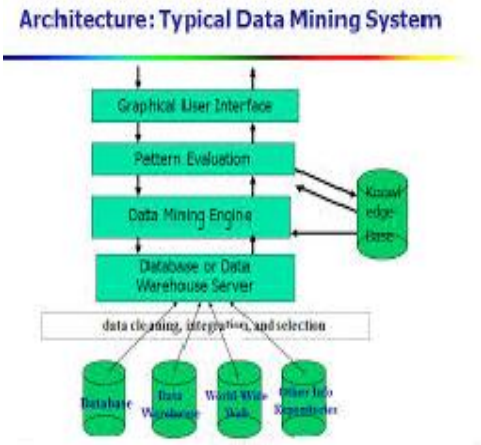

Fig.3: Data Minning architecture

Mining Big Data in Real Time Nowadays, quantity of data that is created every two days is estimated to be 5 ex- a bytes. This amount of data is similar to the amount of data created from the dawn of time up until 2003. Moreover, it was estimated that 2007 was the first year in

which it was not possible to store all the data that we are producing. This massive amount of data opens new challenging discovery tasks. Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring and traffic management, log records or click-streams in web exploring, manufacturing processes, call detail records, email, blogging, twitter posts and others .

Anomaly Detection in Network using Data mining Techniques

As the network dramatically extended security considered as major issue in networks. There are many methods to increase the network security at the moment such as encryption, VPN, firewall etc. but all of these are too static to give an effective protection against attack and counter attack. We use data mining algorithm and apply it to the anomaly detection problem. In this work our aim to use data mining techniques including classification tree and support vector machines for anomaly detection. The result of experiments shows that the algorithm C4.5 has greater capability than SVM in detecting network anomaly and false alarm rate by using 1999 KDD cup data.

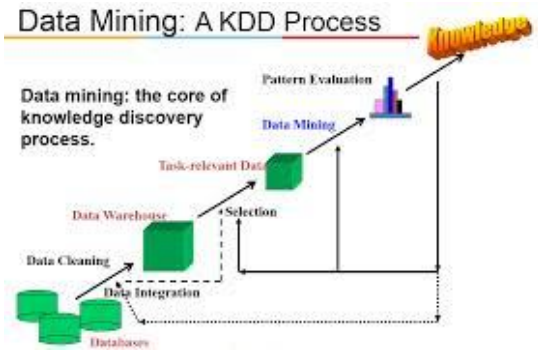

Fig.4: Data Mining with KDD process

3.ALGORITHM:

BIG DATA CHARTERISTICS: HACE THEOREM- HACE Theorem Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naı ¨ve sense, we can imagine that a number of blind men are trying to size up a giant elephant ,which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant according to the part of information he collects during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each

conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that 1) the elephant is growing rapidly and its pose changes constantly, and 2) each blind man may have his own (possible unreliable and inaccu- rate) information sources that tell him about biased knowledge about the elephant. Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

4. RESULT AND DISCUSSION

Autonomous Sources with Distributed and Decentralized Control -Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data-related applica- tions, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/ regions. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors.

Huge Data with Heterogeneous and Diverse Dimensionality - One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. For example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry

detailed examinations. For a DNA or genomic-related test, micro- array expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation.

5.CONCLUSION:

Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, 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. A number of conclusions can be drawn after our preliminary experiments as follows. Data must be collected for a longer time period, allowing thus generation of cleaner and more precise datasets.. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors.

6.REFERENCES:

1. Bartok J., Habala O., Bednar P., Gazak M. & Hluchy L. (2010). Data mining and integration for predicting significant meteorogical phenomena. International Conference on Computational Science, (ICCS 2010), Procedia Computer Science 1, Elsevier, 37-46

2. A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter stream- ing data. In Proc 13th International Conference on Discovery Science, Can- berra, Australia, pages 1–15. Springer, 2010.

3. Liu, L., Kantarcioglu, M., Thuraisingham, B.M.: A Novel Privacy Preserving Decision Tree. In: Proceedings Hawaii International Conf. on Systems Sciences (2009)

4. Wang, Q., Meegan, J., Freund, T., Li, F.T., Cosgrove, M.: Smarter City: The Event Driven Realization of City-Wide Collaboration. 2010 International Conference on Management of e-Commerce and e- Government. 195-199 (2010).

5. R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

6. M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

7. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

8. A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.