IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 11, November 2018

Data Mining Issues and Challenges: A Review

R Ragavi[1], B Srinithi[2], V S Anitha Sofia[3]

Student, Department of Computer Technology, Sri Krishna Arts and Science College, Coimbatore[1,2]

HOD, Department of Computer Technology, Sri Krishna Arts and Science College, Coimbatore[3]

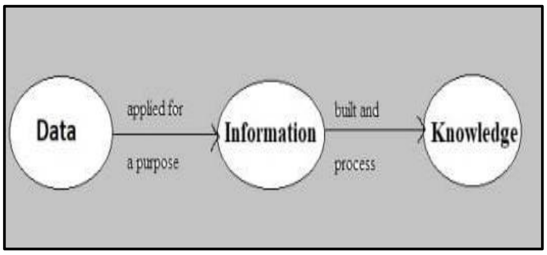
Abstract: Data Mining knowledge and information from high databases has been recognized by many researchers as a key research topic in machine learning and Database system and in many industrial companies as an important area with an opportunity of major revenues .we present a multidimensional view of data mining. The major dimensions are data, knowledge, technologies and application. Researchers in some different fields have shown their great interest in data mining. In this section, we briefly outline methodology, user interaction. Data mining research has strongly impact society and will continue to do so in the future.

keywords: Data Mining, Knowledge discovery

I. INTRODUCTION

Data mining is also called as data or knowledge discovery. It is the process of analyse the data from various perspectives and summarizing it into useful information. Information can be used to highly revenue, break the costs or both. To analyse the data number of analytical tools where used in data mining. It allows users to find the data from many different angles or dimensions, categorize it, and summarize the relationships. Accordingly, data mining is to finding correlations or patterns among bunch of fields in high relational databases.

Data, Information, and Knowledge:



II. KNOWLEDGE MINING

The data mining is refers to extracting or "mining" knowledge from huge amount of data. The team is actually a misnomer. Thinking that the mining of gold from rocks or sand is verified to as gold mining rather than rock or sand mining. Thus the data mining has been highly appropriately named "knowledge mining from data," which is sadly somewhat long. It is a term may not reflect the emphasis on mining from high amount of data.

III. STEPS OF DATA'S

Data preprocessing is an important issue in data mining, as a data can be incomplete, and inconsistent. Data preprocessing techniques can improve the accuracy and efficiency of the subsequent mining process. It is an important step in processing , because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making. Steps in which data's are:

- **Data quality**: Data quality is assumed that the input of KDD is the information from which knowledge is discovered.It is defined in terms of accuracy, completeness, consistency, timeliness, believability, and interpretability. These qualities are assessed based on the intended use of the data.

- **Data cleaning:** In this process noise data and irrelevant data are removed from the collection.it fills missing values, while identifying outliers it will smooth out noise, and correct inconsistencies in the data. Data cleaning involves iterative two-step process consisting of discrepancy detection and data transformation.

DOI 10.17148/IJARCCE.2018.71125

- **Data integration:** It combines data from many sources into a coherent data store, as in data warehousing. These sources may include multiple database, data cubes or flat files etc., store under a unified schema and that usually reside at a single site.

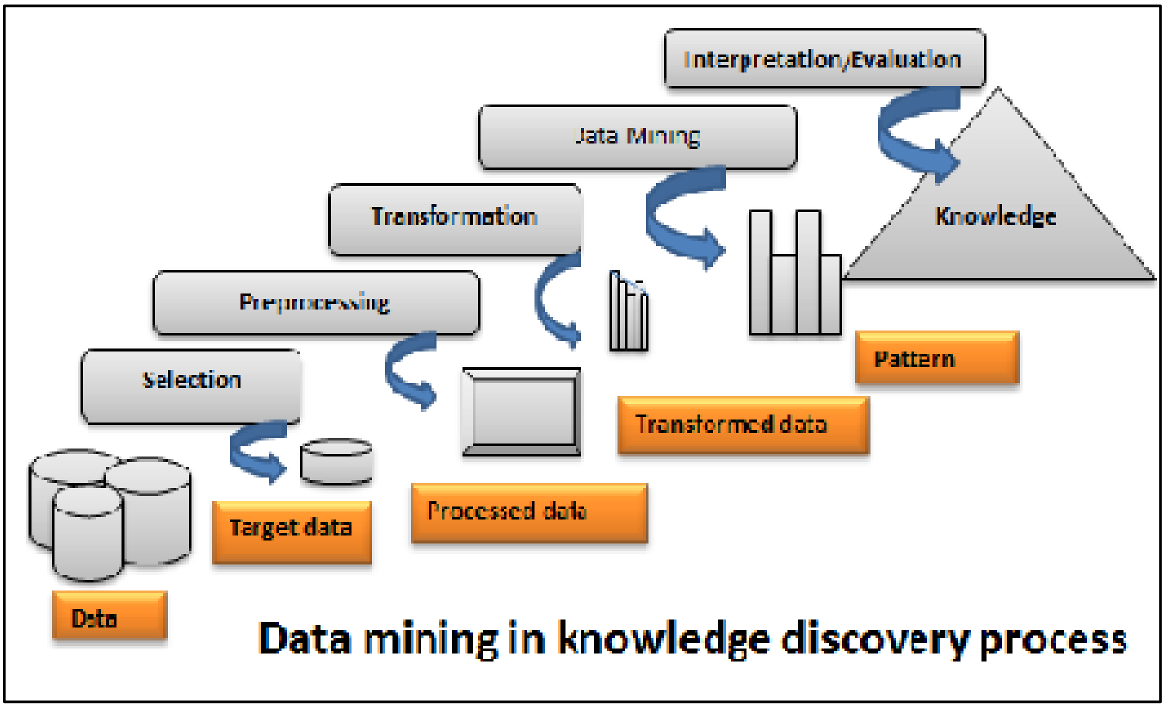
- **Data selection:** In this process data has been relevant to the analysis task are retrieved from its database. With too many variables it is difficult to sense out of results.

- **Data transformation:** Where data has been transformed from appropriate for mining by aggregate operations. The usual process involves performing documents, but data conversions sometimes involve the conversion of a program from one system language to another.

- **Data mining:** A task being performed, where intelligent methods are applied to generate the data patterns which are potentially useful.it is the heart of knowledge discovery process.

- **Pattern evaluation:** The data mining results are presented to the users is extremely important. The interesting patterns representing knowledge are identified based on the given measure. Patterns are represented according to classification rules or classification trees, regression functions, or other knowledge or model representation methods.

- **Knowledge presentation:** It is the final phase where viewing and gathering the representation techniques are used to produce the mined knowledge to the user.



Data mining in knowledge discovery process

IV. DATA MINING ISSUES

Data mining systems depend on the databases to supply the raw input and this raises problems, such as that database tends to be dynamic, incomplete, noisy and large. Other problems arise as a result of the inadequacy and irrelevance of the information stored. In Data mining issues there are so many important implements are there:

- **Over fitting:** Over fitting occurs when the model doesn’t fit future states. When a data mining algorithm searches for the best parameters for a specific model using a set of samples, it may over-fit the data, resulting in poor generalization. Cross-validation, regularization and other sophisticated statistical methods can be applied to overcome the problem.

- **Missing and noisy data:** Missing data are all too common. Ignoring instances with missing values often results in lost information, which is contrary to developing a good data mining model. There are many statistical methods to deal with missing data and identify noisy attribute values.

DOI 10.17148/IJARCCE.2018.71125

- **Size of dataset:** when datasets contains hundreds of fields and tables, millions of records and multi-gigabyte sized files, it is difficult to extract knowledge by a quick mining task. Many algorithms just freeze during the training phase of large datasets. The databases are dynamic and their contents are keep on changing as information is added, modified or removed. The problem with this, from the perspective of data mining, is how to ensure that they are up-to-date and consistent with the most current information.

- **Outliers:** The data entries that do not fit nicely into the derived model. That the model is developed that includes these outliers, then the model may not behave well for data that are not outliers.

- **Higher dimensionality:** This refers not to a large number of records but to a large number of attributes in a datasets-a common situation with bioinformatics datasets where 30000 attributes or more can exits. A high dimensional datasets creates problems in terms of increasing the size of the search space for an efficient model construction to perform the data mining task.

- **Changing data and knowledge:** Rapidly changing data may misguide the data mining users because the developed models may become obsolete before the application. Possible solutions include incremental methods for updating patterns.

- **Mixed dataset:** If a dataset is collated from various sources, the attributes may have values which could be continuous discrete, symbolic, text, images and others types that create problems in constructing good models.

- **Interpretation of results:** Data mining output may require experts to correctly interpret the outputs.

- **Limited Information:** The massive dataset associated with data mining create problems when applying algorithms designed for small datasets. A database is designed for purposes other the data mining and some attributes which are essential for knowledge discovery of the application domain are not present in the data. Thus, it may be difficult to discover significant knowledge about a given domain.

- **Security and social issues:** Security is an important issue with any data collection that is shared and is intended to be used for strategic decision-making. When the data is collected for customer profiling, user behavior understanding, correlating personal data with other information.

- **Efficiency and scalability:** They both are always considered when comparing data mining algorithms. As data amount contains continue to multiply, these two factors are especially critical. Data mining algorithms must be efficient and scalable in order to effectively extract information from high amounts of data in many data repositories or in dynamic data streams. In other words, the running time of an algorithm must be predictable, short and acceptable by applications. Efficiency, scalability, performance, optimization and the ability to execute in real time are key criteria that drive the development of many new data mining algorithm.

- **Human interaction:** User interaction an important role in the data mining process. It describes how a user incorporates background knowledge in mining and to visualize and comprehend data mining result. Since data mining has been interface with both domain and technical expert, its problems are not been stated. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks.

- **Interactive mining:** The user interaction with the system should be an exploratory mining environment. A user may first like to sample a set of data and then explore general characteristic of data and estimate potential

mining result. Interactive mining should allow users to dynamically change the focus of a search, to refine mining request based on the returned result.

- **Incorporation of background knowledge:** Background knowledge, constraints, rules and other information regarding the domain under study must be incorporated into knowledge discovery process. These knowledge are used for pattern evaluation as well as to guide the search towards interesting patterns.

- **Data mining query languages:** Query languages played an important role in flexible searching .this should facilitate specification of the relevant sets of data for analysis, the domain knowledge and constraints to be enforced on the discovered patterns. Optimization of processing flexible mining requests in another promising area of study.

DOI 10.17148/IJARCCE.2018.71125

- **Boosting the power of discovery in a networked environment:** The data objects reside in a linked or interconnected environment, whether it be the Web, database relations, files, or documents. multiple data objects can be used to advantage in data mining by identifying semantic links. to boost the discovery of knowledge, the knowledge derived in one set of objects can be used in semantically linked set of objects.

- **Handling uncertainty, noise or incompleteness of data:** Errors and noise may confuse the data mining process, leading to derivation of erroneous patterns. Data cleaning, data processing, outlier detection and removal are examples of techniques that need to be integrated with the data mining process.

- **Pattern or constraint-guided mining:** The pattern interesting may vary from user to user. Techniques are needed to assess the estimate value of patterns based on subjective measures. These estimates and techniques are needed to access the interestingness of discovered pattern. User specified constraints used to guide the discovery process and generate interesting pattern.

- **Presentation and visualization of data mining results:** The data mining results should be flexible, so that the discovered knowledge can be easily understood .It requires the system to adapt expressive knowledge representations, user-friendly interfaces and visualization techniques. Visualization refers to visual presentation of data. the use of visualization techniques allows the user to summarize, extract, and grasp more complex results .visualization techniques includes graphical, geometric, icon-based, pixel-based, hierarchical, hybrid.

CONCLUSION

This article provides an overview on data mining, the analysis of data includes data quality, data cleansing, data integration, data selection, data transformation, pattern evaluation, knowledge presentation and data mining issues and challenge. Data mining has achieved tremendous success in the late 1980's. The new problems have emerged and it is solved by data mining researchers.

REFERENCES

[1]. M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
[2]. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
[3]. Chieh-Yuan, T. and T. Min-Hong. A dynamic Web service based data mining process system. in Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on. 2005.
[4]. Ding, P. A formal framework for Data Mining process model. In Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on. 2009.
[5]. H. Vernon Leighton and J. Srivastava. Precision Among WWW Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. http://www.winona.msus.edu/isf/libraryf/webind2/webind2.htm, 1997.
[6]. Cooley, B. Mobasher and J. Srivsatava. Web Mining: Information and Pattern Discoveryon the Word Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with AI (ICTAI,97), Nov. 1997.
[7]. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. SIGMOD'93.
[8]. R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98.
[9]. Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen. Efficient Discovery of Functional and Approximate Dependencies Using Partitions. ICDE'98.
[10]. J. W. Shavlik and T. G. Dietterich. Readings in Machine Learning.Morgan Kaufmann, 1990.
[11]. P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining.Addison Wesley, 2005.
[12]. S. M. Weiss and C. A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.Morgan Kaufman, 1991.
[13]. Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,"On the Class Imbalance Problem" Fourth International Conference on Natural Computation, 2008.
[14]. David P. Williams, Member, Vincent Myers, and Miranda Schatten Silvious, "Mine Classification With Imbalanced Data", IEEE Geosciences And Remote Sensing Letters, Vol. 6, No.3, July 2009.
[15]. Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano "A Comparative Study of Data Sampling and Cost Sensitive Learning" , IEEE International Conference on Data Mining Workshops.15-19Dec.2008.

DOI 10.17148/IJARCCE.2018.71125