Analysis and Comparison of Data Mining Tools Using Case Study of Library Management System

Manu Bansal and Mandeep Kaur

Abstract—**The term data mining has been the oldest yet one of the interesting buzzwords. Many organizations often underutilize their already existing databases. There is a need to mine information and interesting patterns from these databases. The focus of the current research is to apply data mining on a library management system. Data mining is usually done on a data warehouse or a data mart. It incurs various cost factors like software, hardware, maintenance and experts. The objective here is to study how the real-time data stored in database can be turned informative without setting up a separate data warehouse. The main emphasize is on understanding the problem perspective, competing objectives and constraints and generating a model for information extraction from the real-time library database using ARM (Association Rule Mining) mining technique. As SQL (Structured Query Language) can also be used for mining data instead of using specialized data mining algorithm [1], the study also compares SQL based mining with ARM. The results shows that association rule mining performs better than SQL based mining as type of pattern to be extracted can be controlled much effectively in ARM as compared to SQL because of the parameters (support and count) used in the data mining algorithm. Algorithms are implemented using SQL and MATLAB (Matrix Laboratory) Tool - ARMADA.**

Index Terms—**ARM, ARMADA, data warehouse, data mart.**

I. Introduction

Data mining is becoming an increasingly important tool for transforming data into information [2]. This information is vital for marketing, analysis and decision making, study trends and demands. There are basically two most important reasons that data mining has attracted a great deal of attention in the recent years. Firstly, the capability to collect and store the huge amount of data is rapidly increasing day by day. The second but the more important reason is the need to turn such data into useful information and knowledge.

The principal purpose of the decision support system for libraries is to provide information regarding patterns generated according to the usage of books, periodicals and electronic services. Indications are there depicting the variance in patterns used by library among various disciplines. The library administration could use such information to adjust and/or justify purchases and licenses, the subscription to serials, contracts with electronic vendor companies and allocation of funds etc. A data warehouse is the core of any decision support system. Typically, the data warehouse is separate from the organization's operational databases. KSUL data warehouse [3] included tools for extracting data from multiple operational databases and external sources; for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse. Data in the warehouse is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of front-end tools: query tools, report writers, analysis tools, and data mining tools. Attempts were also made to apply various architectural alternatives for coupling data mining with relational database systems [3]. Also, a mining computation expressed in SQL is compared to a specialized implementation of the same mining operation i.e. using different alternatives architectures within SQL itself. The vendor database management software are becoming aware of integrating data mining capabilities into database engines (e.g. IBM's DB2 Database and Intelligent Miner or NCR's Teradata Database and Teradata Warehouse miner)[4]. In education sector, ARM is also applied on the software of examination paper evaluation system and provides valuable analysis of the evaluation system [5].ARM is also studied and implemented in distributed database systems [6]. The current research work is to apply data mining on a single database i.e. college library management system, rather than on multiple databases (needed in warehousing), to extract hidden patterns in a cost effective manner so that the hidden information can be discovered and turned into effective usage for future decision making. Association rule mining is applied to extract interesting and hidden patterns from the library management system like most frequently issued book, book issued not more than five times, popular author in a particular field etc.

Data pre-processing is a vital phase in the study, which moulds the real time data base for mining information.

II. Association Rule

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases Given a set of transactions, where each transaction is a set of items, an association rule [7] is an expression X!Y, where X and Y are sets of items. The intuitive meaning of such a rule is that the transactions that contain the items in X tend to also contain the items in Y. An example of such a rule might be that 30% of transactions that contain beer also contain diapers; 2% of all transactions contain both these items". Here 30% is called

Manuscript received September 24, 2012; revised November 30, 2012. This work was done as part of Master of Engineering dissertation.

M. Bansal is working with Department of Information Technology, Shaheed Udham Singh College of Engineering & Technology, Tangori, Punjab , India (e-mail: mrmanubansal@gmail.com).

M. Kaur is with University Institute of Engineering and Technology, Panjab University, Chandigarh, Punjab, India (e-mail: mandeep@pu.ac.in).

the confidence of the rule, and 2% the support of the rule. The problem of mining association rules is to find all rules that satisfy a user-specified minimum support and minimum confidence.

The association rule mining problem can be decomposed into two sub-problems [7]:

- Find all combinations of items, called frequent item-sets, whose support is greater than minimum support.
- Use the frequent item sets to generate the desired rules.

Apriori algorithm is one of the most influential algorithms to mine the frequent item sets of Boolean association rules [8][9]. ARMADA (a MATLAB tool) is used to apply Apriori algorithm for association rule mining on Library database.

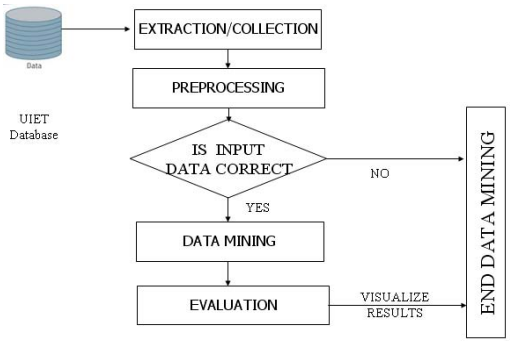

Fig. 1. Proposed model of data mining for UIET library database

III. Methodology

The various considerations are to be recognized and worked upon before applying the mining process on the real-time database.

This research is done for the information extraction from real-time database of library, UIET (University Institute of Engineering and Technology), Panjab University, Chandigarh (India).. There are 2000 students in the department. The database for UIET library composed of around 15000 transactions containing information related to books issued by students of various departments. Currently, around 2000 books as of total is available for students in the library. The work plan is proposed for library that contains various test runs for different cases. These questions include the setting up various queries for information extraction from library database with characteristics (such as book name, edition, publisher name, vendor, etc). This forms the data set for our study, and the purpose of the research is to find interesting relationships among these data and generate more insights from the relationships. It is postulated that data mining, and more specifically, the association rule mining method would be applicable to this type of analysis as compared to other mining techniques. Association rule mining does not impose any assumptions on the relationships among the data, and therefore will not constrain the results to a certain form.

The entire process can be broken down into five distinct phases:

1) Extraction of data from the database, whether or not to carry out data mining on a given data set.
2) Data pre-processing (or preparation), readying the data for analysis.
3) Data verification and data cross-checking.
4) Data Mining is applied on the processed data.
5) Analysis of results which is largely interpreting different results carried out by specifying different mining criteria (minimum support and confidence) .It can be greatly assisted using automated means, such as graphical representations of the results.

This process is illustrated using a flow chart in Fig.1. This will serve as preparation for our implementation.

IV. Preprocessing Data

The pre-processing of the data is a very important phase in mining .The major issues and problems related to data mining on relational database are studied.

Pre-processing influences the results of the mining process very deeply. It involves noise removal, data reduction, etc. Some of the pre-processing steps applied on the Library database system depending on the type of information to be mined are:

1) Removal of noisy data. E.g. time of book issue displays value as "NULL".
2) Conversion of alphanumeric student ID with numeric ones.
3) The unique book-ids given to the books comprising of same title, author and edition are settled with a particular id for all of the copies for same book details (title, author, and edition) in order to have accurate dataset to be mined up.
4) Reassigning edition wise i.e. the year of publishing for book is reassigned. For e.g. instead of having value as "2007", value originally assigned as "07" or "2k7", which is needed to be changed.
5) Data reduction is done in order to remove the unwanted data. E.g.: the information like "place of publishing" and "date of publishing" with "NULL" values is not included in the final data set.
6) Data partitioning is performed i.e. the data is partitioned branch wise with each branch with different years are merged and coded with unique branch id.

The pre-processed data is stored temporarily on a separate file for mining to be performed.

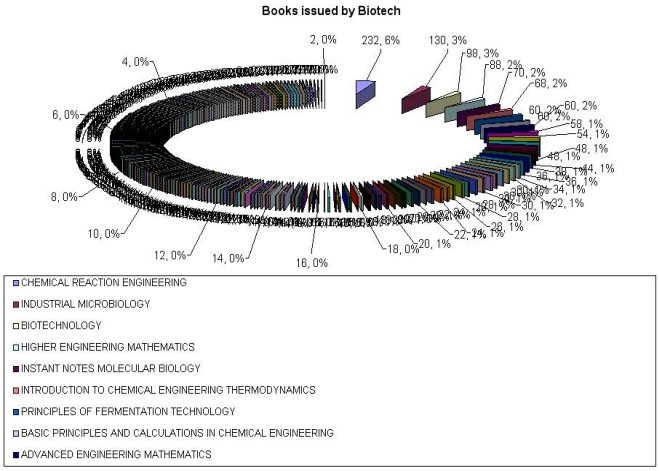

Fig. 2. Books issued by biotech (SQL results)

V. Results for SQL and ARMADA on Real Time Dataset "UIET Library Data"

We implemented the ARM (Association Rule Mining) technique for different categories under same real –time database (UIET Library Data). ARM (Association Rule Mining) is implemented for each branch with same data set using both the tools- Armada and SQL. The results are shown along with data set below representing the useful mined data from library database.

The complexity in the data set for Fig. 2 is lagging in depicting a clear view of extracted information in detail for the branch "Biotech", though we analyzed that the most commonly issued book for this branch is "Chemical Reaction Engineering" with a count of "232" that forms 6% of the overall books issued by this branch.

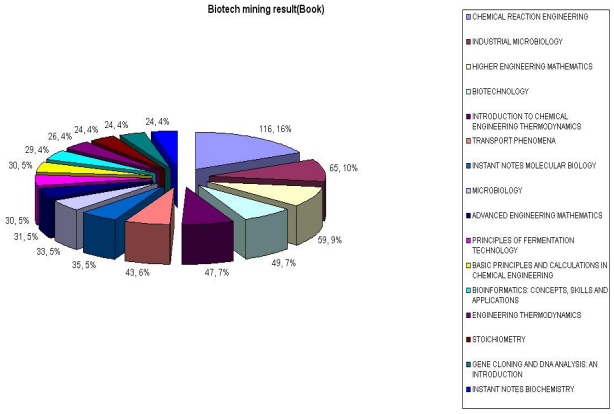

Fig. 3. Books issued by Biotech (ARMADA results)

In Fig. 3, the same data set of Biotech branch (used in SQL) is used for ARMADA. The complexity of data set is reduced and a clear view of extracted information in detail is depicted in Fig: 3 as compared to Fig. 2. We analyzed that the non-frequently issued books have not been entertained and have been excluded from the extracted information set. The minimum issued books have percentage of "0%" for Fig. 2 as compared to "4%" for Fig. 3 respectively, for same data set.

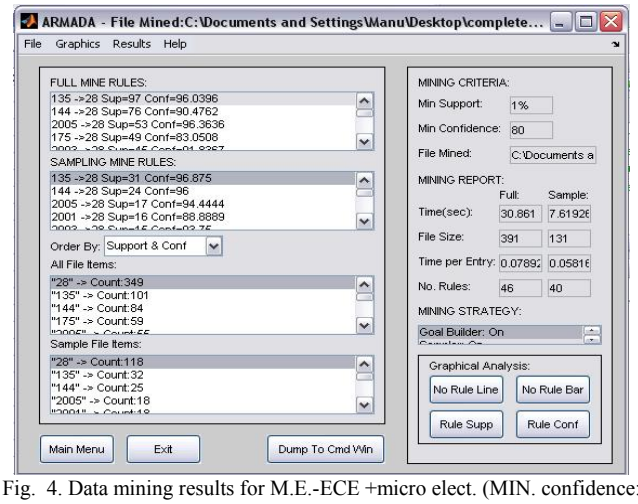

Fig. 4. Data mining results for M.E.-ECE +micro elect. (MIN. confidence: 80%, min support: 1%)

VI. ARAMADA Results (Stream Wise): Full Mining V/S Sampling

The following results are obtained after mining the real-time data as a "Full mine" as well as "Sampling" and the same is performed on the data set partitioned according to the branches.

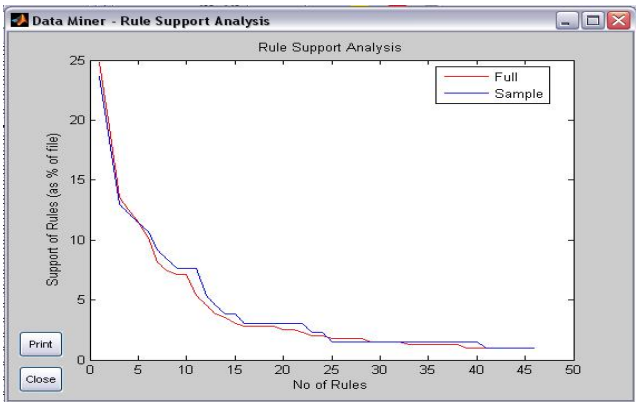

Fig. 5. Rule support analysis (M.E.-ECE +micro elect.) (Min. confidence: 80%, min support: 1%)

The above results for M.E-Electronics (ECE) and M.E-Micro-Electronics branches give us the representation of results as "Full-Mining" and "Sampling" with two critical factors: "Rule Support" and "Rule Confidence". The minimum confidence and minimum support for each branch has been opted as 80% and 1% respectively.

VII. Interpretation of Results with Same Minimum Support and Variations in Minimum Confidence

In Fig. 6, the same data set of Biotechnology branch (used in SQL) is used for ARMADA. The resultant is obtained after applying the data set with same minimum support i.e. 1% but with variance in minimum confidence from 50% to 80%. The results analyzed so far clearly states that less the minimum confidence, more will be the mining execution time, for the same data set.

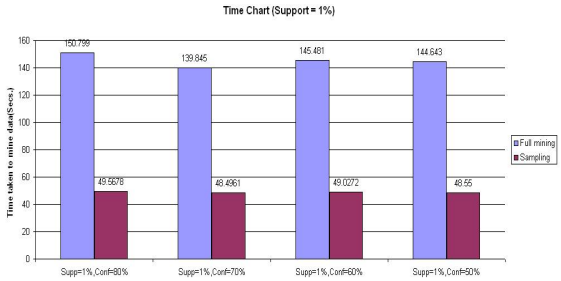

Fig. 6. Execution (Time) result for branch "biotechnology" with same support (1%) and variance in the % of confidence

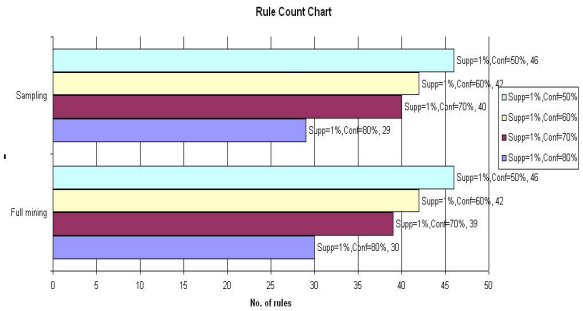

Fig. 7. Rules generation result for branch "biotechnology" with same support (1%) and variance in the % of confidence

In Fig. 7, the same data set of Biotechnology branch (used in SQL) is used for ARMADA. The resultant is obtained after

applying the data set with same minimum support i.e. 1% but with variance in minimum confidence from 50% to 80%. The results analyzed so far clearly states that less the minimum confidence, more will be the rule generated, for the same data set.

V. Conclusion

From the results analysis and model evaluation, it is concluded that data mining methods are applicable for the educational research according to the existing structure/model of the library. The collective information for various departments is available at same point. Also, better standardization of resources (e.g. books, journals, magazines) according to the mined information can be setup helping in better relativeness with the knowledgeable information. In this study, the two important tasks have been performed that forms the part of the solution: firstly, by pre-processing and secondly, by applying association rule mining for same dataset with different tools: SQL and ARMADA. With the refined information, association rule mining can be effectively applied to find the relationships between different factors and free riding behaviors. Combination of objective and subjective methods can generate quality results as well as desired rules.

VI. Future Work

The work presented in the thesis can be extended for multi-level association rule mining and multi-dimensional association rules.

In our study, data refinement, aggregation and accuracy were used as key roles to evaluate and validate the rule mining results. Emphasis was on the study of better result evaluation with more preprocessing applied on real-time data set, mainly because usually, real time data was not in the required form to be mined. If an expert can be brought to interactively put his opinion during the process, the analysis can be conducted on a larger number of rules. Only numeric or categorical data was dealt in during the study. An advantage of data mining technique, compared with statistics is that data mining can process a wider range of data type.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann Publishers, 2006.

[2] S. I. Zutshi, "Logic based Pattern Discovery," *IEEE transaction on Knowledge and Data Engineering*, vol. 22, no. 6, June 2010.

[3] M. Z. Bleyberg, D. Zhu, K. Cole, D. Bates, and W. Zhan, "Developing an integrated library decision support data warehouse," Kansas State University, Manhattan, 1999.

[4] S. Nestorov and N. Jukie, "Ad-hoc Association Rule mining within the Datawarehouse," in *Proceedings of 36th Hawaii International Conference on System Sciences*.

[5] X. Xue, C. Yao, and W. Y. En, "Study on Mining Theories of Association Rules and Its applications," *International Conference on Innovative Computing and Communication 2010 and Asia Pacific Conference on IT and Ocean Engineering*, Macao,2010, pp. 94-97.

[6] Z. Li and M. Xu, "Research on Association Rules in Distributed Database system," in *Proc. 2nd International Asia Conference on Informatics in Control, Automation and Robotics*, vol. 3,Wuhan,China, 2010, pp. 216-219 .

[7] F. Glover. "Improve Linear Programming Models for Discriminant Analysis," *Decision Sciences*, vol. 21, pp. 771-785, 1990.

[8] Y. Liu, "Study on Application of Apriori Algorithm in Data Mining," *Second International Conference on Computer Modeling and Simulation*, 2010.

[9] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *in Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data,* Washington, D.C.,1993, pp. 207-216.

M. Bansal was born in Panchkula district, Haryana in India in the year 1983. He received his B.Tech degree in Computer Science and Engineering from Shaheed Udham Singh College of Engineering & Technology, Tangori, Punjab , India in 2007 and received his M.E degree in Information Technology from Panjab University, Chandigarh, Punjab, in the year 2010. He is currently working as Assistant Professor with Department of Information Technology, Shaheed Udham Singh College of Engineering & Technology, Tangori, Punjab , India.

M. Kaur was born in Pune, India in the year 1977. She received her B.Tech degree in Computer Science and Engineering from BCET, Punjab Technical University, Punjab (India) in 1999 and received her Master's in Engineering (Information Technology) from PEC, Panjab University, Chandigarh (India) in 2004. She is currently working as Assistant Professor with University Institute of Engineering and Technology, Panjab University, Chandigarh, India.