Study of Influence of Data Mining & Data Warehousing

[1]Poonam Yadav , [2]Sarla Kumari
[1]Assistant Professor , [2]Assistant Professor
[1,2]F.L.T.M.S.B.P. Govt. Girls College , Rewari
[1]poonamy162@gmail.com

Abstract: Data Mining is a combination of Database and Artificial Intelligent used to provide useful information to both technical and non-technical users which will help them to make better decisions. It is usually used as a decision support system. A data warehouse is a relational database that is designed for query and analysis rather than transaction processing. It usually contains historical data that is derived from transaction Data in the warehouse can be seen as materialized views generated from the underlying multiple data sources. The aim of this paper is to show the importance of using data warehousing and data mining . It also aims to show the process of data mining and how it can help decision makers to make better decisions.

***Keywords*: Data mining, hidden predictive, warehouse, component, Data Warehousing.**

I. Introduction

The core component of these systems is the data warehouse and nowadays it is widely assumed that the data warehouse design must follow the multidimensional paradigm. Thus, many methods have been presented to support the multidimensional design of the data warehouse. The data warehouse is a huge repository of data that does not tell us much by itself; like in the operational databases, we need auxiliary tools to query and analyze data stored. Data warehouse has more than one definition. The most common one is defined by Bill Inman who defined it as the following: "A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. The past couple of decades have seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the sizes as well as number of databases are increasing even faster. There are many examples that can be cited. Point of sale data in retail, policy and claim data in insurance, medical history data in health care, financial data in banking and securities, are some instances of the types of data that is being collected.

Data mining refers to "using a variety of techniques to identify nuggets of decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The detail often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful" .Data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data. The idea is that it is possible to strike gold in unexpected places as the data mining software extracts patterns not previously discernable or so obvious that no-one has noticed them.

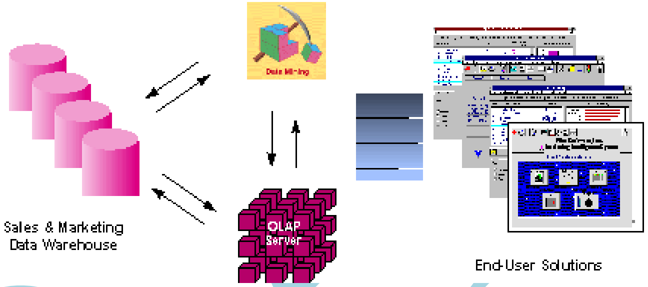

Figure 1 - Integrated Data Mining Architecture

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

II. Comprehensive Survey

Methods here discussed were selected according to three factors: reference papers with a high number of citations according to Google Scholar and Publish or Perish, papers with novelty contributions and in case of papers of the same authors, we have included the latest version of their works. As general rule, each method is described using the terminology presented in the Terminology and Notation section. Finally, we follow a chronological order when introducing the design methods surveyed. As an exception, when a method publications span over several different papers, we place them at the chronological point occupied by their first paper but we quote them by means of the most relevant paper. All in all, this section provides a comprehensive framework of the evolution of multidimensional design methods.

(Kimball et al., 1998) introduced multidimensional modeling as known today. In addition, they also introduced the first method to produce the multidimensional schema. Data marts are essentially defined as pragmatic collections of related facts. Although data sources are not considered, they already suggest to take a look at the data sources to find which data marts may be of our interest .Next step aims to list all conceivable dimensions for each data mart. At this point it is suggested to build an ad hoc matrix to capture our multidimensional requirements. Rows represent the data marts, whereas columns represent the dimensions. A given cell is marked whether that dimension must be considered for a data mart. This matrix is also used to show the associations between data marts by looking at dimensions shared. This process is supposed to be incremental.

First, it is suggested to focus on single-source data marts, since it will facilitate our work and later in a second iteration, look for multiple-sources data marts combining the single-source designs. Being the first approach, it does not introduce a formal design procedure, but a detailed guide of tips to identify the multidimensional concepts and then, give rise to the multidimensional schema. The presentation is quite informal and it relies on examples rather than on formal rules. Kimball's approach follows a demand-driven framework to derive the data warehouse relational schema, as follows. First, we must declare the grain of detail. It is suggested to be defined by the design team at the beginning, although it can be reconsidered during the process. Normally, it must be determined by primary dimensions. Next, we choose the analysis dimensions for each fact table. Dimensions selected must be tested against the grain selected. This must be a creative step. We need to look for the dimension *pieces* in different (and potentially heterogeneous) models and through different documents.

III. Data Mining Techniques

The Most Commonly Used Techniques in Data Mining are:

1) Artificial neural networks
2) Decision trees
3) Genetic algorithms
4) Nearest neighbour method
5) Rule induction

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. 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. The appendix to this white paper provides a glossary of data mining terms

IV. Data warehousing and characteristics

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make better and faster decisions. Data mining potential can be enhanced if the appropriate data has been collected and stored in a data warehouse. A data warehouse is a relational database management system designed specifically to meet the needs of transaction processing systems. It can be loosely defined as any centralized data repository which can be queried for business benefit but this will be more clearly defined later. Data warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats. As well as integrating data throughout an enterprise, regardless of location, format, or communication requirements it is possible to incorporate additional or expert information.

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inman, author of Building the Data Warehouse and the guru who is widely considered to be the originator of the data warehousing concept, is as follows:

a) Subject Oriented
b) Integrated
c) Non volatile
d) Time Variant

Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case makes the data warehouse subject oriented.

Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

For instance, in one application, gender might be coded as "m" and "f" in another by 0 and 1. When data are moved from the operational environment into the data warehouse, they assume a consistent coding convention e.g. gender data is transformed to "m" and "f".Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

V. Conclusion

Data warehouse is used as a central store of a subject oriented, integrated, time-variant and non-volatile collection of data from different sources [1]. For faster performance, data warehousing organizes data in a different architecture – fact table and dimension tables [4]. For that reason modeling the data warehouse is unlike modeling the operational database. Data mining has become an important tool which can extract useful information from the huge amount of data we have nowadays. It is an iterative process which includes feedbacks between the phases and sometimes needs to repeat the entire process from the beginning. The iterations are needed in the mining process in order to provide better answers which will be used by the users to make better decisions.

REFERENCES

[1] Chen, Y. and L.-l. Qu. The Research of Universal Data Mining Model SYSTEM BASED on Logistics Data Warehouse and Application..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] Nimmagadda, S.L. and H. Dreher. On designing multidimensional oil and gas business data structures for effective data warehousing and mining. in Digital Ecosystems and Technologies, 2009. DEST '09. 3rd IEEE International 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] Eisenhardt, K.M., Building Theories from Case Study Research. The Academy of Management Review, 1989. **14**(4): p. 532-550.

[7] Usman, M. and R. Pears. A methodology for integrating and exploiting data mining techniques in the design of data warehouses. in Advanced Information Management and Service (IMS), 2010 6th International Conference on. 2010.

[8] Sung Ho, H. and P. Sang-Chan. Data modeling for improving performance of data mart. in Engineering and Technology Management, 1998. Pioneering New Technologies: Management Issues and Challenges in the Third Millennium. IEMC '98 Proceedings. International Conference on. 1998.

[9] Yuekun, M., et al. Implementation of Metadata Warehouse Used in a Distributed Data Mining Tool. in Challenges in Environmental Science and Computer Engineering (CESCE), 2010 International Conference on. 2010.

[10] Chieh-Yuan, T. and T. Min-Hong. A dynamic Web service based data mining process system. in Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on. 2005.

[11] Ding, P. A formal framework for Data Mining process model. in Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on. 2009.

[12] Tsumoto, S., et al. Exploratory temporal data mining process in hospital information systems. in Cognitive Informatics & Cognitive Computing (ICCI*CC), 2012 IEEE 11th International Conference on.2012.