International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-1, Issue-5, November 2011

Data Mining and It's Approaches towards Higher Education Solutions

Tripti Arjariya*, Shiv Kumar, Rakesh Shrivastava, Dinesh Varshney

Abstract— The major objective of the research is advancement of knowledge and theoretical understanding of the relations among variables for the study and development using data mining in higher education system and its solution for Madhya Pradesh state. New knowledge takes three main forms: Exploratory research: which structures and identifies new problems, Constructive research: develops solutions to a problem, Empirical research: tests the feasibility of a solution using empirical evidence. As per the research methodology, one can have two distinct methods of research either primary or secondary.

This study is a survey type of research followed by the developmental study of data mining in higher education Madhya Pradesh state. The terms basic or fundamental indicate that, through theory generation, basic research provides the foundation for further, sometimes-applied research. As there is no guarantee of short-term practical gain, researchers may find it difficult to obtain funding for basic research. In this research we come to know that how the data mining approaches and issues are helpful for the development and the solutions of higher education in Madhya Pradesh state.

Index Terms— Knowledge development, data mining approaches and issues, higher education system.

I. INTRODUCTION

The hardware and software approaches make changes in technological applications. This makes society more scientific day to day due to generation to generation changes in different areas. In the teaching learning process intervention Information Communication Technology(ICT) able to solve the problem of teacher, learners and administrators with a systematic way. In India and abroad no country is able to solve their basic educational problems even growth of literacy rate from elementary to higher education system. The developed and developing countries worldwide trying to focus on development of educational system of under developed countries. The education systems are also reformed by the interference of International organization like UNESCO[2]. In higher education system it is a great challenge to take advantages of ICT applications in general and find out Root causes of problems, prospects of ennoblement of learners. Still below twenty percent of learners are able to get admission in higher education in Indian context. The new approaches of IT application hindering the factors facilitating or as an obstructed for higher education. The totalities of IT application in the form of data are mined through the different areas of higher education studies.

HOW DO WE CATEGORIZE DATA MINING SYSTEMS

There are many data mining [DM] systems available or being developed [7]. Some are specialized systems dedicated to a given data source or are confined to limited DM functionalities, other are more versatile and comprehensive. DM systems can be categorized according to criteria of classification as following [3]:

A. Classification according to the type of data source mined which categorizes DM systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web etc.

B. Classification according to the data model drawn on that categorizes DM systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional etc.

C. Classification according to the kind of knowledge discovered[5] which categorizes DM systems based on the kind of knowledge discovered or DM functionalities such as characterization, discrimination, association, classification, clustering etc. Some systems tend to be comprehensive systems offering several DM functionalities together.

D. Classification according to mining techniques used employ and provide different techniques, this classification categorizes DM systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, data base or data warehouse-oriented etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems [6].

II. WHAT ARE THE ISSUES IN DATA MINING

DM algorithms embody techniques [3] [7] that have existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While DM is still in its infancy, it is becoming a trend and ubiquitous. Before DM develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed some of these issues are addressed below.

Manuscript received October 22, 2011.

* Corresponding Author

Tripti Arjariya, Department of Computer Science and Engineering, Madhya Pradesh Bhoj (Open) University, Bhopal (M.P.)-462021, India. (E-mail: tripti.arjariya@gmail.com).

Shiv Kumar, Associate Professor, Department of Information Technology, Technocrats Institute of Technology, Bhopal (M.P.)-462021, India. (E-mail: shivksahu@rediffmail.com).

Dr. Rakesh Shrivastava, Professor, Department of Higher Education, Govt. of Madhya Pradesh, Bhopal (M.P.)-462021, India. (E-mail: *rakesh_geol@yahoo.co.in*)

Dr. Dinesh Varshney, Professor, Multimedia Regional Center, Madhya Pradesh Bhoj (Open) University, Indore (M.P.) - 452001, India. (E-mail: *vdinesh33@rediffmail.com*).

A. Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behaviour understanding, correlating personal data with other information etc. large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. DM could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of DM.

B. User interface issues: The knowledge discovered by DM tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of DM results, as well as helps users better understand their needs. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in. The major issues related to user interfaces and visualization is "screen real-estate", information rendering, and interaction. Interactivity with the data and DM results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge [6] from different angles and at different conceptual levels.

C. Mining methodology issues: These issues pertain to the DM approaches applied and their limitations. Topics such as versatility of the mining approaches, diversity of data available, dimensionality of the domain, broad analysis needs (when known), assessment of the knowledge discovered, exploitation of background knowledge and metadata[6], control and handling of noise in data etc. are all examples that can dictate mining methodology choices. Most algorithms assume data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, which may complicate the analysis process and in many cases compromise the accuracy of the results. As a consequence, data pre-processing (data cleaning and transformation) becomes vital and the most important phase in the knowledge discovery process. DM techniques should be able to handle noise in data or incomplete information.

More than the size of data, the size of the search space is even more decisive for DM techniques. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality*. This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

D. Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation[1]. However, these methods were often not designed for the very large data sets DM is dealing today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the DM methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for DM. Linear algorithms are usually the norm. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating DM results when new data becomes available without having to re-analyze the complete dataset.

E. Data source issues: There are various issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of DM is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories.

III. PROPOSED ALGORITHM

Algorithm -1:

```
Input: Data set R, Attribute set Ai
Output: data set R'
R' -> R
For I=1 to n do
 Max (Ai) = the deepest node in the attribute set Ai
If Max(Ai). Distance_to_max<Ii
        Newnode=node.root_path_array[Ii-node.distance_t
        -_max]
 Else
Newnode=max(Ai)
 Endif
 Replace node with new node
Endfor
Remove duplication from R'
End
```

Algorithm -2:

Input : Primitive rules set R
Output Generalized rules set R'

```
R' <- 0
N= | R |
    For I=0 to N-1 do
      r <- ri
     M <- |r|
   For j=0 to M-1 do
            If ri inconsistent with rule rn E  then
            Restore the dropped condition aj
        Endif

         End  for
         Included in rule r
If rule r is not logically include in a rule r' E MRULE then
   MRULE <- r U MRULE
Endif
End
```

R: Data set which is a any college web site because Ontology is used for specific domain.
R': is the output
Ai is the attribute set
Max (Ai) is the function which finds the deepest node in the attribute set Ai.

IV. RESULTS AND DISCUSSIONS

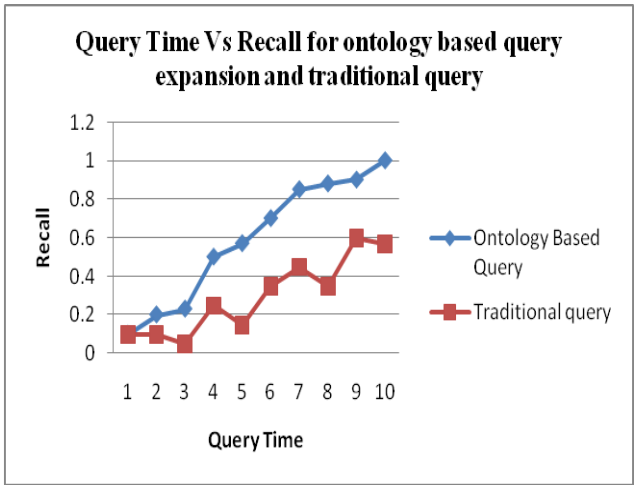

Figure 3 Recall ratios of query methods

User input the query words in ontology as expansion words and its performance can be showed through precision and recall ratios that are calculated from experimental results. Through 10 times different information requests, we compute recall and precision ratios and make comparison with traditional query method.

V. CONCLUSIONS

The DM is directly associated with use of technology for accessing data and to give result as required in a desired way. In Indian contest though computer literacy among the users are very low but its applicability in different sectors of the society is highly demandable day to day. With specific to education sector it has great demand both teaching and learning prospects. The management aspects are highly interference by the Information Communication Technology and DM areas. Higher Education system in India, now a day's totally depended on DM majors. The demand and problem solving abilities within the framework of logical argument and accuracy of result need to explore through research and development procedure. Efforts are made by Government, NGOs and Independent bodies trying to make social problems solve able easily through the DM. Through the algorithm and the experimental results we can conclude that the data mining techniques are very much useful in the development and finding out the solutions of higher education in Madhya Pradesh state.

REFERENCES

1. Cave, M., Kogan, M. and Hanney, S. (1990), "The scope and effects of performance measurement in British higher education, in F. J. R. C." Dochy, M. S. R. Segers and W. H. F. W. Wijnen (Eds.), "Management Information and Performance Indicators in Higher Education,"Van Gorcum and Comp, 48–49.
2. Fielden, J., and Abercromby, K. (2000), "UNESCO Higher Education Indicators Study: Accountability and International Co-operation in the Renewal of Higher Education", Georgia Professional Standards. UNESCO, Paris.
3. Han, J. and Kamber, M. (2001), "Data Mining: Concepts and Techniques", Simon Fraser University, Organ Kaufmann.
4. Johnstone, J.N. (1976), "Indicators of the Performance of Educational Systems. UNESCO", International Institute for Educational Planning, Paris.
5. Luan, J. (2002a), "Data mining and knowledge management in higher education – potential applications", In Proceedings of AIR Forum, Toronto, Canada.
6. Luan, J. (2002b), "Data Mining Application in Higher Education", SPSS Executive Report.