International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

Survey on Data Mining with Big Data

Rakhshanda Noor & Prof. Rakesh V.S
[1]8th semester Student, Cambridge Institute of Technology, Bangalore, Karnataka, India.
E-mail:rakhshanda2396@gmail.com
[2]Professor, Cambridge Institute of Technology, Bangalore, Karnataka, India.
E-mail:rakeshv.cse@citech.edu.in

Abstract:

Big data consists of heterogeneous information from multiple sources. Analyzing this huge amount of data and extracting useful information from it is termed as Data Mining. Nowadays it is being extensively used in science and technology to extract the vast amount of data. Data should be processed to extract some useful knowledge from it. This paper presents the 3 V's architecture of Big data and also the lifecycle of Data mining. Further we analyze the challenges with data mining and some proposed solution.

Keywords

Big Data, data mining, 3 V's architecture.

1. Introduction

Amazon handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. Facebook handles about 80 billion photos from its user base. Google handles roughly 2 trillion searches per year.

The above examples shows the boom of Big Data applications where data collection has grown exponentially and is beyond the ability of commonly used software tools to extract, manage, and process within a tolerable elapsed time.

Data sets grow rapidly because they are increasingly gathered by numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.

2. Big Data

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. The term "big data" is being used to describe an increasing range of technologies and techniques. In essence, big data is data that is valuable but, traditionally, it was not practical to store or analyze it due to limitations of cost or the absence of suitable mechanisms. Big data typically refers to collections of datasets that, due to size and complexity, are difficult to store, query, and manage using existing data management tools or data processing applications.

3. Big Data Architecture

The 3 V's of Big data:

Volume: The quantity of data that is generated and stored is huge. The size of the data determines the value and potential insight and whether it can be considered big data or not.

Variety: The nature or structure of the data varies. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

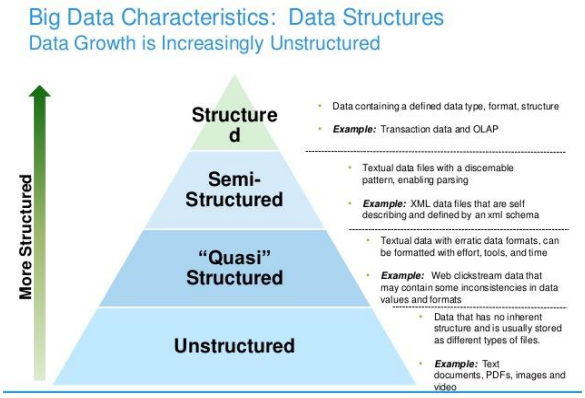
Velocity: In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time.

Huge volume of data is represented by heterogeneous and dissimilar dimensionalities. This is because different information sources prefer their own protocols for data recording. Also the nature of different applications results in variety of data representations. For example, the heterogeneous features like Aadhaar card and PAN card refer to the different types of representations for the same individuals and have different functionalities. Thus, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

A main characteristic of Big Data applications are the Autonomous data sources with distributed and decentralized controls. Each data source is able to generate and collect information without involving (or relying on) any centralized control. For example, 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. Whereas, the enormous volumes of the data also make an application susceptible to attacks or malfunctions, if the whole system has to rely on any centralized

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

control unit. To ensure nonstop services and quick responses for local markets, major Big Data-related applications such as Google, Flicker and Facebook, large number of server farms are deployed all over the world. 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. The complexity and the relationships of data also increases as the volume of the Big Data increases.



The graph shows different types of data structures with 80-90% of the future data growth coming from non-structured data types.

4. Data Mining

Data mining refers to extracting or mining knowledge from large amount of data. It is also defined as finding hidden information from a database. It is used primarily for discovering unknown patterns and that converts raw data into user understandable information. Nowadays it is being increasingly used in science and technology to extract the vast amount of data.

The main purpose of data mining is to collect the needed and useful information from vast and enormous amount of data. The sub domain of data mining is used in extracting the unique text which is termed as text mining which is used to retrieve the main texts from unstructured and semi-structured textual formats. Generally, all data in the web and social media are available in random manner. Therefore, the main aim is to establish a relationship among the texts in order to make it easier for human to understand and process it in an effective manner. This process is known as Knowledge discovery from texts (KDT).

5. Data Mining Lifecycle:

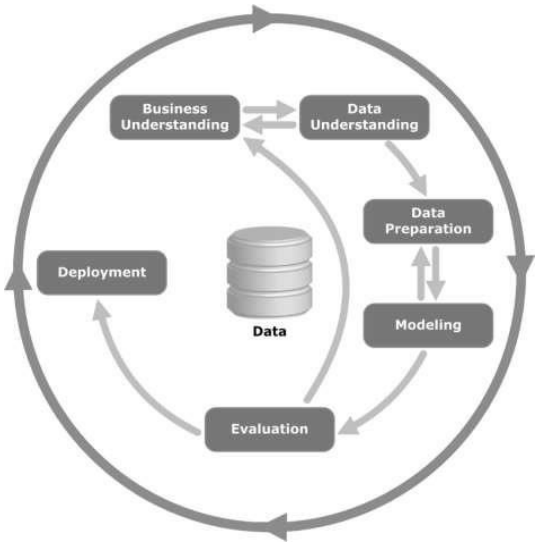


Fig 5.1: lifecycle of Data mining

- **Business Understanding** : This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition. A preliminary plan is designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.

- **Data Understanding** : This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

- **Data Preparation** : This phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

- **Modeling** – In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

have specific requirements on the form of data. Therefore, it is often required to step back to the data preparation phase.

- **Evaluation** − At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model thoroughly and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

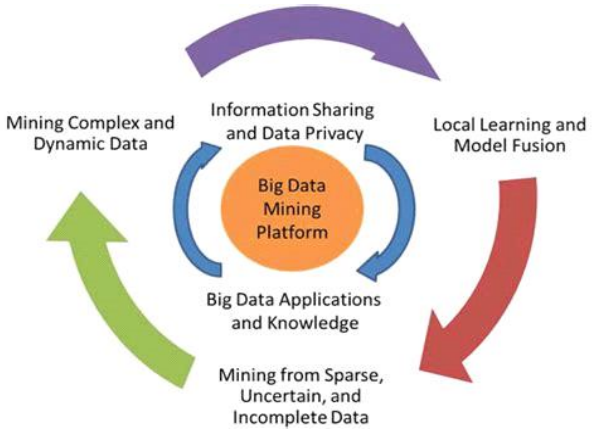
  A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- **Deployment** − Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer.

  Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process.

.

6. Data Mining Challenges with Big Data



.

Fig. 6.1. A Big Data processing framework

The above figure shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Stage 1), data privacy and domain knowledge (Stage 2), and Big Data mining algorithms (Stage 3).

Stage 1: Big Data Mining Platform

The challenge at stage 1 is focused on data accessing and arithmetic computing procedures. Big Data are often stored at different locations and data volumes may continuously grow, therefore, an effective computing platform will have to take distributed large-scale data storage in order to compute. In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For example, typical data mining algorithms require all data to be loaded into the main memory, this, however, is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data for computing. Common solutions are to rely on parallel computing , or collective mining to sample and aggregate data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process. For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with a high-performance computing platform, with a data mining task being deployed by running some parallel programming tools, such as MapReduce or Enterprise Control Language (ECL), on a large number of computing nodes (i.e., clusters).

Stage 2: Big Data Semantics and Application Knowledge

Semantics and domain knowledge for different Big Data applications is the challenge at Stage 2. This information is useful to the mining process, as well as increase technical barriers to the Big Data access and mining algorithms. For example, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different depending on different domain applications. Semantics and application knowledge in Big Data refer to numerous aspects related to the regulations, policies, user knowledge, and domain information. The two most important issues at this tier include 1) data sharing and privacy; and 2) domain and application knowledge. To protect privacy, two common approaches are to either restrict access to the data by adding certification or access control to the data entries so that sensitive

information is accessible by a limited group of users only, or anonymize data fields such that sensitive information cannot be pinpointed to an individual record. Domain knowledge can help identify right features for modeling the underlying data (e.g., blood glucose level is clearly a better feature than body mass in diagnosing Type II diabetes). The domain and application knowledge can also help design achievable business objectives by using Big Data analytical techniques.

Stage 3: Big Data Mining Algorithms

At this Stage, the data mining challenges is to find suitable algorithms in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics. This contains three stages. Initially the sparse, heterogeneous, uncertain and incomplete data are preprocessed by data fusion techniques. Second, complex and dynamic data are extracted from the sources after preprocessing. Third, information and knowledge obtained is tested and relevant information is fed back to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing. A Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two-step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthesized by aggregating patterns across all sites. At the knowledge level, model correlation analysis investigates the relevance between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

Spare, uncertain, and incomplete data are defining features for Big Data applications. For most machine learning and data mining algorithms, high-dimensional spare data significantly deteriorate the reliability of the models derived from the data. Common approaches are to employ dimension reduction or feature selection to reduce the data dimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining. For uncertain data, the major challenge is that each data item is represented as sample distributions but not as a single value, so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters. Incomplete data refer to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission). While most modern data mining algorithms have in-built solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data). Many imputation methods exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance. The emergence of Big Data has also spawned new computer architectures for real-time data-intensive processing, such as the open source Apache Hadoop project that runs on high-performance clusters. The size or complexity of the Big Data, including transaction and interaction data sets, exceeds a regular technical capability in capturing, managing, and processing these data within reasonable cost and time limits.

7. Conclusion

Big data refers to collections of datasets that, due to size and complexity, are difficult to store, query, and manage using existing data management tools or data processing applications. Different information collectors have their own way for data recording. The nature of different applications also results in different representations. This paper also gives the basic idea about different challenges in data mining and their proposed solution.

8. References

[1] 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.

[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2014.

R.Suresh, S.R.Harishni, " Data Mining and Text Mining – A Survey", Sri ManakulaVinayagar Engineering College, Puducherry, India, International Conference on Computation of Power,Energy, Information and Communication (ICCPEIC),2017.

[3] Libina Rose Sebastian, Sheeba Babu, Dr. Jubilant J Kizhakkethottam, "Challenges with Big Data Mining: A Review", International Conference on Soft-Computing and Network Security (ICSNS), Coimbatore, India, 2015.

[4] R.Suresh, S.R.Harishni, " Data Mining and Text Mining – A Survey", Sri ManakulaVinayagar Engineering College, Puducherry, India, International Conference on Computation of Power,Energy, Information and Communication (ICCPEIC),2017.

[5] S.Umadevi, Dr.K.S.Jeen Marseline, " A survey on data mining classification and algorithm", Sri Krishna arts and science college, coimbatore, kerala, India. International Conference on Signal Processing and Communication (ICSPC'17) July 2017.