International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 02, 2021

Applying Datamaing To Big Data: A Review

Sharon Dominick[1], Sherine Dominick[2]

[1]UG Department of Computer Applications, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli-17.
[2]Department of Computer Science, St. Joseph's College, Affiliated to Bharathidasan University, Tiruchirappalli-2.

Email: [1]sharondominick@gmail.com, [2]sherinedominick@gmail.com

Abstract: Big data is all about handling with massive amounts of knowledge. While it has become a highlighted buzzword since last year, "big data mining", i.e., mining from big data, has soon followed up as an emerging, interrelated research area. This paper provides an summary of massive processing and discusses the related challenges and thus the new opportunities. The discussion includes a review of state-of-the-art frameworks and platforms for processing and managing big data also because the efforts expected on big processing. Several issues related to big data and large processing and means opportunities and research topics are addressed well. there's hope that the difficulty which can help reshape the subject area of today's processing technology toward solving tomorrow's bigger challenges emerging in accordance with big data. This paper focuses on extending the utilization of datamining for giant data.

Keywords: Data mining, Big data, knowledge discovery, predictive analysis.

1. INTRODUCTION

The practise of mining data for hidden relationships and forecasting future trends has a long history. The term "data mining," also known as "knowledge discovery in databases," was not coined until the 1990s. However, it is built on the foundation of three linked scientific disciplines: statistics, AI, and machine learning. Advances in processing power and speed have enabled us to shift away from manual, arduous, and time-consuming data analysis and toward quick, easy, and automated data analysis during the previous decade. The more complex the data sets collected, the more likely it is that meaningful insights will be discovered.

Data mining[15] is being used by retailers, banks, manufacturers, telecommunications providers, and insurers to discover relationships between everything from price optimization, promotions, and demographics to how the economy, risk, competition, and social media are affecting their business models, revenues, operations, and customer relationships.

Larger, more complicated data collections, especially from new data sources, are referred to as big data. Because these data sets are so large, typical data processing technologies cannot handle them. However, these vast amounts of data can be leveraged to solve business challenges that previously could not be solved. Big data[16] refers to larger, more intricate data sets, especially those derived from new data sources. Due to the size of these data sets, traditional data processing systems are incapable of handling them. The rate at which data is received and (perhaps) acted on is referred to as velocity. In most cases, data is streamed directly into memory rather than being written to disc. The numerous different sorts of data that are available are referred to as variety. Traditional data formats were well-structured and fit into a relational database with ease. With the rise of big data, new unstructured data kinds have emerged.

2. RELATED WORK

From a data mining perspective, Xindong Wu et al., [1], provide a HACE theorem that characterises the features of the Big Data revolution and suggests a Big Data processing paradigm.

Domenico Talia [2] elaborate how to scale knowledge discovery services, as well as the Data Mining Cloud Framework, which is built for designing and executing distributed data analytics applications as service workflows. Data sets, analysis tools, data mining techniques, and knowledge models are all implemented as single services in this environment.

Hetal Thakkar et al., [3], propose a system with an architecture that addresses challenges by introducing new constructs and synoptic data structures to express complex KDD queries and efficiently support an integrated library of mining algorithms that are fast and light enough to be effective on data streams, as well as support for the Mining Model Definition Language, which allows users to define new mining algorithms.

Longbing Cao [4], propose domain-driven data mining as a framework for developing next-generation approaches, strategies, and tools in preparation for a probable paradigm change from data-centered hidden pattern mining to domain-driven actionable knowledge delivery. Theoretical foundations, many general and adaptable frameworks, research challenges, and prospective directions are all addressed in the D 3 M concept map.

Matthew Herland et al.,[5], highlight latest research on the analysis of Health Informatics data gathered at numerous levels, including molecular, tissue, patient, and population levels, utilising Big Data tools and methodologies. Multiple levels of questions are addressed, in addition to acquiring data at multiple levels: human-scale biology, clinical-scale biology, and epidemic-scale biology.

Dunren Che et al., [6], give an overview of big data mining and analyses the challenges and new opportunities that it presents. A review of state-of-the-art systems and platforms for processing and managing big data, as well as the planned initiatives on big data mining, are included in the debate. Several challenges relating to big data, big data mining, and big data mining indicate to potential and research areas that will be fully fleshed out in the future.

Stančin et al., [7], compare the features of several data mining and large data analysis packages. More than 20 libraries are considered in this article, which are divided into six categories: core libraries, data preparation, data visualisation, machine learning, deep learning, and big data. Larger communities increase the likelihood of quickly discovering a solution to an issue. For data preparation, we propose pandas; for data visualisation, Matplotlib, seaborn, or Plotly; for machine learning, scikit-learn; for deep learning, TensorFlow, Keras, and PyTorch; and for big data, Hadoop Streaming and PySpark.

Vitthal Yenkar et al., [8], provide the reader with a historical and comprehensive overview of the current trend in high-performance computing systems, particularly as it relates to Data Mining and Analytics. There are a number of readings on Big Data and High Presentation that can be found separately. Big Data analytics and algorithms, as well as computing for massively parallel processing databases.

Shengping et al., [9] offer a quick rundown of DM-related subjects is offered. Commonly used data preparation and preprocessing methodologies, DM functions and techniques, and performance metrics are summarised in a DM flowchart and the essential content of the flowchart steps.

Wei Fan et al.,[10] provide a wide review of the topic, its current position, controversy, and a future outlook. We present four articles published by prominent scientists in the area that address the most intriguing and cutting-edge subjects in Big Data mining.
Joseph C. Mellor et al. [11], discuss the process, tools, and interpretation of data mining with a focus on its use in audiology. Modern hearing aids have data-logging technology that records information other than the audio stream, such as the acoustic conditions in which the device was used and how the signal processing responded as a result.
Cheng Ying et al [12], examine the evolution of DMTs in the big data era, as well as the applications of DMTs in production management. Meanwhile, they point out some limits and make some recommendations about the usefulness and future applications of DMTs in production management.
Shoban Babu Sriramoju[13], examine the benefits of Big Data Mining as well as the security considerations. For experiments, benchmark datasets are gathered from the UCI machine learning repository. For association rule mining, Modified Aproiri is utilised. Differential privacy is a data security strategy that requires calculations to be unaffected by changes to data in any database record.

3. MINING BIG DATA

a) Heterogeneous mixture data analysis

It involves breaking up the inherent heterogeneous mixture properties by putting the data in groups with the same patterns or rules allows for accurate analysis of heterogeneous mixture data. However, due to the vast number of data grouping alternatives available, it is almost difficult to check each and every candidate. The following are the three problems.
i)focus on the number of groups
ii)method of grouping
iii)Choice of prediction model

b) Data mining based on heterogeneous mixture learning

By avoiding concerns linked to data grouping or a sudden increase in prediction model combinations, this technology is capable of high-speed optimization of the three issues. The contrasts between learning with the preceding strategies are explained below. This allows for the investigation of models with high prediction accuracies rather than searching for unpromising possibilities in order to discover the best data grouping and prediction model. The latest machine learning theory called factorised asymptotic Bayesian inference supports the sophisticated search and optimization of heterogeneous mixture learning.
Prediction and description are the two main methods for extracting information from data. It is difficult to determine what the data indicates. Data mining is a technique for summarising and simplifying data in a way that we can recognise, allowing us to acquire information about individual cases based on trends. The goal of data mining is usually either prediction or classification. The idea of categorization is to group data into groups. For example, a vendor would be interested in the characteristics of people who responded to an advertisement versus those who did not. There are two groups. The goal of prediction is to forecast the rate of a continuous variable.

4. CONCLUDING REMARKS

On the whole datamining is a very promising field for research and data has been and is always expanding in size and all forms. This has resulted in the massive growth and flow of data in all sectors, there is an urgent need to handle process and filter knowledge out of them. Hence Big data with all its forms is to be mined with some promising methods and approaches of data mining.

5. REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, “Data mining with big data”, IEEE Transactions on Knowledge and Data Engineering, Volume: 26, Issue: 1, Jan. 2014.

[2] Domenico Talia, “Making knowledge discovery services scalable on clouds for big data mining”, 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), 8-10 July 2015.

[3] Hetal Thakkar, “A Data Stream Mining System”, 2008 IEEE International Conference on Data Mining Workshops, 15-19 December 2008.

[4] Longbing Cao, “Domain Driven Data Mining (D3M)”, 2008 IEEE International Conference on Data Mining Workshops, 15-19 December 2008.

[5] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald , “A review of data mining using big data in health informatics”, Journal of Big Data, 24th June 2014.

[6] Dunren Che, Mejdl Safran and Zhiyong Peng, “From Big Data to Big Data Mining: Challenges, Issues, and Opportunities”, International Conference on Database Systems for Advanced Applications, DASFAA 2013.

[7] Stančin, A. Jović, “An overview and comparison of free Python libraries for data mining and big data analysis”, 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 20-24 May 2019.

[8] Vitthal Yenkar et al., “Review on Data Mining with Big Data”, IJCSMC, Vol. 3, Issue. 4, April 2014.

[9] Shengping et al., “A Review of Data Mining with Big Data towards Its Applications in the Electronics Industry”, mdpi, April 2018.

[10] Wei Fan et al., “Mining big data: current status, and forecast to the future”, ACM SIGKDD Explorations Newsletter, Volume 14Issue 2, December 2012.

[11] Joseph C. Mellor et al., “Application of Data Mining to “Big Data” Acquired in Audiology: Principles and Potential”, SAGE, May 31, 2018.

[12] ChengYing et al., “Data and knowledge mining with big data towards smart production”, Journal of Industrial Information Integration, Volume 9, March 2018.

[13] Shoban Babu Sriramoju, “OPPORTUNITIES AND SECURITY IMPLICATIONS OF BIG DATA MINING”, International Journal of Research In Science & Engineering, Volume: 3 Issue: 6 Nov-Dec 2017.

[14] Gandomi, A., Haider, M. et al,, “Beyond the hype: Big data concepts, methods, and analytics,” International Journal of Information Management, 2016.

[15] Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2013), “Addressing big data issues in scientific data infrastructure”, International Conference on Collaboration Technologies and Systems (CTS), 2013.

[16] Richa Gupta, “Journey from data mining to Web Mining to Big Data”, IJCTT,2014.