Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2021, 9, 23-29
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Designing a Model to Study Data Mining in Distributed Environment

Md. Abadur Rahman[1], Masud Karim[2]

[1]School of Business & Tourism, Southern Cross University, Sydney, Australia
[2]Department of Computer Science & Engineering, Daffodil Institute of Science & Technology, Dhaka, Bangladesh
Email: md.abadur.rahman87@gmail.com, masud.mkarim@gmail.com

How to cite this paper: Rahman, Md.A. and Karim, M. (2021) Designing a Model to Study Data Mining in Distributed Environment. *Journal of Data Analysis and Information Processing*, **9**, 23-29.
https://doi.org/10.4236/jdaip.2021.91002

Received: December 30, 2020
Accepted: February 22, 2021
Published: February 25, 2021

Abstract

To make business policy, market analysis, corporate decision, fraud detection, etc., we have to analyze and work with huge amount of data. Generally, such data are taken from different sources. Researchers are using data mining to perform such tasks. Data mining techniques are used to find hidden information from large data source. Data mining is using for various fields: Artificial intelligence, Bank, health and medical, corruption, legal issues, corporate business, marketing, etc. Special interest is given to associate rules, data mining algorithms, decision tree and distributed approach. Data is becoming larger and spreading geographically. So it is difficult to find better result from only a central data source. For knowledge discovery, we have to work with distributed database. On the other hand, security and privacy considerations are also another factor for de-motivation of working with centralized data. For this reason, distributed database is essential for future processing. In this paper, we have proposed a framework to study data mining in distributed environment. The paper presents a framework to bring out actionable knowledge. We have shown some level by which we can generate actionable knowledge. Possible tools and technique for these levels are discussed.

Keywords

Data Mining, Distributed Database, Knowledge Discovery, Classification Algorithm

1. Introduction

This research is to develop a framework, which will help to study data mining in distributed environment. There are many sectors where there are large data base and data warehouse. Agriculture, bank, insurance, medical, education, law, cen-

sus, stock exchange etc. are most of them. These data can be used to make strategic planning [1] [2]. It will also help for decision making and long-term problem solving. The proposed framework can be used for further research and designing algorithms for data mining.

Our objective is to create actionable knowledge [3] [4]. Actionable knowledge is being used to study in data analysis. Data mining application and its research is increasing rapidly. Knowledge discovery is also coming into sharp by using data mining. We have to perform a series of steps in data mining for data cleaning, data selection and transformation, pattern evaluation, visualization, etc. Considering such steps, we are proposing framework to study data mining. Users can do something to bring direct benefits like increase in profits, reduction in cost, improvement in efficiency, building strategic policy, geographical area selection, etc., from the proposed framework. So framework that we proposed can provide the advantages for organizations. Massive amount of data is now available to public [2].

The main subject of this research is to the study is data mining for applying distributed computing for a strategic policy maker and data analyst. The objectives we want to achieve by proposing a new method for designing conceptual framework in distributed environment to gain actionable knowledge. It will demonstrate how actionable knowledge can be generated based on base level data to take decision. It can be applied in Pattern Recognition and Knowledge Discovery (KDD) to supporting management and policy maker [5] [6].

2. Related Work

S. Urmela and others [7] have described a study completely based on the data mining techniques. According to the writers, DDM refers to the distributed data mining. It is an important data mining environment that have the ability to mine larger data in less time. Data mining is a kind of process which is capable to extract effective information from datasets and facts so that reliable decisions can be developed [1]. The writers provided a way where numerous DDM techniques can be reviewed and the level of understanding can be improved. Researchers have classified distributed data mining into major three sections for example agent-based, a classifier based and privacy-preserving. The writers included qualitative designs that helped to gain effective information about DDM techniques and also conducted a literature review in order to evaluate the techniques are applying in recent research. In terms of effectiveness, this paper is more effective that provided complete information about distributed data mining and also included viewpoints of other writers. However, the researchers did not cover benefits and challenges linked with the distributed data mining techniques that produced a research gap but future research will cover such information.

Yuan Huang and others [8] have reviewed data mining programs used for cloud networks and AI systems. This paper provided a platform to demonstrate applications and importance of data mining algorithms among cloud and AI de-

vices [9]. According to the writers, companies are now moving towards emerging technologies in order to develop effective plans and manage complexity. AI is a common technology used in businesses in order to develop automated systems where data-mining programs are helpful for extracting information and managing data handling issues. The researchers highlighted that in the field of cloud computing, data mining has developed a significant platform where collected data can be analyzed easily. For enhancing the performance of data mining among cloud networks, the writers developed the parallel extension of fuzzy c-means. Numerous kinds of data mining algorithms are included in this paper that helped to expand fuzzy c-means clustering for distributed networks and applications. The researchers provided depth information and covered all facts and points related to the distributed data mining. However, it is important to provide information about the challenges faced by the companies while using data mining, cloud networks and AI technology.

David Savage and other [10] proposed novel algorithm based on the distributed mining programs. The objective of this research is to develop a significant algorithm for mining contrast patterns based on distributed techniques [11]. According to the writers, mining for contrast pattern is a critical task but distributed mining is more effective which is capable to extract information and patterns from images effectively. Numerous kinds of methods are included for example qualitative design, inductive approach, and secondary data collection and content analysis techniques. It is found that current data extraction approaches for contrast patterns are not able to manage dense and larger databases but distributed mining-based programs are more effective that helped to manage such problems. The writers provided a significant mining algorithm by which data can be extracted from larger dimensional databases easily. The writers tested a spark cluster using distributed data mining programs. The future research will include applications and different types of algorithms used in distributed data mining companies.

A few research frameworks currently exist [1] [5] [11] for using data mining in distributed fashion. Some of them are designed for general environment to execute mining tasks [11]. Some are implemented by using single data mining algorithms. But more algorithms are required for distributed environment. So a general framework is required for applying data mining tools and techniques [1] [12]. Some researchers have introduced the rationale as new architecture due to huge amount of data and increasing data rapidly. Moreover, data complexity, diversity is also increasing [9].

Some advance data mining in distributed environment like national security and crime detection are essential [1] [11]. Such tasks process knowledge discovery in databases [3] [4] [13]. Data mining in distributed environment is a scientific computing task. It provides facilities of remote computer connected through a network (TCP/IP) to its scientific partners. So using data mining in distributed environment researchers concern in field of Simulation, Data Analysis and Machine Learning [14] [15]. Due to vast development of Internet data mining are

taking place in distributed environment and suitable frameworks are needed in common fashion [5] [6].

There are some important requirements those cannot be handle easily in distributed environment, such as Corporate Decision Handling, Decision support system, Management support system, Strategic policy making, Information retrieval, Knowledge analysis, Intrusion detection, etc.

3. Proposed Framework

This proposed framework can play a significant role in distributed database for effective computational and application regarding knowledge discovery [2] [16]. The proposed framework will provide an environment to perform analysis, clustering, classification, associate etc to find hidden information from large distributed database. Strategic policy or vital decision can be made from that information. A knowledge discovery [4] also works for actionable knowledge. The actionable knowledge provides a good solution to make decision and also provides an idea of path selection to reach in a certain goal. Decision tree is an important issue in this area. Different data mining algorithms [13] can be used. A sample diagram of the proposed framework is shown in **Figure 1**.

The proposed framework mine data in distributed environment. In base line

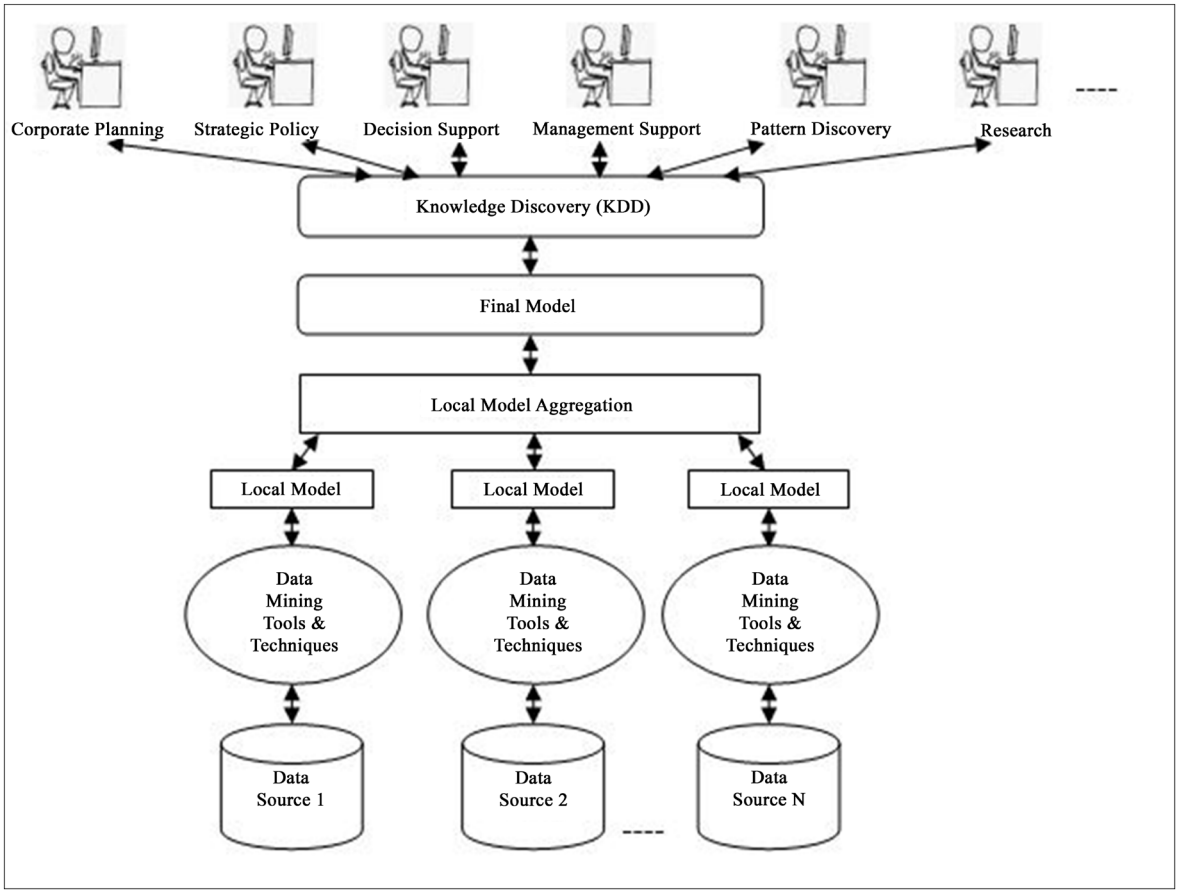

Figure 1. Proposed data mining framework in distributed environment.

data are stored in local database. This may be different work stations or computers those are connected through network [17] [18]. The model provides better algorithms and performance in distributed architecture [16]. The main steps of the framework that we have proposed:

- Perform the data mining query in the distributed database;
- Create a local data mining model in local level;
- Combined the local model in central site;
- Produce a single model based on combination;
- Apply query on the central combined model to produce required result;
- In necessary back to lower level for better analysis.

Our proposed research work can also be seen as a knowledge framework. It will be as a system for knowledge discovery (KDD) [9] in distributed database. KDD plays important role in data mining process. The proposed framework will use actionable knowledge [3] [4] to build specific knowledge discovery services. The framework will provide users with high-level abstractions and a set of services by which is possible to integrate distributed resources to support all the phases of the KDD process. Therefore, it will allow end-users or decision maker to discover knowledge without worrying about distributed environment.

We will use some techniques in the proposed framework. Some techniques may be changed depending on situation whenever developing the framework and also novel techniques can be added. Available techniques are as the followings:

- Classifier learning with meta learning and knowledge discovery framework;
- Collective and collaboration data mining;
- Clustering and association rule mining in distributed environment.

There are many popular algorithms available for distributed model. Most of the model is support by decision tree algorithm and Naïve Bayes algorithm [19]. Due to get some facilities researchers are also using association rule [16] [19]. Our proposed framework to study data mining to deal complex systems [5] [6] [17]. Distributed data sources that we proposed will work with both homogeneous and heterogeneous data [20].

The key requirements to do mining in this framework is using the better applications of algorithms based on Attribute-value description and required output [4] [13]. Also, we want to use data mining algorithms to created data model for data analysis. The algorithms should be studied for specific pattern and trends.

We will try to use at agriculture and bank sector in Bangladesh. Using this novel technique long term strategic policy in agriculture or bank will be made. So, these sectors will be developed more and there will be a big change for well-being [20] [21].

4. Conclusions

The proposed framework is designed for learning association rules [1], classifier and meta classifier [9], link analysis, frequent episodes [18] for sequence analysis and supporting distributed environment. While, analyzing and summarizing raw

data is an essential task for this framework. There are some similar successful projects as Nimrod, Keplar, Taverna, WEKA, GridMiner, RapidMiner, R Language, etc.

Advanced tools and techniques play significant roles in distributed environments over the wireless and wired network. The environment deals with many sources of data and unlimited computer nodes with unlimited user. To analysis data, suitable framework is required to work with the complex system.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Thitiprayoonwongse, D., Suriyaphol, P. and Soonthornphisaj, N. (2012) A Data Mining Framework for Building Dengue Infection Disease Model. *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, Japan, 12-15 June, 2012, 1-8.

[2] Devale, A.B. and Kulkarni, R.V. (2012) Applications of Data Mining Techniques in Life Insurance. *International Journal of Data Mining & Knowledge Management Process*, **2**, 31-40.

[3] Alam, M. and Akhlaque Alam, S. (2012) Actionable Knowledge Mining from Improved Post Processing Decision Trees. *International Conference on Computing and Control Engineering* 2012, Chennai, 12-13 April 2012, 1-8.

[4] Patil, R.A., Ahire, P.G., Patil, P.D. and Golande, A.L. (2012) Decision Tree Post Processing for Extraction of Actionable Knowledge. *International Journal of Engineering and Innovative Technology*, **2**, 152-155.

[5] Talia, D., Trunfio, P. and Verta, O. (2005) Weka4WS: A WSRF-Enabled Weka Toolkit for Distributed Data Mining on Grids. *European Conference on Principles of Data Mining and Knowledge Discovery*, Porto, 3-7 October 2005, 309-320. https://doi.org/10.1007/11564126_32

[6] Talia, D., Trunfio, P. and Verta, O. (2008) The Weka4WS Framework for Distributed Data Mining in Service-Oriented Grids. *Concurrency and Computation: Practice and Experience*, **20**, 1933-1951. https://doi.org/10.1002/cpe.1311

[7] Urmela, S. and Nandhini, M. (2017) Approaches and Techniques of Distributed Data Mining: A Comprehensive Study. *International Journal of Engineering and Technology*, **9**, 63-76. https://dx.doi.org/10.21817/ijet/2017/v9i1/170901408

[8] Huang, Y., Cheng, Z., Zhou, Q.Y., Xiang, Y.X. and Zhao, R.X. (2020) Data Mining Algorithm for Cloud Network Information Based on Artificial Intelligence Decision Mechanism. *IEEE Access*, **8**, 53394-53407. https://doi.org/10.1109/ACCESS.2020.2981632

[9] Atkinson, M.P., van Hemert, J.I., Han, L.X., Hume, A. and Liew, C.S. (2009) A Distributed Architecture for Data Mining and Integration. *Proceedings of the 2nd International Workshop on Data-Aware Distributed Computing*, Munich, June 9-10 2009, 11-20. https://doi.org/10.1145/1552280.1552282

[10] Savage, D., Zhang, X.Z., Chou, P., Yu, X.H. and Wang, Q.M. (2017) Distributed Mining of Contrast Patterns. *IEEE Transactions on Parallel and Distributed Sys-*

tems, **28**, 1881-1890. https://doi.org/10.1109/TPDS.2016.2637914

[11] Lee, W., Stolfo, S.J. and Mok, K.W. (1999) A Data Mining Framework for Building Intrusion Detection Models. Supported in Part by Grants from DARPA (F30602-96-1-0311) and NSF (IRI-96-32225 and CDA-96-25374), Computer Science Department, Columbia University, New York.

[12] Thavavel, V. and Sivakumar, S. (2012) A Generalized Framework of Privacy Preservation in Distributed Data Mining for Unstructured Data. *International Journal of Computer Science Issues*, **9**, 434-441.

[13] Kulkarni, S.V. (2011) Mining Knowledge Using Decision Tree Algorithm. *International Journal of Scientific & Engineering Research*, **2**, 1-6.

[14] Cannataro, M., Pugliese, A., Pugliese, A., Talia, D. and Trunfio, P. (2004) Distributed Data Mining on Grids: Services, Tools, and Applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **34**, 2451-2465. https://doi.org/10.1109/TSMCB.2004.836890

[15] Alli, P., SelvaLakshmi, C.B. and Murali, S. (2012) Distributed Data Mining in the Grid Environment. *International Journal of Engineering and Innovative Technology*, **2**, 240-243.

[16] Sujni, P. (2010) An Optimized Distributed Association Rule Mining Algorithm In Parallel and Distributed Data Mining With Xml Data For Improved Response Time. *International Journal of Computer Science and Information Technology*, **2**, 88-102. https://doi.org/10.5121/ijcsit.2010.2208

[17] Vuda Sreenivasa, R. (2009) Multi Agent-Based Distributed Data Mining: An Overview. *International Journal of Reviews in Computing*, **2009-2010**, 83-92.

[18] Padhy, N., Mishra, P. and Panigrahi, R. (2012) The Survey of Data Mining Applications and Feature Scope. *International Journal of Computer Science, Engineering and Information Technology*, **2**, 43-58. https://doi.org/10.5121/ijcseit.2012.2303

[19] Pardeep Kumar, N., Vivek Kumar, S. and Durg Singh, C. (2012) A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems. *International Journal of Data Mining & Knowledge Management Process*, **2**, 25-42. https://doi.org/10.5121/ijdkp.2012.2503

[20] Venkatadri. M. and Reddy, L.C. (2011) A Review on Data mining from Past to the Future. *International Journal of Computer Applications*, **15**, 19-22. https://doi.org/10.5120/1961-2623

[21] Oueslati, W. and Akaichi, J. (2010) A Survey on Data Warehouse Evolution. *International Journal of Database Management Systems*, **2**, 11-24. https://doi.org/10.5121/ijdms.2010.2402