INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

An Implementation Of DM And DWH Concepts In Enterprises

S.M.Hansa
Research Scholars, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai, India
L.Chitra Devi
Research Scholars, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai, India
Dr.G.N.K.Suresh Babu
Professor and Head, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai, India

Abstract:
With the development and penetration of data mining within different fields and industries, many data mining algorithms have emerged. The selection of a good data mining algorithm to obtain the best result on a particular data set has become very important. What works well for a particular data set may not work well on another. The requirements for data mining systems for large organisation and enterprises range from logical and physical distribution of large data and heterogeneous computational resources to the general need for high performance at a level that is sufficient for interactive work. Data mining has many advantages across different industries. It allows large historical data to be used as the background for prediction. The interpretation and evaluation of the patterns obtained by data mining produces new knowledge that decision-makers can act upon. Data mining provides a means to obtain information that can support decision making and predict new business opportunities. For example, telecommunications, stock exchanges, and credit card and insurance companies use data mining to detect fraudulent use of their services; the medical industry uses data mining to predict the effectiveness of surgical procedures, medical tests, and medications; and retailers use data mining to assess the effectiveness of coupons and special events.

***Keywords** : Data mining, Prediction, Association Rules, Clustering*

1.Introduction

Data are pervasive. Machines record our program schedules, preferences, events, achievements, buying and selling. Even our comings and goings are recorded today. As the volume of data increases, it becomes a more difficult task to comprehend it. Information technology has made it possible to manage a huge volume of data electronically and to be able to search for potentially very useful knowledge hidden inside this deep ocean of data. Data Mining, the methodology for the extraction of knowledge from data, seems the only solution to this ever growing problem. With the technological advancements, businesses have gone online. The World Wide Web has, since then, been the ultimate and vast source of information. For example, people today can buy desired things by just clicking on a button in the computer. Because of the growing popularity of the World Wide Web, many websites typically experience thousands of visitors everyday. Analysis of who browsed what can give important insight into, for example, what are the buying patterns of existing customers. Interesting information extracted from the visitors browsing data help analysts to predict, for example, what will be the buying trends of potential customers. Correct and timely decisions made based on this knowledge have helped organizations in reaching new heights in the market. It has many web pages that provide information such as courses being taught in the computer science department, research topics, timetables and international programs. Web servers store information of each page requested by web visitors is called the web access log. Web Usage Mining addresses the problem of extracting behavioral patterns from one or more web access logs. The first step, preprocessing, is the task of accurately identifying pages accessed by web visitors. This is a very difficult task because of page caching and accesses by web crawlers. The second step, pattern discovery, involves applications of data mining algorithms to the preprocessed data to discover patterns. The last step, pattern analysis, involves analysis of patterns discovered to judge their interestingness.

Data mining techniques such as association rules, classification, clustering and attribute selection are considered very useful in web usage mining. Association rules and the correlations among the items in large data sets. Examples of association rules in accessing an educational website can be:

55 % of the web visitors who visited ``college'' page also visited ``scholarship'' page.

48 % of the web visitors who visited ``degrees'' page also visited ``subjects'' page.

70 % of the web visitors who visited ``data'' page also visited ``data-mining'' page.

3.Basics Data Mining Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

The most commonly used techniques in data mining are:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree

methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

4.How Data Mining Works

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship’s captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure.

This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer. For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. In neither case would you achieve the results you desired and of course you have the opportunity to do much better than random - you could use your business experience stored in your database to build a model.

As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long distance calling usage. The good news is that you also have a lot of information about your prospective customers: their age, sex, credit history etc. Your problem is that you don't know the long distance calling usage of these prospects (since they are most likely now customers of your competition). You'd like to concentrate on those prospects who have large amounts of long distance usage. You can accomplish this by building a model. Table 1 illustrates the data used for building a model for new customer prospecting in a data warehouse.

	Customers	**Prospects**
General information (e.g. demographic data)	Known	Known
Proprietary information (e.g. customer transactions)	Known	Target

Table 1: Data Mining for Prospecting

The goal in prospecting is to make some calculated guesses about the information in the lower right hand quadrant based on the model that we build going from Customer General Information to Customer Proprietary Information. For instance, a simple model for a telecommunications company might be:

98% of my customers who make more than $60,000/year spend more than $80/month on long distance

5.An Architecture For Data Mining

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

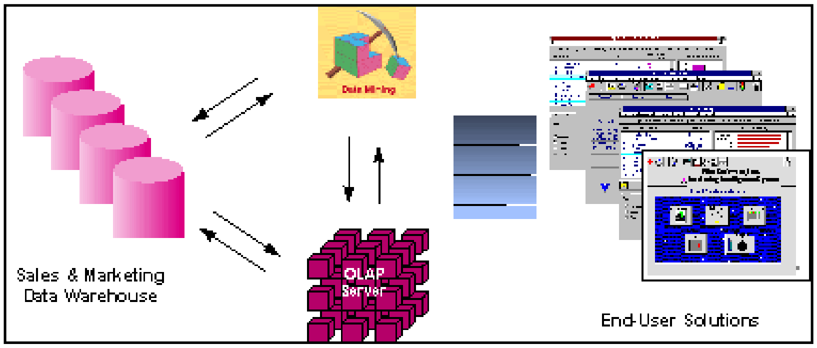

Figure 1: Integrated Data Mining Architecture

6.Mining Techniques And Pattern Discovery

Once the data are prepared and formatted, the data mining techniques are applied to discover patterns. For the experiments of this thesis, a data mining tool, WEKA is used. The WEKA is a collection of machine learning algorithms for solving real-world data mining problems. It is open source software issued under the GNU General Public License. Data mining tools for preprocessing, classification, clustering, attribute selection and visualization are implemented in the WEKA. The data mining techniques used for the experiments are described below.

1. Classification: Classification is a technique of supervised learning meaning that the number of categories or classes that the instances fall into are known in advance before the pattern discovery process starts. In the data set, an attribute that represents the classes is known as a class variable.

In supervised learning, a model is first built using the training data. Once the model is built, it is then applied to the test data. One of the problems found in supervised learning is that the difference in the accuracies at correctly classifying instances of training and test data can be considerable. This problem is referred to as over writing. One of the solutions to this problem is to rerun the algorithm with different values of the parameters attempting to make the results less specific to the training data. The technique of cross-validation, which is based on re-sampling" , is used for estimating the error rate. In a p-fold cross-validation, data are divided into p subsets of equal size. The model is trained p times, with a different subset omitted from the training set each time. The accuracy is measured only for the omitted subset.

7.Enterprise Data Mining Requirements

Data mining system architectures for enterprises have to meet a range of demands from the field of data analysis and the additional needs that arise when handling large amounts of data inside an organisation. Modern data mining applications are expected to provide a high degree of integration while retaining flexibility. In this way they can efficiently support different types of analyses over the organisation's data. Data mining is understood to be an iterative process for the analyst, especially in the initial exploratory phases of the analytical task. Therefore, a high degree of interactivity is required, often combined with the need for visualisation of the data and the analytical results.

The field of data mining is developing rapidly, and the methods applied in a tool today may be superseded by more advanced algorithms in the near future. Furthermore, the convergence with statistical methods has only just started, and will grow in pace over the next few years. The need for enhancement of the existing tool set has to be reflected by a software architecture that enables the straightforward integration of new analytical components. In a similar vein, the results from the analytical functions need to be presented in portable formats, as most analysts will want to use different specialist packages to further refine or report the results.

In large organisations, the amount and the distribution of the data become an additional challenge. The size of the data may make it impractical to move it between sites for individual analytical tasks. Instead, data mining operations are required to execute \close to the database". In the absence of dedicated support for data mining and other analytical algorithms in the database management systems, this can be achieved by setting up high-performance servers in close proximity to the databases. The overall data mining system will then have to manage the distributed execution of the analytical tasks and the combination of the partial results into a meaningful total. Also, this approach can sometimes benefit from the improved generalisation power that often occurs when combining analytical models [CS96].

Three kinds of scalability requirements arise in enterprise environments:

- data sets can be very large
- there may be many sites on which data is accumulated
- there may exist many users who need access to the data and the analytical results

An enterprise data mining architecture should be exible enough to scale well in all these cases. This will require access to high-performance analytical servers (case 1), the ability to distribute the application (case 2) and the capability to provide multiple access points (case 3).

8.Conclusion and Future Work

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs. Building models is only one step in knowledge discovery. It’s vital to properly collect and prepare the data, and to check the models against the real world. The “best” model is often found after building models of several different types, or by trying different technologies or algorithms. Choosing the right data mining products means finding a tool with good basic capabilities, an interface that matches the skill level of the people who’ll be using it, and features relevant to your specific business problems. After you’ve narrowed down the list of potential solutions, get a hands-on trial of the likeliest ones. Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs. Building models is only one step in knowledge discovery. It’s vital to properly collect and prepare the data, and to check the models against the real world. The "best" model is often found after building models of several different types, or by trying different technologies or algorithms.

Choosing the right data mining products means finding a tool with good basic capabilities, an interface that matches the skill level of the people who'll be using it, and features relevant to your specific business problems.

9.Reference

1. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 12-23, 2000.
2. K.W. Tan, H. Han, and R. Elmasri. Web data cleansing and preparation for ontology extraction using WordNet. In First International Conference on Web Information Systems Engineering (WISE'00), volume 2.
3. Feng Tao and Fionn Murtagh. Towards knowledge discovery from www log data. In Proc. of the International Conference on Information Technology: Coding and Computing, 2000.
4. J. Chattratichat, J. Darlington, M. Ghanem, Y. Guo, H. Huning, M. Kohler, J. Sutiwaraphun, H. W. To, and D. Yang. Large scale data mining: Challenges and responses. In Proceedings of Third International Conference on Knowledge Discovery and Data Mining, pages 143-146, 1997.
5. J. Chattratichat, J. Darlington, Y. Guo, S. Hedvall, M. Kohler, A. Saleem, J. Sutiwaraphun, and D. Yang. A software architecture for deploying high-performance solutions on the internet. In High-Performance Computing and Networking, 1998.
6. P. K. Chan and S. J. Stolfo. Sharing learned models among remote database partitions by local meta-learning. In E. Simoudis, J. Han, and U. Fayyad, editors, The Second International Conference on Knowledge Discovery and Data Mining, pages 2-7. AAAI Press, 1996.
7. A. Berson, S. Smith, Data Warehousing, Data Mining, and OLAP, McGraw Hill, 1997.
8. M. Bertero, T. Poggio and V. Torre, Ill-Posed Problems in Early Vision, Proceedings of the IEEE, Vol.76, No.8, pp.869-902, 1990.
9. M. Bertold and D. Hand, Intelligent Data Analysis: An Introduction, Springer Verlag, 1999.
10. Wong, W., Moore, A., Cooper, G. & Wagner, M. (2003). Bayesian Network Anomaly Pattern Detection for Detecting Disease Outbreaks. Proc. of ICML03, 217-223.
11. Yamanishi, K., Takeuchi, J., Williams, G. & Milne, P. (2004). On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. Data Mining and Knowledge Discovery **8**: 275-300.

12. Yeung, D. & Ding, Y. (2002). User Profiling for Intrusion Detection Using Dynamic and Static Behavioural Models. Proc. of PAKDD2002, 494-505.
13. Yoshida, K., Adachi, F., Washio, T., Motoda, H., Homma, T., Nakashima, A., Fujikawa, H. & Yamazaki, K. (2004). Density- Based Spam Detector. Proc. of SIGKDD04, 486-493.