An Approach to Implement Data Mining in Service Oriented Methodology towards Amelioration of Society

Tamasree Biswas
Department of Information technology
Narula Institute of Technology, Kolkata, West Bengal, India

Mousumi Saha
Department of Computer Science And Engineering
Narula Institute of Technology, Kolkata, West Bengal, India

Soumya Bhattacharyya
Department of Information technology
Narula Institute of Technology, Kolkata, West Bengal, India

Abstract- Data mining is the process of identifying valid and potential information from huge amount of data. In this age of Information Technology, Data mining has emerged as a technology that automates the pattern discovery process in large database. Today, this decision support system forms a multi-billion dollar industry. Speed and efficiency only cannot decide the success of a business organization. Daily, a huge amount of data is generated which is to be stored and mined for decision making. So data mining is helpful for all types of organization. In this paper we will be mainly focusing on the real-time application of data mining. This is a survey paper covering the previous works on data mining applications ranging from educational field to medicine, sales/marketing, banking/finance etc.

Keywords – Knowledge Discovery in Database (KDD), Educational data mining (EDM), Medical data mining, Customer Relationship Management (CRM).

I. INTRODUCTION

With the advent of technology, the collection of data is increasing enormously, so is the size of the database. This huge amount of data is to be stored in the datawarehouse so that we can derive information that will be useful to derive patterns and trends in the data to formulate rules. Once these rules are formulated, the user can use it to support, review and examine decisions in some related business or scientific areas. Thus, it creates opportunities to interact with databases and data warehouse.

Data mining, most commonly is defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from the data [2]. In certain cases data mining and knowledge discovery in database (KDD) is used synonymously. We can also say data mining and KDD is a new interdisciplinary field which merges the idea from machine learning, parallel computing, statistics and databases. The process of inferring rules from data takes place through a series of steps. Following is a brief description of the steps in KDD.

KDD Stages:

Step 1: Integration

Firstly the data is collected from different sources and stored in a single point. This process of collecting data from heterogeneous sources and storing at a homogeneous place is known as the process of Integration.

Step 2: Selection

All the data that are collected in Stage 1 will not be required for mining process. So in this stage only those data are selected that are relevant in some aspects. So, choosing data from the entire list of available data is known as the Selection process.

Step 3: Cleaning

The collected data may contain missing values, error, noisy or inconsistent data. So this unnecessary information must be removed before proceeding further. This task is achieved in this stage. This stage reconfigures the data to ensure a consistent format, as there is a possibility of inconsistent formats. Cleaning stage is thus also known as the preprocessing stage.

Step 4: Transformation

In this stage data is transformed in order to be suitable for the task of data mining. Smoothing, aggregation, normalization, etc. Are some of the techniques used to accomplish this. The primary objective of this stage is to make the data usable and navigable.

Step 5: Mining

In this stage data mining techniques are used to discover meaningful, new correlation patterns and trends. Some of the commonly used data mining techniques are association analysis, clustering, classification rules, etc.

Step 6: Interpretation and Evaluation

The patterns that are obtained in the mining stage are converted into knowledge to support decision making. In other words we can say this stage involves visualization, transformation, removing redundant patterns, etc.

Step 7: Data Visualization

Data visualization helps the analyst to gain a deeper, more intuitive understanding of the data. So it helps the user to make use of data mining in taking better decision.

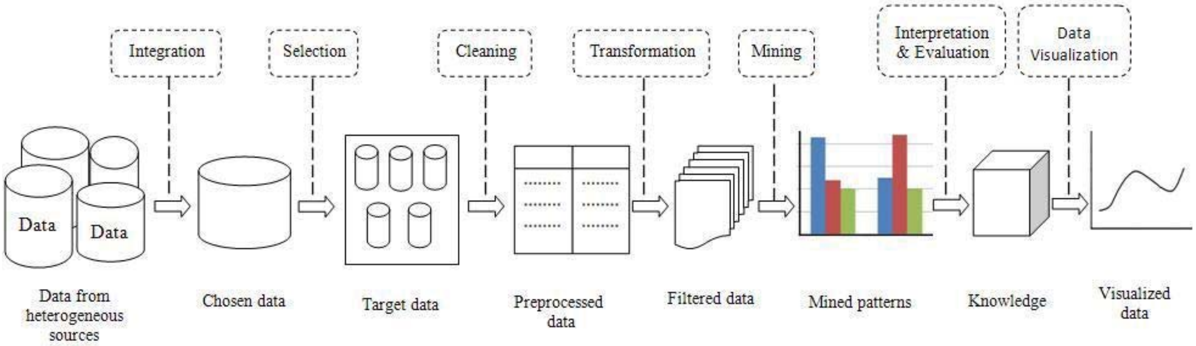

Figure 1 : Stages of knowledge Discovery in Database (KDD)

We can conclude that though in many cases data mining and KDD is used synonymously but data mining is only one of the stages of KDD.

In section II we will be discussing about the different applications of data mining in various fields. The reason behind this is to give the readers a thorough knowledge about the works that have already been in this field so that new ideas open up for future. Future scope will be discussed in section III followed by the conclusion in section IV and bibliography in section V.

II. DATA MINING APPLICATIONS

After gaining an insight into the data mining process now we can discuss some of the applications or past work in this field. In the subsequent section we will be discussing different works in various sectors under the sub sections.

A. Educational Data Mining–

In higher educational institution there is always an overwhelming pressure to provide up-to-date information for their accountability on student success. In this respect data mining could be used to solve problems such as student retention and attrition, personal recommender system (PRS) and analysis of course management system data. Baker and Yacef et. al [3] defined EDM as a completely new field emphasizing on all types of data that are retrieved from an educational institution used for the betterment of the students and improving the teaching learning process. In the present scenario all the educational institutions aim at upgrading the quality of the students, in which EDM can be a powerful tool. In the following section a literature survey on EDM is provided.

Student Retention and Attrition
Luan et. al [15] used data mining to predict which student would drop out and return later. By applying classification and regression tree technique Luan made an effort to identify the students who are likely to return in future. His work was effective in EDM as it suggested ways to improve student retention rate by applying both quantitative and qualitative research techniques.

In another research by Vandamme et. al [25] suggested to categorize the students in three groups, viz. low risk, medium risk and high risk from the beginning of the academic session. By using data mining techniques like neural network, random forest and decision tree it was observed that the students in high-risk group tend to fail and drop-out. This research provide a way out to identify the weak student and take effective measure to improve them from the very beginning and thus control the attrition rate. Other related works were done by Lin et al [14] by predicting which type of student would be benefitted from student retention program by using machine learning algorithms. Furthermore, Chacon et al [4] worked on practical environment to improve the student retention effort utilized at Bowie State University.

Course Management System
Romero et al [20] established the application of data mining techniques to Moodle usage data. By doing so, trend of student online behavior could be recorded. This data could be analyzed to identify the student strength and weakness in every area.

Another work by Wang and Lio [26] suggests that instead of undergoing a static course content the student can go through the course at his own pace. This could be achieved by adapting learning exercises based on student progress through a course.

Personal Recommender System (PRS)
The main aim of the PRS is to provide services, tools and artifacts so that the system could recognize the need of the student. While applying personal recommender system in educational context, two goals must be achieved. Firstly, it should aim at fulfilling the goal of the learner and secondly, the faculty members should be capable of controlling the recommendations to the student. A research on personal recommender system or PRS by Huang, Chen and Cheng [10] suggests student specific recommendations by learning student behaviour in an online course. This could be achieved by frequent item set mining to provide user specific recommendation for improved learning efficiency. Another work by Su, Tseng, Lin and Chen [23] suggested highly personalized, dynamic and fast learning recommendation to mobile users. This study focuses on how data mining could be beneficiary to mobile to mobile learning. Ecommerce is a domain where PRS has been widely used. Example, Amazon.com uses the recommendation system to customize the browsing history and display those products on the top which has a high chance of being purchased.

B. Medical Data Mining–

Data mining techniques is widely used in medical data to trace patterns or trends that supports knowledge discovery or decision making process. Clustering, classification, association rule mining and regression are some commonly used techniques for this purpose. Mahmud Khan et. al [21] suggested classification of x-ray images for diagnosis of lung cancer. In his work he used decision tree algorithm for image analysis. Wang et al [22] used the same method to classify mammography reports as normal or abnormal case. Xing et al [28] combined Support Vector Machine (SVM), Artificial Neural Network (ANN) and decision tree for predicting the survival of Coronary Heart Disease (CHD) patient. Chen et al. [1] suggested the side effects of drug exposure by mining pregnancy data. By using Smart Rule, a mining technique for association rule generation, he considered two groups, one exposed to active drugs and the other including preterm cases. The generated rules were used to build hierarchical model for making interesting deductions. Bethel et al. [7] used a tool named "Clinical Trial Assignment Expert System" to develop an association rule learner based on the past records of breast cancer patients. Froelich and Wakulicz-Deja [27], using adaptive fuzzy cognitive maps, mined drugs and health effects for improving decision support and planning in healthcare. Tu et al. [17] used decision tree algorithm C4.5 and Naive Bayes algorithm to propose an intelligent medical support system for diagnosis of heart diseases. For setting questions to extract knowledge from medical data, a language known as knowledge Discover Question Language was introduced by Hogl [19]. Saraee et al. [16] proposed a method to determine mortality rate in children due to accidents by using decision tree generated by CART algorithm.

B. Banking and Finance–

A banking system contains an enormous amount of data both historical and operational. From the collected data it has to take vital decisions including default decisions, relationship start up, credit decision, investment decision, etc. In order to take such decisions data mining could be used efficiently to deduce inference depending on past history or patterns. Customer Relationship Management (CRM) is another aspect where data mining is used in any organization. In this age of cut-throat competition retention of customer is the toughest challenge faced. So, if the needs and expectations of the customer is satisfied then it could be achieved. Data mining is used to mine the previous transaction history of the customer to take decision about the expectation of the customer before-hand.

Some important areas where data mining could be used in banking and finance sector includes risk management, marketing, defaulter detection, investment banking, money laundering detection, fraud detection, etc.. Costa et al. [8] suggested that future defaults can be predicted by using the historical default patterns. Kazi and Ahmed [12] suggested a probable way of identifying the defaulter to help the credit managers decision before granting loan. Ingle and Meshram [11] proposed the use of K-means clustering algorithm for choosing the best investments based on the profile of the customer. Naeini et al. [18] proposed the use of prediction techniques like linear regression and neural networks for stock price prediction. This type of prediction for historic values helps in better investment decision. Now-a-days, in such a competitive market, retaining the old customers and making new is the most challenging task for the banks. Chopra et al. [6] suggested to use data mining techniques to learn the expectation of every customer. If the expectations are known beforehand customers can be retained easily. Sundari and Thangadurai [24] suggested the analysis of sequential patterns for investigating changing customer preferences for pro-active approach towards the customer. Chen et al. [5] proposed to identify the risk factors while lending by Credit Approval authorities. This could be achieved by analyzing the data based on nationality, repayment capacity, etc.

Dheepa and Dhanapal [9] proposed that the probability of credit card faults can be detected by mining the probability density of credit card users previous behaviour.

III. CONCLUSIOn

From the above study we can conclude that data mining is used in almost all aspects of human life. Whether a user wants to buy a product online or to choose the reference books, knowledge discovery from previous history is needed. Similarly, for a business organization data mining plays a crucial role in holding the success. If data mining is used properly, it proves to be beneficial for humanity.

Due to the presence of practically relevant problems and scalable solutions researchers have always been attracted towards this field. Business intelligence, customer relationship management (CRM) and e-commerce all hinge on data mining. In this paper we have put forward some of the works in the field of data mining. This review work will be helpful to the researchers for learning the progress in the field of data mining in the past. It also opens up new areas were data mining can be implemented for the betterment of humanity.

In our next work we will be focusing on some of the algorithms of data mining that will be used to solve the problem of frequent call drops in mobile phone network. This has become a serious issue in the telecom sector. Whenever there is a change in the location of the mobile user, the call drops. If data mining could be used efficiently then all would be benefitted.

REFERENCES

[1] Adepele Olukunle and Sylvanus Ehikioya, (n.d). A Fast Algorithm for Mining Association Rules in Medical Image Data. IEEE. p1-7.

[2] Arun K Pujari “Data Mining Techniques”, Edition 2001.

[3] Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1.

[4] Chacon, F., Spicer, D., & Valbuena, A. (2012). Analytics in Support of Student Retention and Success (Research Bulletin 3, 2012ed.). Louisville, CO: Educause Center for Applied Research.

[5] Chen, S.C. and M.Y. Huang, 2011. Constructing credit auditing and control and management model with data mining technique. Expert Syst. Applic., 38: 5359-5365. DOI: 10.1016/j.eswa.2010.10.020

[6] Chopra, B., V. Bhambri and B. Krishnan, 2011. Implementation of data mining techniques for strategic CRM issues. Int. J. Comput. Technol. Appli., 2: 879-883.

[7] Cindy L. Bethel and Lawrence O. Hall and Dmitry Goldgof (n.d). Mining for Implications in Medical Data. IEEE. p1-4.

[8] Costa, G., F. Folino, A. Locane, G. Manco and R. Ortale, 2007. Data mining for effective risk analysis in a bank intelligence scenario. Preccedings of the 23rd International Conference on Data Engineering Workshop, Apr. 17-20, IEEE Xplore Press, Istanbul, pp: 904-911. DOI: 10.1109/ICDEW.2007.4401083 .

[9] Dheepa, V. and R. Dhanapal, 2009. Analysis of credit card fraud detection methods. Int. J. Recent Trends Eng., 2: 126-128.

[10] Huang, Y.-M., Chen, J.-N., & Cheng, S.-C. (2007). A Method of Cross-Level Frequent Pattern Mining for Web-Based Instruction. *Educational Technology & Society*, 10(3), 305-319.

[11] Ingle, D.R. and B.B. Meshram, 2012. E-Investment banking: NextGen investment. Int. J. Advanced Res. Comput. Eng. Technol.,

[12] Kazi, I.M. and. Q.B. Ahmed, 2012. Use of data mining in banking. Int. J. Eng. Res. Appli., 2: 738-742.

[13] Khac, N.A.L. S. Markos, M. O'Neill, A. Brabazon and M. Kechadi, 2011. An investigation into data mining approaches for anti money laundering. Proceedings of the International Conference on Computer Engineering Applications, (EA' 11), Lacsit Press, Singapor, pp: 504-508.

[14] Lin, S.-H. (2012). Data mining for student retention management***.*** *J. Comput. Sci. Coll., 27* (4),92- 99.

[15] Luan, J. (2002). ***Data Mining and Knowledge Management in Higher Education- Potential Applications.*** Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada.

[16] Mohammad Saraee, George Koundourakis, Babis Theodoulidis. (n. d). EASYMINER: DATA MINING IN MEDICAL DATABASES. IEEE. p1-3.

[17] My Chau Tu AND Dongil Shin (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. IEEE. P1-5.

[18] Naeini, M.P., H. Taremian and H.B. Hashemi, 2010. Stock market value prediction using neural networks. Proccedings of the International Conference on Computer Information Systems and Industrial Management Applications, Oct. 8- 10, IEEE Xplore Press, Krackow, pp: 132-136. DOI: 10.1109/CISIM.2010.5643675

[19] Oliver Hogl, Michael Müller (2001). On Supporting Medical Quality with Intelligent Data Mining. IEEE. p-10.

[20] Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384. doi: 10.1016/j.compedu.2007.05.016

[21] Safwan Mahmud Khan Md. Rafiqul Islam Morshed U. (n.d). Medical Image Classification Using an Efficient Data Mining Technique. IEEE, p1-6. 1.

[22] Shuyan Wang Mingquan Zhou Guohua Geng (n.d). Application of Fuzzy Cluster Analysis for Medical Image Data Mining. IEEE. p1-6.

[23] Su, J.-m., Tseng, S.-s., Lin, H.-y., & Chen, C.-h. (2011). A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments. User Modeling and User - Adapted Interaction, 21(1-2), 5-49.doi: 10.1007/s11257-010-9094-0.

[24] Sundari, P. and K. Thangadurai, 2010. An empirical study on data mining applications. Global J. Comput. Sci. Technol., 10: 23-27.

[25] Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, 15(4), 405-419.

[26] Wang, Y.-h., & Liao, H.-C. (2011). Data mining for adaptive learning in a TESL learning system. *Expert Systems with Applications, 38*(6), 6480-6485. doi:10.1016/j.eswa.2010.11.098

[27] Wojciech Froelich, Alicja Wakulicz-Deja (2009). Mining Temporal Medical Data Using Adaptive Fuzzy Cognitive Maps. IEEE. P1-8.

[28] Yanwei Xing, Jie Wang and Zhihong Zhao (2007). Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease. IEEE. p1-5.