Evaluate Various Techniques of Data Warehouse and Data Mining with Web Based Tool

M. Aarthi,
MCA., M.Phil., M.Tech., [1]
Assistant Professor, Department of Computer Science,
PRIST University, Thanjavur.

J. Priyadharshini,
MCA., [2]M.Phil– Research Scholar,
Department of Computer Science,
PRIST University, Thanjavur.

***Abstract*:-** All enterprise has a crucial role to play proficiently and productively to maintain its survival in the market and increase its profitability shares. This challenge becomes more complicated with advancement in information technology along with increasing volume and complexity of information. Currently, success of an enterprise is not just the result of efforts by resources but also depends upon its ability to mine the data from the stored information. Data warehousing is a compilation of decision making procedure to integrate and manage the large variant data efficiently and scientifically. Data mining shores up organizations, scrutinize their data more effectively and proficiently to achieve valuable information, that can reward an intelligent and strategic decision making. Data mining has several techniques and maths algorithms which are used to mine large data to increase the organization performance and strategic decision-making.

Clustering is a powerful and widely accepted data mining method used to segregate the large data sets into group of similar objects and provides to the end user a sophisticated view of database. This study discusses the basic concept of clustering; its meaning and applications, especially in business for division and selection of target market. This technique is useful in marketing or sales side and, for example, sends a promotion to the right target for that product or service.

Association is a known data mining techniques. A pattern is inferred based on an affiliation between matter of same business transaction. It is also referred as relation technique. Large enterprises depend on this technique to research customer's buying preferences. For instance, to track people's buying behavior, retailers might categorize that a customer always buy sambar onion when they buy dal, and therefore suggest that the next time that they buy dal they might also want to buy onion.

Classification – it is one of the data mining concept differs from the above in a way it is used on machine learning and makes use of techniques used in maths such as linear programming, decision trees, neural network. In classification, enterprises try to build tool that can learn how to classify the data items into groups. For instance, a company can define a classification in the application that "given all records of employees who offered to resign from the company, predict the number of individuals who are likely to resign from the company in future." Under such a scenario, the company can classify the records of employees into two groups that namely "separate" and "retain". It can use its data mining software to classify the employees into separate groups created earlier.

Fuzzy logic resembles human reasoning greatly in handling of imperfect information and can be used as a flexibility tool for soften the boundaries in classification that suits the real problems more efficiently. The present study discusses the meaning of fuzzy logic, its applications and different features.

A tool to be build to check data mining algorithms and algorithm behind the model, apply clustering method as a sample in tool to select the training data out of the large data base and reduce complexity and time while computing. K-nearest neighbor method can be used in many applications from general to specific to find the requested data out of huge data.

Decision trees – A decision tree is a structure that includes a root node, branches, and leaf nodes. Every one interior node signify a test on an attribute, each branch denotes the result of a test, and each leaf node represents a class label. The topmost node in the tree is the root node. Within the decision tree, we start with a simple question that has multiple answers. Each respond show the way to a further query to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

Outlier detection technique refers to observation of data items in the dataset which do not match an expected pattern or expected behaviour. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

Sequential Patterns technique helps to find out similar patterns or trends in transaction data for definite period.

*****

I. Introduction

Currently, the technological revolution in data capture, processing power, data transmission and storage capabilities are available to organizations to assimilate their databases into a central repository so that data can be managed effectively and systematically. Before the advent of data warehouses, operational databases were used to satisfy their functional requirements, like data processing, analysis and reporting, however informational needs were the secondary considerations. But with the advent of information technology and increased complexity of data, business houses started demanding an information tool to improve their decision making capabilities. Now the data becomes heterogeneous (mixture of text, symbolic, numeric, texture, image), huge (both in dimension and size), scattered and growing at a phenomenal rate. Data warehousing has advanced to meet these disputes without affecting operational processing. Warehouses optimize database query and reporting tools because of their ability to analyze data, often from disparate databases and in remarkable ways. Data warehousing technology helps the managers and decision makers to extract information quickly and easily to retrieve the patterns in data, hidden but useful facts and relationships between the data items. Data mining is a powerful technique for the knowledge discovery and extraction of predictive information from data warehouse to help an enterprise to catch and focus on the vital information while making decisions. There are huge varieties of data mining methods and algorithms for information extraction and prediction. These different technological aspects are discussed in the following sections.

Various techniques of data warehouse

Data warehouse is a repository of an organization's electronically stored data [Wikipedia, 2008]. A data warehouse is the consistent store of data which is made available to end users, so that they can understand and use in a business context [Gatziu, and Vavouras, 1999]. Data warehouse is used in the businesses to convert data into business intelligence and making management decisions, based on the facts and not on intuition. Analysts make use of warehouse to build strategic consideration, forecasting, competitive analysis, and targeted market research. Data warehouse is one of the steps on the long road towards the ultimate goal of accomplishing the objectives of a concern.

Data warehousing is defined as a process of centralized data management and retrieval [Palace, 1996]. It is expected to present the right information at the right place and at the right time with the right cost in order to support the right decision [Jarke and Yannis, 1997]. Data warehousing is about molding data into information, and storing this information based on the subject rather than application. Centralization of data is needed to maximize user access and analysis [Palace, 1996]. As mentioned by W.H. Inmon, in one of his articles, “the data warehouse environment is the foundation of decision support systems (DSS) [Inmon, 1995]”.

II. Proposed Work

It involves critically evaluating the quality of the output of the data mining algorithm from the different data mining techniques. It includes the predictions of the model as well as the interpretation of the fitted model itself. Evaluation of a model can be performed by it applicability on different test data sets and also by considering time and cost factors involved in providing results. Time to time evaluations may also be carried out to maintain its performance. This may sound like a simple operation, but in fact, it sometimes involves an elaborate process. In the present research work, proposed model will be evaluated by applying the designed algorithms based on nearest neighbor; clustering and decision tree methods and implement the proposed model using java and jsp to build a web based evaluation tool.

III. Research Methodology

Exclusive literature survey has been carried out to provide a platform for the initiation of present research work. It’s found that most of authors, either gave a theoretical approach to mining, or too complex methods have been used to implement. After revision of several papers and the methods available, the present research work has been taken-up. Literature survey as part of this thesis work presents a brief view of data warehouse, data mining and their methods. The study elaborated different definitions, methods articulated so far. It is concluded that data warehousing is a subject oriented, integrated, time varying, non-volatile collection of data and not just a central repository. Data mining is technique of warehousing to analyze data from different perspective to find meaningful relationships, patterns or other significant correlations between data. The mining technique helps in extraction of hidden information, allow the business house to make proactive and knowledge driven decisions. There are various techniques of data mining like; clustering, fuzzy logic, decision tree and k-nearest neighbor but each with its own limitations. It is well known fact that every task starts from collection of data and its classification, to provide a platform for future projections to work upon. Thus clustering may be an important technique for the present study as the same is used for classification of data. Fuzzy logic resembles human reasoning, hence can be used in conjunction with clustering for smooth classifications.

Data Mining Web Tool

It is Java/J2EE based newly developed web application with user interface to analyse various data sources with already implemented data mining algorithms like clustering, associate, predictive, K-means and decision tree. It has feature to implement new algorithms. As security measure, tool does not store input data source.

IV. Conclusion

Implementation of data warehousing and data mining includes the conversion of data from various source systems into a common format with accuracy. This study focuses on its most important application that is the use of clustering technique in business in which it may provide an aid in analysis and reporting for discovering various important and hidden facts. Main disadvantage lies in Clustering algorithms is the rigidness in the clustering which does not provides the natural decision making, for smoothening of the boundaries, fuzzy logic is a good method that classify the data in a more simplified and likely way. The utilization of fuzzy logic in clustering is that it offers flexibility in classification of data and gives a more acceptable and obvious decision making. Fuzzy logic works on the basis of degree of membership and classify the data on the basis of magnitude of this degree. To make the decision more factual and valid, nearest neighbor method is useful to mine the data in such a way that the unknown information for an input case can be predicted. Pruned multi-path tree method includes the advantage of clustering and nearest neighbor method to find missing values as well as outliers along with certain assumptions to optimize this proposed method. If there is a single attribute as the predictor, it will be easy to locate an outlier by clustering the values and remove the row value that is far from the clusters by calculating their relative distance from the centroid of disjoint clusters but if the more than one attributes are used as predictors

V. Future Work

As every work have its own limitations and a scope for future findings. By the above discussions, the following areas can be considered as the future scope of research work:

The current thesis focuses on few major techniques of data mining say, clustering, fuzzy logic, k-nearest neighbor and decision trees. The other techniques of mining like neural networks, optimization and visualization may also be explored and implemented in tool for future research.

In the current thesis, a dummy ‘example database’ is used that has a finite size. As the real world data is very huge with plenty of records and attributes, implementation of proposed methods on live data may also be taken-up by the research scholars. Industry data or a bank data will be sufficient to make the proposed methods more factual and valid. The present work concentrated on the data mining methods, model-building, validation and implementation in a web based application.

Being a vast area of research, sub-domain i.e. web mining of key domain- data mining could not be covered to the entire range. It is a sub-section of data mining and can be considered as another important area for the future research work.

Other issues such as handling of categorical values, large size database, time complexity in handling of categorical values as well as self-updating capability of the proposed models via learning through knowledge e.g. neural networks, may also be helpful for future research work.

References

[1] Chen, Y. and L.-l. Qu. The Research of Universal Data Mining Model SYSTEM BASED on Logistics Data Warehouse and Application. in Management Science and Engineering, 2007. ICMSE 2007. International Conference on. 2007.

[2] Viqarunnisa, P., H. Laksmiwati, and F.N. Azizah. Generic data model pattern for data warehouse. in Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. 2011.

[3] ElDahshan, K.A. and H.M.S. Lala. Mining uncertain data warehouse. in Internet Technology and Secured Transactions (ICITST), 2010 International Conference for. 2010.

[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] Trifan, M., et al. An ontology based approach to intelligent data mining for environmental virtual warehouses of sensor data. in Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS 2008. IEEE Conference on. 2008.

[6] Dongkwon, J. and M. Songchun. Scalable Web mining architecture for backward induction in data warehouse environment. In TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. 2001.

[7] Bora, S. Data mining and ware housing. In Electronics Computer Technology (ICECT), 2011 3rd International Conference on. 2011. IEEE.

[8] Yi, L. and P. Yongjun. Application of Digital Content Management System Based on Data Warehouse and Data Mining Technology. In Computational Intelligence and Communication

Networks (CICN), 2012 Fourth International Conference on. 2012.

[9] LePine, J.A. and A. Wilcox-King, EDITORS'COMMENTS: DEVELOPING NOVEL THEORETICAL INSIGHT FROM REVIEWS OF EXISTING THEORY AND RESEARCH. Academy of Management Review, 2010. 35(4): p. 506-509.

[10] Huifang, Z. and P. Ding. A knowledge discovery and data mining process model in E-marketing. In Intelligent Control and Automation (WCICA), 2010 8th World Congress on. 2010.

[11] Zhen, L. and G. Minyi. A proposal of integrating data mining and online analytical processing in data warehouse. In Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on. 2001.

[12] Nimmagadda, S.L., et al. On new emerging concepts of modeling petroleum digital ecosystems by multidimensional data warehousing and mining approaches. In Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on. 2010.