A New Approach towards Active Data Retrieval Technique Warehouse and Mining, Big Data and Data Warehousing

Amit Kumar
Research Scholar (IT)
Faculty Member BCA Deptt.,
M.P.S.Sc. College,
B.R.A. Bihar University,
Muzaffarpur, Bihar

Raj Kumar
Ph. D (IT),
B.R.A. Bihar University,
Muzaffarpur, Bihar

Alok Ranjan Tripathi
Head of Department,
Deptt .of Mathematics,
M.P.S.Sc. College,
B.R.A. Bihar University,
Muzaffarpur, Bihar

Abstract

In this article, Business intelligence is the use of information that enables organizations to best decide, measure, manage and optimize performance to achieve efficiency and financial benefit. We employ the data warehousing technology as a preprocessing step to apply piecewise regression as a derivative data mining technique that fits a data model which will be used for prediction. Also, correlation analysis is used to qualify the best of the proposed models for that purpose. The conclusion of this paper is that a data warehouse accompanied with a suitable data mining technique represents an effective platform for data mining.

Keywords: Data Mining, Data Warehouse, Derivative, Marketing.

Introduction

During the past 10 years, business intelligence (BI) has changed and expanded to support a broader set of applications, business requirements, information stores (data warehousing, data mart, operational data store and transactional database management systems) and users (analyst, executives, knowledge workers, sales managers, IT managers, partners and more). At Gartner's Symposium conference in Australia, we conducted an informal survey with 150 IT and business users. The survey was given to clients attending a diverse set of presentations and supplied 14 options of possible definitions for BI — including enabling users to fill in their own definition. Interestingly, 43 percent of the clients surveyed viewed BI as "The use of information that enables organizations to best lead, decide, measure, manage and optimize to achieve efficiency and financial benefit." In addition, 16 percent of those surveyed viewed BI as "The ability to leverage data/information in a specific functional process (or application) to enable context-specific insight that can be translated into action." Another 16 percent viewed BI as "Access to and analysis of quantitative information sources to deliver insight to its users to better align people and processes with business objectives." Fewer than 5 percent viewed BI as "Tools and technologies (reporting, mining) that help analysts analyze data." The two critical insights we noted from this survey were that clients viewed BI's value as more than information dissemination; BI is highly linked to achieving business goals. We believe that BI capabilities will become more pervasive in operational and workplace applications as organizations seek to leverage BI to lead, support decisions, explore measure, manage and optimize their businesses.

Bottom Line

Seek new opportunities to leverage BI applications, tools, methods and practices to support a broad base of business needs beyond information dissemination. Plan for BI to take on a more-pervasive role in supporting business value as it emerges as a core capability of new workplace applications. Plan an overall BI application portfolio to match your overall business objectives rather than enabling silos of tools and technology. In addition to using BI to manage and optimize their businesses, organizations will increasingly use new capabilities to help discover and explore as BI becomes more pervasive in operational and workplace applications.

Any good decision is based on information. It is therefore crucial that businesses have a good source of business data that can be used quickly and flexibly by managers. Data Warehousing and Knowledge Discovery technology is emerging as a key technology for enterprises that wish to improve their data analysis, decision support activities, and the automatic extraction of knowledge from data. Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane (such as supermarket transaction data, credit card usage

records, telephone and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database It is estimated that the amount of information in the world doubles every 20 months. That is, many scientific, government and corporate information systems are being overwhelmed by a flood of data that are generated and stored routinely, which grow into large databases amounting to giga (and even tera) bytes of data. These databases contain potential gold mine of valuable information, but it is beyond human ability to analyze massive amounts of data and elicit meaningful patterns. The data mining is emerged to emphasize the challenges of searching for knowledge in large databases. It comes from the idea that large databases can be viewed as data mines containing valuable information that can be discovered by efficient knowledge discovery techniques. Data mining will continue to take place in environments not supporting a data warehouse. However, as volumes of data continue to be collected for purposes of decision support, the need for organized, efficient data storage and retrieval architectures has become quite apparent. The result of this need has sparked the birth of the data warehouse. The construction of data warehouses can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Furthermore, many other data mining functions, such as classification, prediction, association, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. Hence, the data warehouse has become an increasingly important platform for data analysis and on-line analytical processing and will provide an effective platform for data mining. Building a data warehouse is a very challenging issue because compared to software engineering it is quite a young discipline and does not yet offer wellestablished strategies and techniques for the development process. A lot of projects fail due to the complexity of the development process. As yet there is no common strategy for the development of data warehouses.

We use the data warehousing technology as a preprocessing step to apply piecewise regression as a derivative data mining technique that fits a data model which will be used for prediction

Case of Study

We apply this new approach to “marketing” data aiming at understanding consumer behavior, forecasting product demand, managing and building the brand, tracking performance for customers or products in the market and driving incremental revenue from transforming data into information and information into knowledge. In this case study, the subject area would be Sales. The dimensions of analysis would be Customer, Product, and Area. The requirement is to analyze sales by customer, sales by product, and sales by area. Time is regarded as a necessary dimension of analysis (time variance is a characteristic of data warehouses) and so is always included as one of the dimensions of analysis. We use the snowflake schema as a multidimensional model for building the data warehouse as shown in figure1.

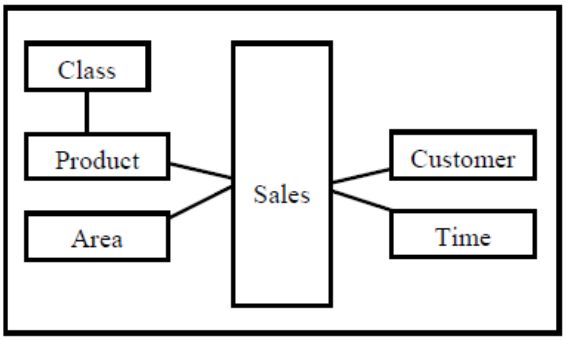

Figure 1: the snowflake schema for food mart

Methodology

Data mining can be used for a variety of purposes. There are many legitimate ways of describing the types of task that can be performed using data mining techniques, but even across differing categorizations five tasks consistently emerge:

Classification

The act of distributing objects into predefined classes or categories.

Estimation

Predicting the value of an unknown, continuous variable.

Clustering

Identifying logical groups in which to place similar objects.

Prediction

Classifying, estimating or clustering about a value or behavior that has yet to occur.

Affinity Analysis

Determining which objects can be expected to co-occur with other objects.

These tasks can be applied to solve a wide variety of problems across different industries. In many cases, different approaches, or a combination of approaches, will be needed to solve the problem.

The goal of this research was performance amelioration in the data mining process. This is done using the data warehousing technology as a preprocessing step for data mining in conjunction with regression analysis as a derivative data mining technique to fit a data model and to make predictions based on such a model. Also, correlation analysis is used to qualify the best of the proposed models for that purpose.

Step1: Building the summary tables (Preprocessing step)

The heart of any data warehouse is its database, where all the information is stored. Most queries will need to access hundreds and thousands of rows in order to answer questions about sales over time. It will take quite a long time. Some queries are quite complex, involving multiple join paths, and this will seriously increase the time taken for the result set to be presented back to the user, perhaps to several hours. The problem is exacerbated when several people are using the system at the same time, each with acomplex query to run.

Since almost all queries would be summarizing large numbers of rows together and returning a result set with a smaller number of rows. So if we can predict the types of queries the users will mostly be executing, we can prepare some summarized fact tables so that the users can access those if they happen to satisfy the requirements of the query. The snowflake schema principles still apply, but the result is that we have several.

One-Dimensional Summary Tables

In this type we build summary tables moving through one dimension of analysis. For our first summary table, to build one-dimensional summary table analyzing sales by Time dimension, the individual sales for each quarter have to be added together to form a total for that quarter for each promotion event (we might also summarize over time monthly, but here we summarize over time quarterly which is a higher level of summarization).

Two-Dimensional Summary Tables

In this type we build summary tables moving through two dimensions of analysis. i.e., to build a twodimensional summary table analyzing sales by Time and Area dimensions respectively, the individual sales for each quarter in each customer-district have to be added together to form a total for this customerdistrict for that quarter for each promotion event.

Three-Dimensional Summary Tables

In this type we build summary tables moving through three dimensions of analysis. We build a threedimensional summary table analyzing sales by Time, Product and Area dimensions respectively, the individual sales for each quarter for each product subcategory in each customer-district have to be added together to form a total for this product subcategory that have been sold in this customer district for that quarter of year for each promotion event.

Step2: Applying Piecewise Regression

We use regression to forecast how sales may respond to various advertising campaigns.

Our approach is to apply the piecewise regression technique on the summary tables. This is a recurring theme in data mining; the idea of composing complex global structures from relatively simple local components. That is, the locality also provides a framework for decomposing a complex model into simpler local patterns.

We applied piecewise regression technique through the dimensions of analysis in different types as follows:

One-Dimensional Piecewise

In this type we have applied the piecewise regression technique reading from one of the one-dimensional summary tables that have been constructed at the preprocessing stage.

Two-Dimensional Piecewise

In this type we have applied the piecewise regression technique reading from one of the two-dimensional summary tables that have been constructed at the preprocessing stage.

Three-Dimensional Piecewise

In this type we have applied the piecewise regression technique reading from one of the three-dimensional summary tables that have been constructed at the preprocessing stage.

Experimental Results

The local patterns that have been detected when we studied the problem through Time, Product and Area dimensions together were relatively well at modeling the data. The coefficient of correlation for the local patterns that have been detected when we applied the piecewise regression technique reading from the corresponding summary table

Conclusion

Using interactive and advanced visualization enables users to explore data by dynamic linking by color, selection and filtering. This usage model completely surpasses the more traditional, linear model of data interaction, whereby users must predetermine where they want to gorilla navigate and they are often restricted by the types and amounts of data to which they have access. By simply pointing-and-clicking, business users are able to rapidly create their own views of the data.

References

1. David Hand, Heikki Mannila, and Padhraic Smyth; “Principles of data mining”. MIT Press, 2001.
2. Jiawei Han and M. Kamber; “Data Mining: Concepts and Techniques”. Morgan Kaufmann publishers, 2001
3. Beate List, Robert M. Bruckner, Karl Machaczek, and Josef Schiefer; “A Comparison of Data Warehouse Development Methodologies - Case Study of the Process Warehouse”, proceedings of the 13th International Conference on Database and Expert Systems Applications (DEXA 2002), pp. 213-225, Springer 2002.
4. Kimball, Ralph and Ross, Margy. The Data Warehouse Toolkit Second Edition (2002) John Wiley and Sons, Inc. ISBN 0-471-20024-7
5. Linstedt, Graziano, Hultgren. The Business of Data Vault Modeling Second Edition (2010) Dan linstedt, ISBN 978-1-4357-1914-9
6. William Inmon. Building the Data Warehouse 2005) John Wiley and Sons, ISBN 978-8-1265-0645-3