www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-7, Issue-2, Apr 2019
Review Paper

Int. J. Sc. Res. in Network Security and Communication

ISSN: 2321-3256

Data Mining tools and challenges for current market trends-A Review

Rakesh Kumar Saini

Department of Computer Application, DIT University, Dehradun, Uttrakhand, India

Received: 10/Mar/2019, Accepted: 15/Apr/2019, Published: 30/Apr/2019

Abstract- Data mining is a promising and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence (AI) and statistical. Data mining is used regularly in a variety of industries and is continuing to gain in both popularity and acceptance. However, applying data mining methods to complex real-world tasks is far from straight forward and many pitfalls face data mining practitioners. Data mining tools can analyze massive databases to deliver answers to questions such as, which clients most likely to respond to my next promotional mailing, and why? However, most research in the field tends to focus on the algorithmic issues that arise in data mining and ignores the human element and process issues that are often the cause of these pitfalls. Data mining have many advantages but still data mining systems face lot of problems and pitfalls. In this paper approximately 23 research papers were collected concerning various fields in data mining and discussed each and categorized them under the few areas and the trends was interpreted based on the area of research and applications.

Keywords- Data mining tools, Database, Data warehouse, Knowledge Discovery in Database.

I. INTRODUCTION

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining is the analysis of observational data sets to find. An Increasing amount of work is beginning to focus on the characteristics of real-world problems that makes data mining difficult there is still relatively little work that describes the many practical issues that arise when addressing real data mining problems. Data mining is an interdisciplinary field bringing together techniques from Machine

Learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases [1, 2].

The growth in the field of data mining and knowledge discovery has been fastened by a variety of factors:

- The growth in data collection, as exemplified by the supermarket.
- The storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database.
- The availability of increased access to data from Web navigation and intranets.
- The competitive pressure to increase market share in a globalized economy.
- The tremendous growth in computing power and storage capacity [1].

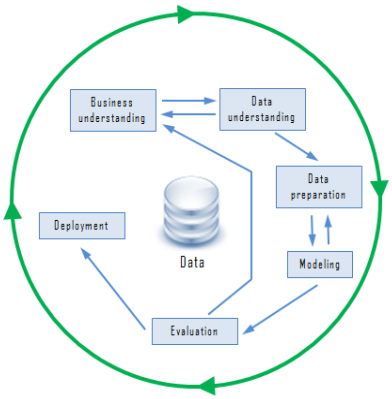

Figure 1. Life cycle of data mining

II. EVOLUTION OF DATA MINING

The evolution of data mining 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 [4]. Data mining is ready for application in the business world because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at rapid rate. The accompanying need for improved computational engines can now be met in a cost effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods. In the evolution from business data to business information, each new step has built upon the previous one. From the user's point of view, the four steps listed below were revolutionary because they allowed new business questions to be answered accurately and quickly.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments[2][3]. The data mining is having various challenges and issues which are discussed below:

III. DATA MINING CHALLENGES

The shift towards intrinsically distributed complex problem solving environments is prompting a range of new data mining research and development problems [1, 5]. These can be classified into the following broad challenges:

1. **Distributed data:** The data to be mined is stored in distributed computing environments on heterogeneous platforms. Both for technical and for organizational reasons it is impossible to bring all the data to a centralized place. Consequently, development of algorithms, tools, and services is required that facilitate the mining of distributed data.

2. **Distributed operations:** In future more and more data mining operations and algorithms will be available on the grid. To facilitate seamless integration of these resources into distributed data mining systems for complex problem solving, novel algorithms, tools, grid services and other IT infrastructure need to be developed.

3. **Data privacy, security, and governance:** Automated data mining in distributed environments raises serious issues in terms of data privacy, security, and governance. Grid-based data mining technology will need to address these issues.

4. **Massive data:** Development of algorithms for mining large, massive and high-dimensional data sets (out-of-memory, parallel, and distributed algorithms) is needed.

Complex data types: Increasingly complex data sources, structures, and types (like natural language text, images, time series, multi-relational and object data types etc.) are emerging. Grid-enabled mining of such data will require the development of new methodologies, algorithms, tools, and grid services.

5. **User-friendliness:** Ultimately a system must hide technological complexity from the user. To facilitate this, new software, tools, and infrastructure development is needed in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces.

IV. DATA MINING ISSUES

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. There are many important implementation issues associated with data mining [1, 6]:

1. **Human interaction:** Since data mining problems are often not precisely stated, interfaces may be need with both domain and technical experts. Technical experts are used to formulate the queries and assist in interpreting the results. Users are needed to identify training data and desired results.

2. **Over-fitting:** When a model is generated that is associated with a given database state, it is desirable that the model also fit future database states. Over-fitting occurs when the model does not fit future states. This may be caused by assumptions that are made about the data or may simply be caused by the small size of the training database.

3. **Outliers:** There are often many data entries that do not fit nicely into the derived model. This becomes even more of an issue with very large databases. If a model is developed that includes these outliers, then the model may not behave well for data that are not outliers.

4. **Social One:** One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual

privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

5. **Data integrity:** Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources.

6. **Interpretation of results:** Currently, data mining output may require experts to correctly interpret the results, which might otherwise be meaningless to the average database user.

7. **Visualization of results:** To easily view and understand the output of data mining algorithms, visualization of the results is helpful.

8. **Complex Data:** Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on. It is really difficult to handle these different kinds of data and extract required information. Most of the times, new tools and methodologies would have to be developed to extract relevant information.

9. **Multimedia data:** Most previous data mining algorithms targeted to traditional data types (numeric, character, text, etc.). The use of multimedia data such as images, audio, video complicates or invalidates many proposed algorithms.

10. **Relational or Multidimensional databases:** A hotly debated technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments. And, with the explosion of the Internet, the world is becoming one big client/server environment [12, 13].

11. **Noisy data:** Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors.

12. **Irrelevant data:** Some attributes in the database might not be of interest to the data mining task being developed.

13. **Missing data:** During the pre-processing phase of knowledge discovery in databases (KDD), missing data may be replaced with estimates. This and other approaches to handling missing data can lead to invalid results in the data mining step [1, 10].

14. **Changing data:** Databases cannot be assumed to be static. However, most data mining algorithms do assume a static database. This requires that the algorithm be completely rerun anytime the database changes.

V. DATA MINING APPLICATIONS

There are some applications of data mining are:

1. **Sales/Marketing:** Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way. Retail companies' uses data mining to identify customer's behavior buying patterns [11].

2. **Banking / Finance:** Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection. Data mining is used to identify customer's loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. Credit card spending by customer groups can be identified by using data mining [8, 14].

3. **Health Care and Insurance:** The growth of the insurance industry entirely depends on the ability of converting data into the knowledge, information or intelligence about customers, competitors and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. Data mining enables to forecasts which customers will potentially purchase new policies [1, 11].

VI. CONCLUSION

Data mining brings a lot of benefits to businesses, society, governments as well as the individual. However, privacy, security, and misuse of information are the big problems if they are not addressed and resolved properly. In this paper, number of issues were raised—issues that often are not discussed in research papers on data mining. The concept of data mining, role of data mining its major challenges, issues and application have been focused which help in business strategy formulations, decision making and analysis to the business, society and governments. Hopefully this discussion will provide some insight into the challenges one may encounter when using data mining to solve complex real-world problems.

REFERENCES

[1] Raj Sharma, Kaur Daljeet, A. Manju, "A Review on Data Mining: Its Challenges Issues and Applications." International Journal of Current Engineering and Technology, Vol.3, Issue. 2, 2013.

[2] R. Agrawal. R Shrikant, "Fast algorithms for mining association rule", In the proceeding of 20th international conference on VLDB, pp.487-499, 1994.

[3] W. Yanthy, T. Sekiya., K .Yamaguchi.,,"Mining Interesting Rules by Association and Classification Algorithms", In the proceeding of International Conference on Frontier of Computer Science and Technology, pp. 177-182,2009.

[4] Bansal, Rashi, Akansha Mishra, and Shailendra Narayana Singh. "Mining of educational data for analysing students' overall performance." 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence. IEEE, 2017.

[5] Priyanka, L. T., and Neethu Baby. "Classification approach based Customer Prediction Analysis for Loan Preferences of Customers." International Journal of Computer Applications 67.8 (2013).

[6] Filipova, Biljana Teohareva, and Cveta Martinovska. "Analysing customer profiles using data mining techniques." In Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces, pp. 1-6. IEEE, 2012.

[7] Ping, Zhao Li, and Shu Qi Liang. "Data mining application in banking-customer relationship management." In 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), vol. 6, pp. V6-124. IEEE, 2010.

[8] Rachburee, Nachirat, Jedsada Arunrerk, and Wattana Punlumjeak. "Failure Part Mining Using an Association Rules Mining by FP-Growth and Apriori Algorithms: Case of ATM Maintenance in Thailand." In IT Convergence and Security 2017, pp. 19-26. Springer, Singapore, 2018.

[9] Simha, Jay B., and S. S. Iyengar. "Fuzzy data mining for customer loyalty analysis." In 9th International Conference on Information Technology (ICIT'06), pp. 245-246. IEEE, 2006.

[10] Mishra, Sushruta, Pamela Chaudhury, Brojo Kishore Mishra, and Hrudaya Kumar Tripathy. "An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction." In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, p. 42. ACM, 2016.

[11] Meena, Deepak, and Hitesh Gupta. "A Review: Data Mining over Multi-Relations." International Journal of Computer Applications 62, no. 8, 2013.

[12] Jain, Nikita, and Vishal Srivastava. "Data mining techniques: a survey paper." IJRET: International Journal of Research in Engineering and Technology 2, no. 11 (2013): 2319-1163.

[13] Sharma, Aarti, Rahul Sharma, Vivek Kr Sharma, and Vishal Shrivatava. "Application of Data Mining–A Survey Paper." International Journal of Computer Science and Information Technologies 5, no. 2 (2014): 2023-2025.

[14] Raval, Kalyani M. "Data mining techniques." International Journal of Advanced Research in Computer Science and Software Engineering 2, no. 10, 2012.

Author Profile

Rakesh Kumar Saini received the MCA degree from UPTU, Lucknow, India in 2005 and M.Tech (Computer Science and Engineering) degree from UTU, Dehradun, India in 2012 and PhD from DIT University, Dehradun, India in 2017. He is having 14 Years of teaching experience. He is author of around 10 books. His research interests include cross-layer modification, Energy-Efficiency in Wireless Sensor Network.

