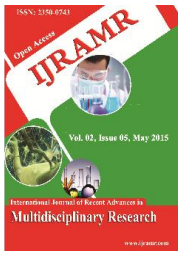
International Journal of Recent Advances in Multidisciplinary Research
Vol. 02, Issue 05, pp.0413-0417, May, 2015



Research Article

DATA MINING USING HACE THEOREM

***Abdul Ahad, Ravali, M. and Durga Bhavani, S.**

Department of ECM, KL University, Vaddeswaram, India

ARTICLE INFO

Article History:
Received 19th February 2015
Received in revised form
21th March, 2015
Accepted 25th April, 2015
Published online 31th May, 2015

Keywords:
Big Data,
Data mining,
Hace theorem.

ABSTRACT

Big Data consists of huge modules, difficult, growing data sets with numerous and, independent sources. With the fast development of networking, storage of data, and the capacity of data gathering, Big Data is now rapidly increasing in all science and engineering domains, as well as animal, genetic and biomedical sciences. This paper detailed a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining view. This data-oriented model contains demand-driven aggregation of information sources, mining and study, user data modeling, and security and privacy problems .We examine the difficult issues in the data-oriented model and also in the Big Data revolution.

INTRODUCTION

Every day 3 billion kilobytes of data are produced and today 90 percent of the data in the web were created within the last two years. Our ability for information creating has ne'er been therefore dominant and large since the creation of the knowledge technology within the early 19th century. In an instance, a public picture distribution site, flicker, this achieved 2.5 million photos per day. Each photo is assumed the size of 2 megabyte, this needs 5 terabytes storage every single day and as an old axiom elaborates that a single picture has value of lacks of words. The pictures on Flicker square measure an enormous tank for U.S.A. To look the human civilization, social proceedings, public relationships, disasters, and so on, on condition that we've the facility to connect the huge quantity of knowledge. These instances shows the rise of BIG DATA applications where data gathering has grown extremely and is beyond the capability of usually used software tools to catch, control, and make the procedure. The most essential challenge for BIG DATA applications is to discover the huge volume of data and mine useful information for future events. In many occurrences, the information mining process has to be very capable and close to real time because storing all practical data is nearly in flexible. For a instance, the square kilometer arraycapability of usually used software tools to catch, control, and make the procedure.

Corresponding author: Abdul Ahad
Department of ECM, KL University, Vaddeswaram, India

The most essential challenge for BIG DATA applications is to discover the huge volume of data and mine useful information for future events. In many occurrences, the information mining process has to be very capable and close to real time because storing all practical data is nearly in flexible. For a instance, the square kilometer array

However, with a 50 gigabytes (GB) second data volume, the data delivers from the Square Kilometer array are especially large. Although scientist have confirmed that attractive patterns, such as temporary radio anomalies can be exposed from the Square Kilometer array data, existing processes can boosts in an offline manner and are incapable of handling this

However, with a 50 gigabytes (GB) second data volume, the data delivers from the Square Kilometer array are especially large. Although scientist have confirmed that attractive patterns, such as temporary radio anomalies can be exposed from the Square Kilometer array data, existing processes can boosts in an offline manner and are incapable of handling this

Big Data

Big Data is a comprehensive term for any collection of data sets so large and multifarious that it becomes difficult to process them using conventional data processing applications. The challenges embrace analysis, capture, search, sharing, storage, transfer, revelation, and privacy violations. The tendency to giant information sets is as a result of the extra

data derived from analysis of one large set of connected information, as compared to separate smaller sets with identical total quantity of knowledge, allowing correlations to be found to combat crime and then on. Here there are 2 types of Big Data: structured and unstructured.

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured knowledge additionally embraces things like sales res, account balances, and dealings knowledge.

Unstructured data include more multifarious information, such as customer reviews from feasible websites, photos and other multimedia, and comments on social networking sites. These data cannot be separated into categorized or analyzed numerically.

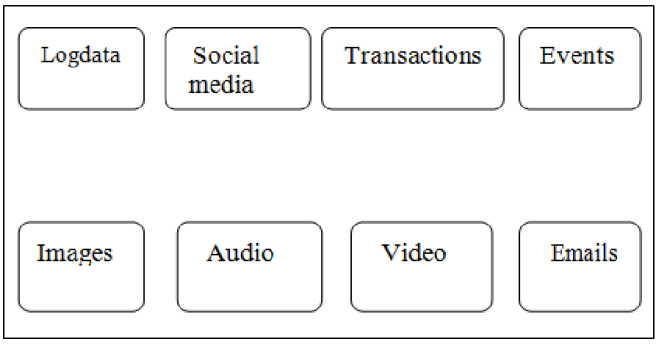


Fig.1. Sources of BIG DATA

Big Data Characteristic (HACE Theorem)

HACE theorem is theorem to model the BIG DATA characteristics. Big Data starts with large-volume, Heterogeneous; Autonomous sources with distributed and decentralized control, and seeks to explore Complex and Evolving relationships among data

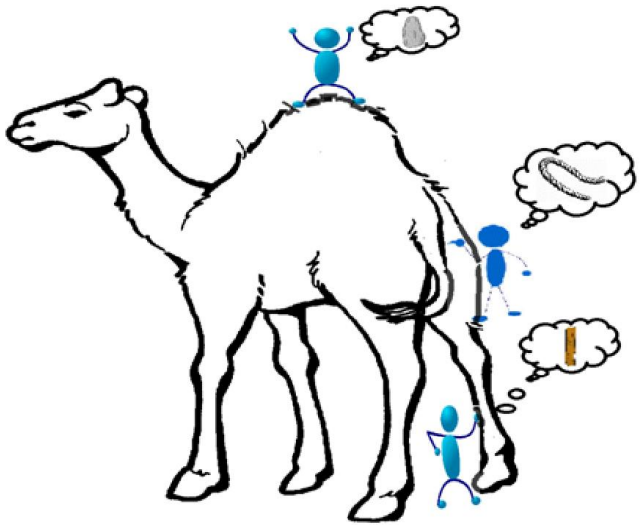


Fig. 2. The blind men and the enormous camel: the restricted view of each blind man leads to a biased conclusion

These characteristics make it an intense challenge for discovering useful knowledge from the Big Data. In a native sense, we can imagine that a number of blind men are trying to size up a giant camel (see Fig. 2), which will be the Big Data in this context. The aim of each blind man is to extract conclusion of the camel according to the part of information he collects during the procedure. Because each individual's opinion is restricted to his native area, it is expected that the blind men will each conclude independently that the camel "feels" like a rope, a stone, a stick, depending on the part each of them is limited to. To make the problem even more complex, let us accept that 1) the camel is increasing quickly and its posture varies frequently, and 2) each blind man may have his own information sources that tell him about subjective knowledge about the camel (e.g., one blind man may exchange his feeling about the camel with another blind man, where the exchanged knowledge is intrinsically subjected). Exploring the Big Data in this scenario is equivalent to form various information from different sources (blind men) to help to draw a best possible illustration to uncover the actual sign of the camel in a actual way. Certainly, this job is not as simple as enquiring each blind man to designate his spirits about the elephant and then getting an skilled to draw one single picture with a joint opinion, regarding that each separate may express a different language (varied and diverse information sources) and they may even have confidentiality concerns about the messages they measured in the information exchange procedure. The term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are

A. Huge with various and miscellaneous data sources: One of the fundamental characteristics of the Big Data is the large volume of data represented by various and miscellaneous dimensionalities. This large volume of data comes from various sites like Twitter, MySpace, Orkut and Linked In etc. This is as a result of completely different data collectors like their own illustration or procedure for knowledge recording, and therefore the nature varied applications also ends up in various knowledge representations

B. Autonomous Sources with circulated & disperse Control: Autonomous Sources with circulated & disperse Control are a main characteristic of Big Data applications. Being an autonomous, each data source is able to produce and collect information without connecting 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 depending on other servers. On the other hand, the massive volumes of the data also make an application susceptible to attacks or failure, if the whole system has to depend on any centralized control unit. For example, Asian markets of Wal-Mart are inherently completely different from its North yank markets in terms of seasonal promotions, high sell things, and client behaviors. a lot of specifically, the regime laws conjointly impact on the wholesale management method and end in restructured knowledge representations and knowledge

C. Complex and Evolving associations: In associate early stage of {information} centralized information systems, the main target is on finding best feature values to represent every observation. This type of sample feature representation inherently treats each individual as an independent entity

without considering of their social connections, which is the most important factors of the human society. The correlations between individuals inherently complicate the whole data representation and any reasoning process on data. In this dynamic world, the features are used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Examples of complex data types are bills of the materials, processing of word documents, maps, the time-series, images and video. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values.

The 5V'S of Big Data

In the past, the term "Big Data" has served as a catch all phrase for the massive amounts of knowledge on the market and picked up within the digital world. Today, massive information is being known as the rising power of the twenty first century and helps the maximum amount over nonsensicality, acting as an enormous d High Performance. Inter-Thread electronic messaging Library within the technology channel. With a rate of fifty a year, harnessing all of the elements of huge information presents a true challenge. (See Figure 3) It shows five V's in massive information. In a 2001 publication of metaGroup, Gartner analyst Doug Laney introduced the three V's of information management, shaping the three main elements of information as volume, speed and selection. Variety - Refers to the various types of information that we have a tendency to collect and use. Information comes in numerous formats, like structured and unstructured. to form matters even tougher, due to the explosion of knowledge generated by social media sources, eighty to eighty five p.c of the complete world's information is currently unstructured (text, audio, video, click streams, log files then on).

As the years have continuing and also the quantity of knowledge created considerably will increase, we have a tendency to currently grasp way more regarding what defines huge information, and IBM has introduced a fourth V, Veracity, as made public in their info graphic. Velocity - Velocity refers to the speed at that new information is generated and also the speed at that it moves around. as an example, The big apple stock market captures regarding 1 TB of trade info daily. Reacting quick enough and analyzing the streaming information is distressful to businesses, with speeds and peak periods typically inconsistent.

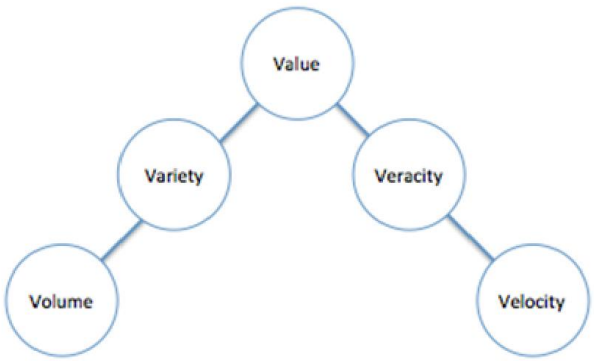


Fig.3. Five Vs of BIG DATA

Value-Through effective data processing and analytics, the high quantity of knowledge that we have a tendency to collect throughout the traditional course of doing business may be place to smart use and yield price and business opportunities. By applying processing and analytics to reveal valuable business knowledge embedded in structured, unstructured, and streaming information and information warehouses, this insight could also be accustomed facilitate the revamp offer chains, improve a program planning, track sales and promoting activities, live performance across channels, rework into associate on-demand business. The enormous knowledge strategy provides businesses the aptitude to higher analyze this knowledge with a goal of fast profitable growth.

Veracity- The average billion dollar company is losing $130 million a year owing to poor knowledge management. truthfulness refers to the uncertainty close knowledge, that is owing to knowledge inconsistency and wholeness that ends up in another challenge, keeping huge knowledge organized. The volume, velocity, selection and truthfulness of information that's being generated nowadays goes on the far side what ancient analytics systems will handle during a timely and economical manner. This ends up in the fifth V that organizations are scuffling with, finding the worth inside their knowledge. Velocity- Velocity refers to the speed at that new information is generated and also the speed at that it moves around. as an example, The big apple stock market captures regarding 1 TB of trade info daily. Reacting quick enough and analyzing the streaming information is distressful to businesses, with speeds and peak periods typically inconsistent.

Challenges with Big Data

For an intelligent knowledge database system (Wild list Organization, 2000) to handle massive information, the essential key is to scale up to the extremely huge volume of data and provide actions for the characteristics featured by the HACE theorem. Figure. four shows a abstract read of the massive processing framework, which has 3 tiers from within out with concerns on information accessing and computing (Tier I), information isolation and domain information (Tier II), and large data processing algorithms (Tier III).

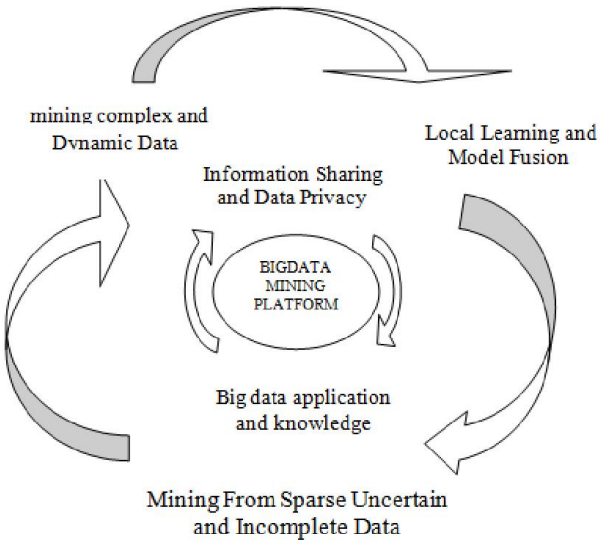


Fig.4. A conceptual view of the Big Data Processing framework

Tier I: Big Data Mining Platform

In typical data processing systems, the mining procedures need machine thorough computing units for information analysis and comparisons. for giant data processing, as a result of quantity of knowledge is huge so one laptop computer (PC) cannot handle, a typical massive processing framework can rely upon cluster computers with a superior computing platform, with an information mining task being dead by running some parallel Computing tools, like Map Reduce or Enterprise management Language (ECL), on an oversized range of clusters. The operate of the computer code module is to create positive that one data processing task, like finding the most effective match of a query from a info with billions of records, is split into several tiny tasks every of that is running on one or multiple cluster.

Tier II: data isolation and domain knowledge

In Big Data, Semantic & Application knowledge refer to several aspect related to the rules, policies, user information & application information. The most important aspect in this tier contain 1) Information sharing and its confidentiality; and 2) domain and application knowledge.

V 2.1: Information Sharing and its confidentiality

Information sharing is an crucial goal for all systems relating multiple parties (Lemke *et al*., 2003). Whereas the Goal of sharing is obvious, a real-world concern is that huge information applications square measure associated with sensitive data, like banking transactions and medical records. Straightforward information interactions don't resolve privacy considerations (Mangiameli *et al*., 2004; Baylis and Philip, 1999 and Bhavani Thuraisingham, 2008), however public revelation of AN individual's personal locations/movements over time will have serious repercussion for privacy. To safeguard privacy, 2 common approaches square measure to

1) limit access to the info, like adding certification or access management to the info entries, therefore sensitive data is accessible by a restricted cluster of users only and

2) Remove information fields such sensitive data cannot be pinpointed to a personal record (Walus *et al*., 1997).

Domain and Application Knowledge

Domain and application information (Jenn-Lung Su *et al*., 2001) provides necessary info for coming up with huge data processing algorithms and systems. in a large straight forward case, Application information will facilitate to spot right options for modeling the essential knowledge. The domain and application information may facilitate style possible business objectives by exploitation huge knowledge analytical techniques.

Local knowledge and Model synthesis for Multiple Information Sources. As Big Data applications are featured with independent sources and decentralized controls, collecting all distributed data sources to a centralized site for mining is thoroughly excessive due to the possible transmission cost and privacy concerns. More specifically, the global mining can be featured with a two-step process, at data, model, and at the knowledge stages. At each data level, local site can calculate the data statistics locate on the local data sources and exchange the statistics between the sites to achieve a global data distribution view. At the pattern level, every website will perform the native mining activities, with reference to the localized knowledge, to find native patterns. By exchanging patterns between multiple sources, new global patterns can be synthesized by aggregating patterns across all sites (Hosseinkhah *et al*., 2009). At the knowledge level, model correlation analysis finds out the importance between models generated from different data sources to determine how relevant the data sources are connected with each other, and how to form the accurate decisions based on the models built from autonomous sources

Mining from meager, tentative, and partial Data

Meager, tentative, and partial knowledge area unit process options for large knowledge applications. Being meager, the quantity of knowledge points is just too few for etymologizing consistent conclusions. Tentative knowledge area unit a special variety of knowledge reality wherever every knowledge set isn't any longer settled however is subject to some casual/inaccurate distributions. The absent valuesis caused by completely different realities, like the failure of a device node, or some regular policies to by design skip some values. Where as latest data processing algorithms have in-built solutions to handle absent values, knowledge attribution is a longtime analysis field that seeks to attribute absent values to supply increased models

Mining Complex and Dynamic

Data The growth of massive information is driven by the quick growing of advanced information and their changes in volumes and in nature (Mangiameli *et al*., 2004). Documents denote on WWW servers, net backbones, social networks, communication networks, and transportation networks, and then on square measure all featured with advanced information. Whereas advanced dependency structure supported the info elevate the issue for our data systems, however, huge information quality is conferred in several aspects, as well as advanced various information varieties, advanced essential linguistics relations in information, and sophisticated association networks among information. In huge information, information varieties embody structured information, unstructured information, and semi structured information, and so on. Significantly, there square measure relative databases, text, hyper-text, image, audio etc.

Conclusion

While the term massive knowledge specifically associated with knowledge volumes, our HACE theorem applies the key characteristics of the large knowledge are 1) large with varied and various knowledge sources, 2) freelance with scattered and redistributed management, and 3) difficult and developing in knowledge and information associations. To take care of massive data processing, superior computing platforms are necessary, that enforce organized styles to line free the total power of the large knowledge. At the information level, the

freelance info sources and also the vary of the information assortment environments; usually end in knowledge with advanced conditions, like unsure values. In various things, isolation problems, noise, and errors area unit typically introduced into the data, to construct distorted data copies. Mounting a secure and sound info sharing procedure may be a main challenge. At the model level, the clue challenge is to provide world models by change of integrity domestically searched patterns to create a uniform read. At the system level, the required challenge is that an enormous data processing framework wishes to suppose troublesome interaction between samples, models, and, at the side of their development changes with time and alternative potential factors. A system requests to be rigorously designed so formless knowledge are often joined through their troublesome relationships to form helpful patterns, and also the growth of knowledge volumes and item relationships ought to facilitate type legal patterns to guess the trend and future.

REFERENCES

Abbasi, M. M. and S. Kashiyarndi, 2006. "Clinical Decision Support Systems: A discussion on different methodologies used in Health Care."

Baylis, Philip, 1999. "Better health care with data mining." SPSS White Paper, UK.

Bhavani Thuraisingham, 2008. Data Mining for Security Applications, IEEE/IFIP international Conference on Embedded and Ubiquitous Computing.

Christodorescu, M. and S. Jha, Testing Malware Detectors, International Symposium on Software Testing and Analysis archive. Boston, Massachusetts, USA.

Hosseinkhah, Fatemeh, *et al*. 2009. "Challenges in Data Mining on Medical Databases." 1393-1404.

Jenn-Lung Su, Guo-Zhen Wu and I-Pin Chao, 2001. The Approach of Data Mining Methods for Medical Database. IEEE. P1-3.

Johannes Kinder, "Detecting Malicious Code by ModelChecking"pure.rhul.ac.uk/portal/files/17566588/mcode dimva05.pdf.

Koh, Hian Chye and Gerald Tan, 2011. "Data mining applications in healthcare."*Journal of Healthcare Information Management*— Vol 19.2: 65.

Lemke, Frank and Johann-Adolf Mueller, 2003. "Medical data analysis using self-organizing data mining technologies." Systems Analysis Modeling Simulation43.10: 1399-1408.

Mangiameli, Paul, David West and Rohit Rampal, 2004. "Model selection for medical diagnosis decision support systems." Decision Support Systems 36.3: 247-259.

Miller, Randolph, A. 1994. "Medical Diagnostic Decision Support Systems—Past, Present, and Future A Threaded Bibliography and Brief Commentary." *Journal of the American Medical Informatics Association* 1.1: 8-27.

Ozer, Patrick, 2008. "Data Mining Algorithms for Classification."

Walus, Y. E., H. W. Ittmann and L. Hammer, 1997. "Decision support systems in health care." Methods of information in medicine 36.2: 82.

Wild list Organization. Virus descriptions of viruses in the wild. Online publication, 2000. http://www.fsecure. com/virus-info/wild.html.

****